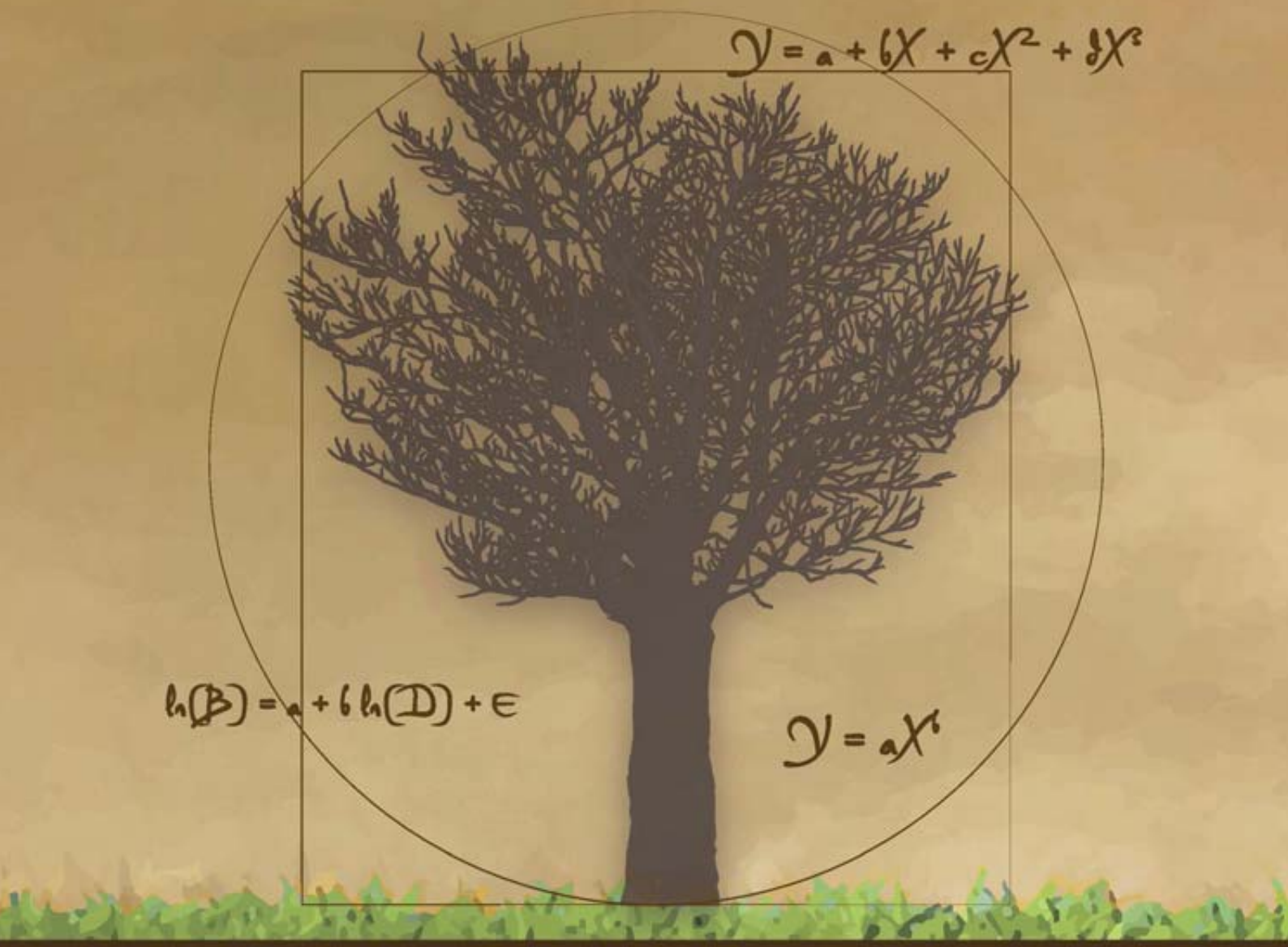
Manual de construcción de ecuaciones alométricas para estimar el volumen y la biomasa de los árboles

Del trabajo de campo a la predicción



cirad

FAO

Manual de construcción de ecuaciones alométricas para estimar el volumen y la biomasa de los árboles

Del trabajo de campo a la predicción

Nicolas Picard
Departamento del Medioambiente y Sociedad
Centre de Coopération Internationale en Recherche Agronomique pour le Développement

Laurent Saint-André
UMR Eco&Sols
Centre de Coopération Internationale en Recherche Agronomique pour le Développement
&
UR1138 BEF
Institut National de la Recherche Agronomique

Matieu Henry
Departamento Forestal
Organización de las Naciones Unidas para la Alimentación y la Agricultura

Agosto de 2012

E-ISBN 978-92-5-307347-4

Las Naciones Unidas para la Alimentación y la Agricultura (FAO)
Viale delle Terme di Caracalla
00153 Rome, Italie

Centre de Coopération Internationale en Recherche Agronomique pour le Développement (CIRAD)
Campus international de Baillargeut
34 398 Montpellier Cedex, France

Créditos fotográficos: Stephen Adu-Bredu (Foto 3.5), Rémi D'Annunzio (Foto 3.4 y Figura 3.2), Astrid Genet (Fotos 3.13 y 3.14), Matieu Henry (Fotos 3.8 y 3.10), Christophe Jourdan (Fotos 3.11 y 3.12 y Figura 3.8), Bruno Locatelli (Foto 1.2), Claude Nys (foto 3.7 y Figura 3.2), Régis Peltier (Foto 3.9), Jean-François Picard (Foto 3.15 y Figura 3.2), Michaël Rivoire (Fotos 3.3, 3.5 y 3.14 y Figura 3.2), Laurent Saint-André (Fotos 1.1, 3.3, 3.4, 3.6, 3.8 y 3.11 y Figura 3.2).

Citación recomendada: Picard N., Saint-André L., Henry M. 2012. Manual de construcción de ecuaciones alométricas para estimar el volumen y la biomasa de los árboles: del trabajo de campo a la predicción. Las Naciones Unidas para la Alimentación y la Agricultura y el Centre de Coopération Internationale en Recherche Agronomique pour le Développement, Rome, Montpellier, 223 págs.

Índice general

Índice de figuras

Índice de fotos

Índice de cuadros

Índice de líneas rojas

Prefacion

En la Convención Marco de las Naciones Unidas sobre el Cambio Climático (CMNUCC), los beneficios potenciales para las Partes no incluidas en el Anexo I que disminuyan sus emisiones de gases de efecto invernadero se basarán en los resultados mensurables, notificables y verificables. La precisión de dichos resultados tendrá una enorme influencia en las posibles compensaciones financieras. Las mediciones de las reservas de carbono forestal adquieren así una importancia mayor para los países que prevén contribuir a mitigar los cambios climáticos gracias a sus actividades forestales. Estas mediciones recurren actualmente a técnicas que funcionan a escalas diferentes, desde los inventarios de campo realizados a escala local hasta las mediciones de teledetección por satélite que funcionan a escala nacional o subregional, pasando por el láser o el radar aerotransportado. Las mediciones indirectas de las reservas de carbono forestal, como las derivadas de los índices satelitales, Lidar o radar, se basan en las relaciones calibradas a partir de las mediciones efectuadas sobre el terreno. Lo mismo ocurre con los inventarios. Al final de cuentas lo que es seguro es que toda medición del carbono forestal exige que se pesen los árboles sobre el terreno; esta etapa constituye la piedra angular sobre la que se apoya todo el edificio de la estimación de las reservas de carbono forestal, independientemente de las escalas consideradas.

De este modo las ecuaciones alométricas, que permiten predecir la biomasa de un árbol a partir de las características dendrométricas más fáciles de medir (como su diámetro o su altura) son elementos clave para estimar la contribución de los ecosistemas forestales al ciclo del carbono. El presente manual se propone abarcar todas las etapas de su construcción, a partir de la medición de la biomasa de los árboles sobre el terreno. Debería resultar particularmente útil para los países que no disponen aún de mediciones y de modelos de ecuaciones adaptados a sus formaciones forestales.

El presente *Manual de construcción de ecuaciones alométricas para estimar el volumen y la biomasa de los árboles* constituye una guía práctica para estudiantes, técnicos e investigadores que trabajan sobre la evaluación de los recursos forestales tales como el volumen, la biomasa y las reservas de carbono, con objetivos comerciales, bioenergéticos o de mitigación del cambio climático. Los métodos propuestos en este manual se aplican a la mayoría de los bosques y zonas ecológicas, haciendo especial hincapié en los bosques tropicales, los que hoy necesitan, quizás más que los demás, un esfuerzo por parte de la comunidad internacional para medir las reservas de carbono. Se propone una "Línea roja" para guiar al lector: se trata de un caso concreto que ilustra las distintas cuestiones asociadas a la construcción de las ecuaciones alométricas, el muestreo, las mediciones sobre el terreno, el registro de datos, la exploración gráfica de los datos, el ajuste de las ecuaciones y su utilización para la predicción. Los datos utilizados proceden de tres lugares muy diferentes en términos de estructura forestal y de medios disponibles. A partir de ello se dan consejos prácticos que deberían permitir a los lectores hacer frente a la mayoría de los problemas que se encuentran habitualmente. También resultará de interés para los especialistas en biometría forestal en

la medida en la que contiene no sólo referencias exhaustivas a la teoría matemática de la regresión y sus recientes desarrollos sino también numerosos consejos sobre la selección y la utilización de modelos de regresión lineal.

223 páginas. Numerosas ilustraciones. Bibliografía de 255 obras.

Francis Cailliez

Agosto de 2012

Agradecimientos

Los autores desean agradecer a la Organización de las Naciones Unidas para la Alimentación y la Agricultura por haber financiado la edición y la traducción del presente manual.

También desean agradecer a las personas mencionadas a continuación por haber contribuido a las misiones de campo y con los datos utilizados para la línea roja, por haber enriquecido con su propia experiencia el contenido del manual y por haber aceptado releerlo y traducirlo: Dr. Stephen Adu-Bredu, Angela Amo-Bediako, Dr. Winston Asante, Dr. Aurélien Besnard, Fabrice Bonne, Noëlle Bouxiero, Emmanuel Cornu, Dr. Rémi D'Annunzio, Dra. Christine Deleuze, Serge Didier, Justice Eshun, Charline Freyburger, Dominique Gelhaye, Dra. Astrid Genet, Dickson Gilmour, Hugues Yvan Gomat, Dr. Christophe Jourdan, Dr. Jean-Paul Laclau, Dr. Arnaud Legout, Lawrence y Susy Lewis, Dra. Fleur Longuetaud, Dr. Raphaël Manlay, Jean-Claude Mazoumbou, Adeline Motz, Dr. Alfred Ngomanda, Dr. Yann Nouvellon, Dr. Claude Nys, Charles Owusu-Ansah, Thierry Paul, Régis Peltier, Dr. Jacques Ranger, Michaël Rivoire, Gaël Sola, Luca Birigazzi, Dr. Olivier Roupsard, Dr. Armel Thongo M'bou y Prof. Riccardo Valentini.

Los autores agradecen a quienes, a pesar de los breves plazos impuestos, aportaron sus comentarios y sugerencias sobre correcciones así como sus palabras de aliento. Esta obra se benefició mucho con su valiosísimo aporte. Un agradecimiento especial a Miguel Cifuentes Jara, quien se desempeñó como revisor técnico y asesor de idioma de la publicación. Sin embargo, la responsabilidad de su contenido le incumbe únicamente a sus autores.

Los métodos sintéticos presentados en el presente manual fueron elaborados durante las misiones de medición financiadas por los siguientes proyectos: ATP Carbone (CIRAD), MODELFOR (ONF), BIOMASSE OPE (ANDRA), SOERE F-ORE-T (GIP ECOFOR), EMERGE (ANR), WAFT (UE), ULCOS (UE), CarboAfrica (UE, contrato no INCO-CT-2004-003729).

Preámbulo

El presente manual se destina a estudiantes, investigadores o ingenieros que deseen aprender la metodología necesaria para elaborar las tablas de volumen, biomasa o mineralomasa. Estos modelos están reunidos en una obra única porque todos se basan en el mismo principio: estimar un dato difícil de medir en todos los árboles de un rodal (por ejemplo, el volumen) a partir de las características más simples como el diámetro del árbol a 1,30 m, su altura o su edad.

Basado en un conjunto de publicaciones de referencia, este manual no presenta todos los casos posibles sino que propone técnicas que permiten construir las ecuaciones. Las referencias en el texto son precisas (en la medida de lo posible: autor, año, página) para que el lector pueda encontrar fácilmente la información. Un ejemplo concreto (llamado "Línea roja") guía al lector para adquirir los conocimientos mediante la práctica.

Los requisitos previos para la utilización de este manual son escasos. Los programas informáticos utilizados en la "Línea roja" son Microsoft Excel para la preparación de los archivos y R (R Development Core Team, 2005) para adaptar los modelos. Las instrucciones de R utilizadas se reproducen en la línea roja.

1

Las bases de la estimación de la biomasa

A escala de una población existe una relación estadística entre las diferentes medidas de un individuo (Gould, 1966). Esta relación se deriva del desarrollo ontogénico de los individuos que es la misma para todos, salvo la variabilidad asociada a la historia personal de cada uno. Así, las proporciones entre altura y diámetro, entre tamaño de la copa del árbol y el diámetro, entre la biomasa y el diámetro, obedecen a una regla que es la misma para todos los árboles que viven en las mismas condiciones, desde el más pequeño al más grande (King, 1996; Archibald & Bond, 2003; Bohlman & O'Brien, 2006; Dietze *et al.*, 2008). Se trata del principio básico de la alometría que permite predecir una medida de un árbol (lo típico es su biomasa) en función de otra medida (por ejemplo, su diámetro). Una ecuación alométrica es una fórmula que formaliza de forma cuantitativa dicha relación. En el caso de la predicción del volumen, de la biomasa o de la masa mineral, hablaremos en el presente manual de modelos de volumen, biomasa y mineralomasa respectivamente. Existe una definición más restrictiva de alometría que consiste en una relación de la proporcionalidad entre los aumentos relativos de las medidas (Huxley, 1924; Gayon, 2000). Si se observan B la biomasa y D el diámetro, esta segunda definición significa que existe un coeficiente a que corresponde a:

$$\frac{\mathrm{d}B}{B} = a\frac{\mathrm{d}D}{D}$$

que se integra en una relación de potencia: $B = b \times D^a$. Con esta definición restringida, una ecuación alométrica resulta sinónimo de una ecuación de potencia (White & Gould, 1965). El parámetro a da el coeficiente de alometría (proporcionalidad entre los aumentos relativos) mientras que el parámetro b indica una proporcionalidad entre las magnitudes acumuladas. A veces hace falta agregar una intersección en esta relación que se convierte en $B = c + bD^a$, donde c representa la biomasa del individuo antes de que alcance la altura a la cual se mide el diámetro (por ejemplo 1,30 m si D se tomó a 1,30 m). La relación de potencia hace referencia a la idea de autosimilaridad durante el desarrollo de los individuos (Gould, 1971). Partiendo de este principio y apoyándose en la "pipe theory" (Shinozaki *et al.*, 1964a,b), fue elaborada una teoría fractal de la alometría (West *et al.*, 1997, 1999; Enquist *et al.*, 1998, 1999). Si consideramos ciertas hipótesis de limitaciones biomecánicas, de estabilidad de los árboles y de resistencia hidráulica en las redes de células conductoras, esta teoría predice una relación de potencia con un exponente igual a $a = 8/3 \approx 2{,}67$ entre la biomasa y el

diámetro de los árboles. Esta relación es interesante porque se fundamenta en los principios físicos y una representación matemática de las redes de células de los árboles. No obstante, es objeto de un amplio debate en el que a veces se critica su carácter excesivamente general (Zianis & Mencuccini, 2004; Zianis *et al.*, 2005; Muller-Landau *et al.*, 2006), aunque es posible usar otros coeficientes de alometría según las hipótesis biomecánicas e hidráulicas utilizadas (Enquist, 2002).

En el marco de este manual adoptaremos la definición más amplia de la alometría que hace referencia a una relación (lineal o no) entre los aumentos de las medidas de los árboles. La relación de potencia será considerada simplemente una relación alométrica entre otras. Independientemente de la definición adoptada, la alometría se refiere al desarrollo ontogénico de los individuos, es decir, al crecimiento de los árboles.

1.1. La "biología": ley de Eichhorn, site index...

El crecimiento de los árboles es un fenómeno biológico complejo (Pretzsch, 2009) que resulta de la actividad de las yemas (crecimiento primario o aumento de la longitud de los ejes) y del cambium (crecimiento secundario o aumento del espesor de los ejes). Este crecimiento de los árboles es obviamente variable ya que depende del patrimonio genético del individuo, de su entorno (suelo, atmósfera), de la etapa de desarrollo (envejecimiento de los tejidos) y de la acción del hombre (modificación del medio ambiente o del propio árbol como las entresacas o las podas).

Para los estudios de biomasa se suelen dividir los árboles en partes o compartimientos homogéneos: la madera del tronco, la corteza, las ramas vivas, las ramas muertas, las hojas, las raíces gruesas y medianas, y por último las raíces finas. La biomasa es un volumen multiplicado por una densidad mientras que la mineralomasa es una biomasa multiplicada por una concentración de elementos minerales. El volumen, la densidad y la concentración evolucionan no sólo en función de los factores antes mencionados (véase, por ejemplo, la reseña de Chave *et al.*, 2009 sobre la densidad de la madera) sino también en el interior de los árboles: entre partes pero también en función de la posición radial (cerca de la médula o cerca de la corteza), y de la posición longitudinal (cerca del suelo o cerca de la copa), incluso, por ejemplo, para las concentraciones en elementos minerales: Andrews & Siccama (1995); Colin-Belgrand *et al.* (1996); Saint-André *et al.* (2002b); Augusto *et al.* (2008); o para la densidad de la madera: Guilley *et al.* (2004); Bergès *et al.* (2008); Henry *et al.* (2010); Knapic *et al.* (2011). Todo esto tiene consecuencias en las ecuaciones de biomasa y mineralomasa y este Capítulo tiene por objeto recordar algunas nociones importantes en silvicultura que permitirán luego pensar en los modelos de este manual en términos "biológicos" (¿cuáles son los factores de variación potenciales?) y no en términos puramente estadísticos (¿cuál es la mejor ecuación posible, independientemente de su grado de verosimilitud con respecto a los procesos biológicos?). La combinación de ambos objetivos es, al final de cuentas, lo que pretende lograr el presente manual.

1.1.1. Caso de las masas homogéneas y monoespecíficas

Este tipo de rodal se caracteriza por una relativa homogeneidad de la población arbórea: los árboles son de la misma edad y mayoritariamente de la misma especie. El crecimiento de estos rodales fue estudiado muy ampliamente (de Perthuis, 1788 *in* Batho & García, 2006) y los principios descritos a continuación tienen un aplicación prácticamente universal (Assmann, 1970; Dhôte, 1991; Skovsgaard & Vanclay, 2008; Pretzsch, 2009; García, 2011). Se suele distinguir el rodal de su conjunto del árbol dentro de él. Esta distinción permite

disociar los diferentes factores que intervienen en el crecimiento de los árboles: fertilidad del lugar, presión global en el seno del rodal y clasificación sociológica. La fertilidad del lugar en su sentido amplio comprende la capacidad del suelo de alimentar a los árboles (en nutrientes y en agua) así como el clima general de la zona (iluminación, temperatura y pluviometría medias, recurrencia habitual de períodos de heladas o de sequía, etc.). La presión entre los árboles en el seno del rodal se mide con diferentes índices de densidad del mismo. Por último, la clasificación social de cada individuo define su capacidad de movilizar los recursos en su entorno próximo.

Foto 1.1 – *Plantación de eucaliptos en el Congo. Arriba, zona de Kissoko, ejemplo de los mosaicos de sabana y plantaciones. Abajo, zona de Kondi en curso de explotación que muestra las principales salidas para la madera de eucalipto (trozas para pasta de papel y producción de carbón vegetal para la ciudad de Pointe-Noire) (Fotos: L. Saint-André).*

Crecimiento del rodal

La noción de producción en silvicultura comprende el volumen (o la biomasa) en pie así como todo lo que se ha retirado del rodal a lo largo de su vida (por mortalidad o por raleo). En general esta noción de producción tal como figura en los modelos de producción o en la mayoría de los modelos de crecimiento con base dendrométrica, no incluyen la hojarasca (hojas, ramas, corteza) ni el ciclaje de las raíces. En cambio en los modelos con base ecofisiológica o en los estudios cuyo objeto son los balances de carbono y de elementos minerales en los rodales, la producción incluye también esta renovación de los órganos. Más adelante en este Capítulo consideramos la producción en su acepción restringida.

La producción de una masa homogénea y monoespecífica, para una especie dada, en una región dada y en una amplia gama de la silvicultura (siempre y cuando el dosel esté cerrado), está totalmente determinada por su altura media. A este postulado se le conoce con el nombre de ley de Eichhorn (1904), o ley de Eichhorn ampliada, cuando considera la altura dominante en vez de la altura media (Decourt, 1973). Significa que la fertilidad de los diferentes sitios de una misma región sólo modifica la velocidad de crecimiento en altura del rodal sin modificar la relación existente entre productividad y altura media. Aun cuando

se la ponga en tela de juicio (véase Assmann, 1970), está claro que la altura de los árboles dominantes (H_0) constituye el principal motor de la mayoría de los modelos de crecimiento con base dendrométrica (por ejemplo Dhôte, 1996; García, 2003; Saint-André *et al.*, 2008; Skovsgaard & Vanclay, 2008; Weiskittel *et al.*, 2009; García, 2011). Alder (1980 *in* Pardé & Bouchon, 1988) resume el principio en la frase siguiente: "la relación altura / edad / índice de fertilidad constituye el elemento fundamental para predecir el aumento de rodales homogéneas. Se lo suele expresar como un grupo de curvas de fertilidad". El hecho de que el crecimiento en altura dominante sólo dependa de la fertilidad del sitio (en su sentido amplio, lo que en inglés se denomina "site index") y de la edad de los rodales es válido, en una primera aproximación, en la mayoría de los ecosistemas monoespecíficos y homogéneos templados o tropicales. Esto se debe a dos factores principales: los árboles dominantes, por su posición, son menos sensibles a la competencia que los árboles suprimidos y, además, el crecimiento en altura también es menos sensible a la silvicultura (salvo en el caso de un régimen de raleo especial) que el crecimiento en diámetro de los árboles. Esto implica que el crecimiento en altura de los árboles dominantes refleja mucho mejor la fertilidad del sitio que el crecimiento medio en altura o en diámetro. Para llegar a la ley de Eichhorn hace falta luego combinar el aumento del área basal (o del volumen) con el aumento de la altura dominante. Esta relación también es estable para una especie y una región dadas, bajo una amplia gama de tipos de silvicultura (para comenzar, cuando el dosel es lo suficientemente denso). Dhôte (1996) da un ejemplo para las hayas en Francia.

No obstante hay varios ejemplos donde la relación estricta $H_0 = f(\text{edad y fertilidad})$ no se cumple: el pino laricio en el centro de Francia (Meredieu *et al.*, 2003) y el eucalipto del Congo (Saint-André *et al.*, 2002a). En ambos casos, el crecimiento en altura dominante también es función de la densidad del rodal. La hipótesis subyacente está asociada a la escasa fertilidad de los suelos que conllevaría una fuerte competencia por el acceso a los recursos hídricos y minerales, incluso en los árboles dominantes. Desde hace algunos años, con respecto a esta relación y a aquella que asocia el crecimiento en área basal al aumento en altura dominante, se ha puesto claramente de manifiesto un efecto "fecha" debido a los cambios globales (véanse, por ejemplo Bontemps *et al.*, 2009, 2011; Charru *et al.*, 2010).

En resumen, aunque se la ponga en tela de juicio y no sea necesariamente tan invariable como se esperaba, esta primera "ley" es importante porque permite introducir luego, en los modelos "parametrizados" de biomasa, la noción de fertilidad por medio de la edad y la altura dominante de los rodales inventariados (y también la densidad) para aumentar el carácter genérico de las ecuaciones elaboradas.

Crecimiento de los árboles en el rodal

Cuando se calcula el crecimiento del volumen o de la biomasa de todo el rodal, hay que repartirlo luego entre los distintos árboles. Las relaciones utilizadas para el crecimiento en diámetro individual suelen ser del tipo potencial × reductor, donde el potencial lo da el crecimiento en área basal y / o altura dominante y los reductores son función (*i*) de un índice de densidad y (*ii*) de la clasificación social del árbol. Un índice de densidad puede ser simplemente la densidad del rodal pero los investigadores elaboraron otros índices como el espaciamiento de Hart-Becking, basado en el crecimiento de los árboles fuera del rodal (crecimiento libre) o el IDR ("índice de densidad de Reinecke"), basado en la ley de auto raleo (crecimiento de los árboles en rodales excesivamente densos). Ambos presentan la ventaja de depender menos de la edad del rodal que de la propia densidad (véase Shaw, 2006 o, en líneas más generales, la reseña bibliográfica de Vanclay, 2009). La posición social de los árboles se expresa generalmente por las relaciones de tipo H/H_0 o D/D_0 (donde D

es el diámetro del árbol, H su altura y D_0 el diámetro dominante del rodal) pero también pueden usarse otras relaciones. Por ejemplo, Dhôte (1990), Saint-André *et al.* (2002a) y más recientemente Cavaignac *et al.* (2012) utilizan un modelo lineal segmentado para traducir el crecimiento diamétrico de los árboles: por debajo de cierto umbral de circunferencia, los árboles están totalmente por debajo del dosel y no crecen más; más allá de él, el crecimiento en área basal es una función lineal de la circunferencia de los árboles. Esta relación refleja bien el hecho de que los árboles dominantes crecen más que los suprimidos. El umbral y la pendiente de la relación evolucionan en función de la edad de los rodales y los tratamientos silviculturales (raleos). El crecimiento en altura puede estimarse también por medio de las relaciones de tipo potencial × reductor pero, en general, los modeladores usan relaciones de altura–circunferencia (Soares & Tomé, 2002). Estas relaciones están saturadas (la asíntota es igual a la altura dominante del rodal) y curvilineales. Los parámetros de esta relación evolucionan también en función de la edad y de la silvicultura (Deleuze *et al.*, 1996).

Resumiendo, para estas otras dos relaciones que dan la dimensión de cada árbol dentro del rodal, hay que recordar lo siguiente: los índices de densidad y de competencia (clasificación social) que determinan en gran medida el crecimiento individual de los árboles dentro del rodal son los factores que se pueden integrar también a los modelos de biomasa. Las variables interesantes desde este punto de vista pueden ser: la densidad del rodal, el espaciamiento de Hart-Becking, el IDR, y luego, a escala individual: el coeficiente de rectitud (H/D), la robustez del árbol ($D^{1/2}/H$ — Vallet *et al.*, 2006; Gomat *et al.*, 2011), o su clasificación sociológica (H/H_0 o D/D_0).

Distribución de la biomasa en el árbol

Por último, una vez distribuida la biomasa del rodal entre los árboles, para cada uno hay que asignarla a cada compartimiento y repartirla a los largo de los ejes. Para el tronco, la relación que se suele utilizar es la ley de Pressler (o, para los especialistas en ecofisiología, su equivalente dado por el "pipe-model" de Shinozaki *et al.*, 1964a,b): (*i*) la superficie transversal de los anillos aumenta en forma lineal desde lo alto del árbol hasta la base funcional de la copa; (*ii*) luego se mantiene constante desde la base de la copa hasta la base del árbol. Por lo tanto, a medida que el árbol crece, el tronco se volverá cada vez más cilíndrico puesto que la distancia entre los anillos será mayor cerca de la copa que en la base. Esta ley de Pressler no expresa una distribución promedio de la madera en el árbol (Saint-André *et al.*, 1999). En efecto, para los árboles dominantes, la superficie del anillo puede seguir aumentando por debajo de la copa y para los dominados/suprimidos, puede disminuir mucho. En casos extremos, también es posible que el anillo no esté completo en la base del árbol, incluso puede faltar, como por ejemplo en las hayas (Nicolini *et al.*, 2001). Además, cualquier acción sobre la copa (densidades importantes o escasas, raleos, podas o entresacas) tendrá consecuencias en el apilado de los anillos y, en consecuencia, sobre la forma del tronco (véanse la reseña de Larson, 1963, o los ejemplos dados por Valinger, 1992; Ikonen *et al.*, 2006). La densidad de la madera también es diferente en la parte alta y la parte baja del árbol (madera joven cerca de la copa y mayor proporción de madera madura en la parte inferior — Burdon *et al.*, 2004) pero también variará según las condiciones de crecimiento de los árboles (mediante los cambios de proporción entre la madera tardía y aquella temprana, o los cambios de estructura y las propiedades celulares, véanse Guilley *et al.*, 2004; Bouriaud *et al.*, 2005; Bergès *et al.*, 2008 para mencionar algunas publicaciones recientes). En consecuencia, la biomasa será diferente o no para troncos de dimensiones iguales (altura, diámetro, edad), en función de las condiciones de crecimiento de los árboles. Es posible que, por ejemplo, un aumento del volumen se acompañe de una baja de densidad

(es el esquema clásico para las resinosas) y no resulte por tanto en grandes diferencias en la biomasa de los troncos. Para las ramas y las hojas, la biomasa dependerá mucho de la arquitectura de los árboles y, por ende, de la densidad del rodal: a dimensiones iguales (altura, diámetro y edad), los árboles que hayan crecido en rodales abiertos tendrán más ramas y hojas que aquellos que hayan crecido en rodales densos. Lo que se procura con las investigaciones actuales sobre la biomasa es determinar la parte asociada al desarrollo intrínseco del árbol (ontogenia) distinguiéndola de aquella asociada a factores ambientales (Thornley, 1972; Bloom *et al.*, 1985; West *et al.*, 1999; McCarthy & Enquist, 2007; Savage *et al.*, 2008; Genet *et al.*, 2011; Gourlet-Fleury *et al.*, 2011). Con respecto a las raíces, su biomasa depende del bioma, de la biomasa sobre el suelo, de la etapa de desarrollo y de las condiciones de crecimiento (cf. por ejemplo Jackson *et al.*, 1996; Cairns *et al.*, 1997; Tateno *et al.*, 2004; Mokany *et al.*, 2006).

De estas últimas nociones cabe destacar que las condiciones de crecimiento no sólo influirán en la cantidad global de biomasa producida sino también en su distribución dentro de los árboles (proporción sobre el suelo/subterránea; apilado de anillos, etc.). Por lo tanto será absolutamente necesario tener en cuenta estas variaciones posibles en el muestreo (en especial para el troceo de troncos y la toma de las diferentes alícuotas) pero también en la elaboración de las ecuaciones de forma que reflejen correctamente las distintas relaciones de biomasa (sobre el suelo/subterránea; tronco/ramas; hojas/raíces finas) en función de las condiciones de crecimiento.

1.1.2. Caso de masas homogéneas y/o pluriespecíficas

Las nociones descritas anteriormente siguen siendo válidas también para las masas pluriespecíficas y homogéneas pero resulta difícil integrarlas a una ecuación y la mayoría de las veces es imposible con la forma anterior (Peng, 2000). Por ejemplo, la noción de altura dominante es difícil de cuantificar para los rodales heterogéneos y / o pluriespecíficos (¿cabe establecer una altura dominante para todas las especies? ¿o bien una para cada especie?). Del mismo modo, cabe plantearse qué significa el área basal para una masa demasiado heterogénea como las que hay en el bosque tropical húmedo. Por último, ¿cómo tener en cuenta el hecho de que la edad de los árboles suele ser desconocida (Tomé *et al.*, 2006)? Los modelos de crecimiento elaborados para estos rodales desglosan pues con menor detalle las diferentes escalas (producción de biomasa a escala del rodal, distribución entre los árboles y también dentro de ellos) que aquellos usados en las masas uniformes. Se pueden distinguir tres tipos de modelos: (1) los modelos matriciales de rodales; (2) los modelos centrados en los individuos que, en general, dependen de las distancias entre los árboles; (3) los modelos de corta por grupos (véanse las diferentes reseñas realizadas por Vanclay, 1994; Franc *et al.*, 2000; Porté & Bartelink, 2002). Los modelos de tipo matricial reúnen a los árboles por grupos funcionales (grupos con una estrategia de crecimiento común) y por clases de dimensión homogéneas (en general, el diámetro) y aplican un sistema de matrices que incluyen el relutamiento, la mortalidad y el paso de individuos de un grupo a otro (véanse, por ejemplo, Eyre & Zillgitt, 1950; Favrichon, 1998; Namaalwa *et al.*, 2005; Picard *et al.*, 2008). Para los modelos centrados en los individuos, se suele cartografiar la población de árboles y el crecimiento de un árbol depende de sus vecinos (véanse, por ejemplo,Gourlet-Fleury & Houllier, 2000 para un modelo en bosque tropical, o Courbaud *et al.*, 2001 para un modelo en bosque templado). Pero, así como con los modelos elaborados para las masas uniformes, también hay modelos centrados en los individuos que son independientes de las distancias (por ejemplo Calama *et al.*, 2008; Pukkala *et al.*, 2009; Vallet & Pérot, 2011; Dreyfus, 2012) e incluso modelos intermedios (véanse Picard & Franc, 2001; Verzelen *et al.*,

2006; Perot *et al.*, 2010). Por último, en los modelos de corta por grupos, se representa el bosque con un conjunto de células en diferentes etapas del ciclo silvogenético. La mortalidad y la incorporación de nuevos árboles inventariables se simula en forma estocástica mientras que el crecimiento de los árboles sigue las mismas leyes que las de los modelos centrados en individuos e independientes de las distancias (véase una reseña en Porté & Bartelink, 2002).

El hecho de que estos rodales sean más complicados cuando se trata de traducirlos en ecuaciones no contradice los principios evocados anteriormente para la elaboración de los modelos de volumen, de biomasa o de mineralomasa: (*i*) introducir la fertilidad para ampliar la zona de validez de los modelos de biomasa; (*ii*) utilizar los índices de densidad para tomar en cuenta el grado de competencia entre los árboles; (*iii*) tener en cuenta la clasificación social además de las características básicas de los árboles (altura, diámetro).

Además de las dificultades asociadas a la elaboración de las ecuaciones concebidas para los bosques monoespecíficos, la estimación de la biomasa en bosques pluriespecíficos se enfrenta a dificultadas adicionales: la preparación de un muestreo adecuado (¿cuáles especies? ¿cómo reunirlas en grupos funcionales?) y el acceso al terreno (sobre todo en la zona tropical donde estos rodales suelen encontrarse en áreas protegidas donde la corta de árboles está muy reglamentada, incluso prohibida para ciertas especies).

Foto 1.2 – *Rodales heterogéneos. Izquierda, caso de rodales pluriespecíficos en el Monte Saint-Anne en Quebec; derecha, rodales plurispecíficos y multietáneos en Costa Rica (Foto: B. Locatelli).*

1.2. Elección del método

1.2.1. Estimación de la biomasa de una bioma

No existe "un" método unico para estimar una reserva de biomasa sino varios, según la escala considerada (Gibbs *et al.*, 2007). A escala nacional y más allá de ella, suelen utilizarse valores medios por bioma (FAO, 2006): la cantidad de biomasa se estima multiplicando la superficie de cada bioma por la cantidad de biomasa media por unidad de superficie para dicho bioma. Las cantidades medias por bioma son las estimadas a partir de medidas tomadas a una escala más restringida. La teledetección permite estimar la biomasa de la escala nacional a la escala del paisaje. Ya se trate de sensores ópticos satelitales (Landsat, MODIS),

de imágenes satelitales de alta resolución (Ikonos, QuickBird) o no (fotografías aéreas), sensores de radar o microondas satelitales (ERS, JERS, Envisat, PALSAR), o sensores de láser (Lidar), todos estos métodos parten del supuesto de que se dispone de las medidas de campo para ajustar las relaciones que predicen la biomasa en función de las observaciones hechas por los sensores. En el caso de los sensores ópticos satelitales, se necesitan datos de campo para calibrar la relación entre la biomasa y los índices de vegetación obtenidos por satélite (NDVI, NDFI, AVI, GVI, etc.) (Dong *et al.*, 2003; Saatchi *et al.*, 2007). Las imágenes de alta resolución y las fotografías aéreas aportan informaciones sobre el tamaño de las copas y la altura de los árboles. A continuación se necesitan datos de campo para vincular estas informaciones a la biomasa (por ejemplo Bradley, 1988; Holmgren *et al.*, 1994; St.-Onge *et al.*, 2008; Gonzalez *et al.*, 2010). Lo mismo vale para las informaciones sobre la estructura vertical del bosque, aportadas por el Lidar, o por las informaciones sobre la distribución vertical del agua contenida en la vegetación, suministradas por el radar o las microondas (por ejemplo Lefsky *et al.*, 2002; Patenaude *et al.*, 2004). No obstante, los métodos de teledetección siguen siendo limitados en cuanto a la precisión de las mediciones de biomasa (especialmente las superficies) y la diferenciación de los tipos de bosques en función de los medios técnicos y financieros, los recursos humanos disponibles, la nubosidad y el riesgo de saturación de las señales utilizadas para ciertos tipos de vegetación.

De esta forma los métodos de estimación de la biomasa a escala del paisaje y más allá se basan en las mediciones de campo, tomadas a una escala comprendida entre el paisaje y la parcela. En este tipo de escala, las estimaciones de la biomasa se basan en los datos del inventario forestal: inventario de una muestra de árboles si la superficie es grande, o un censo completo en caso contrario (en particular en las parcelas permanentes de algunas hectáreas). Por debajo de esta escala las mediciones individuales de biomasa pueden obtenerse pesando árboles y la vegetación del sotobosque).

1.2.2. Estimación de la biomasa de un bosque o de un conjunto de bosques

Las estimaciones de biomasa o mineralomasa forestales basadas en los inventarios forestales exigen que se disponga de

1. un inventario exhaustivo o estadístico de los árboles presentes;
2. modelos para evaluar las reservas a partir de las dimensiones de los individuos medidos;
3. una evaluación de la biomasa contenida en la necromasa (madera muerta en pie) y en la vegetación de sotobosque.

En el presente manual nos concentramos en el segundo aspecto sabiendo que la parte del inventario o la evaluación cuantitativa de la parte bajo cubierta no son necesariamente fáciles de realizar, en particular en bosques altamente heterogéneos.

A partir de los inventarios se pueden usar dos grandes opciones para estimar las reservas de carbono o de elementos minerales en los árboles (MacDicken, 1997; Hairiah *et al.*, 2001; AGO, 2002; Ponce-Hernandez *et al.*, 2004; Monreal *et al.*, 2005; Pearson & Brown, 2005; Dietz & Kuyah, 2011): (1) uso de modelos de biomasa/mineralomasa: esta solución suele adoptarse porque permite establecer rápidamente balances de carbono o de elementos minerales dentro de una parcela en un momento dado. En general se consideran todas las partes del ecosistema (sobre el suelo, subterráneas, hojarasca en el suelo, etc.). Se cortan árboles específicamente para estas operaciones. La definición de compartimientos (trozos cortados) pueden variar según la aplicación y el ámbito de interés (véase el Capítulo 3). (2) El uso de modelos para estimar sucesivamente el volumen de los árboles, la densidad de la

madera y el contenido de nutrimentos. La ventaja de este método es que disocia las distintas partes, permitiendo analizar la influencia de la edad y las condiciones de crecimiento independientemente sobre uno u otro componente. En general sólo el tronco puede usarse para una modelización detallada (entre y dentro de los anillos). La biomasa de los otros compartimientos se estima mediante coeficientes de expansión volumétrica, valores de densidad media de la madera y el contenido de nutrimentos. En todos los casos estos métodos utilizan ampliamente un gran modelo tipo que reagrupa indiferentemente los "modelos de cubicación, modelos de biomasa, modelos de mineralomasa, etc." y que es objeto del presente manual.

Los modelos de biomasa o de mineralomasa se parecen mucho a los de volumen, modelos que se vienen estudiando comúnmente desde hace casi dos siglos. Los primeros modelos para las hayas (*Fagus sylvatica*) fueron publicados por Cotta en 1804 (*in* Bouchon, 1974). El principio es vincular una magnitud difícil de medir (como el volumen del árbol, su masa o su contenido de nutrimentos) a magnitudes más fáciles de determinar como el diámetro a 1,30 m o la altura del árbol. Si se utilizan ambas características, se habla un modelo de dos entradas; si sólo se utiliza el diámetro, se habla entonces un modelo de una entrada. En general las correlaciones son buenas y las funciones más usadas son de tipo polinómico, logarítmico o de potencia. Para conocer más detalles al respecto, se pueden consultar las reseñas propuestas por Bouchon (1974); Hitchcock & McDonnell (1979); Pardé (1980); Cailliez (1980); Pardé & Bouchon (1988), y más recientemente por Parresol (1999, 2001).

Estas funciones son relativamente simples pero representan tres dificultades mayores. Primero, son bastante poco genéricas: si se cambia de especie o si uno se aleja del ámbito de calibración, hay que utilizar las ecuaciones con precaución. El Capítulo sobre el muestreo da algunas explicaciones sobre cómo mitigar este problema. El principio fundamental es cubrir al máximo la variabilidad de las cantidades estudiadas.

El segundo obstáculo de estas funciones reside en el carácter mismo de los datos que se tratan (volúmenes, masas, mineralomasas). En particular, pueden presentarse problemas de heterocedasticidad (es decir, varianza no homogénea de las biomasas en función del regresor). Esto tiene poca influencia sobre el valor de los parámetros estimados: cuanto mayor sea el número de árboles del muestreo, más rápida será la convergencia hacia los verdaderos parámetros (Kelly & Beltz, 1987). No obstante, todo lo que tiene que ver con el intervalo de confianza de las estimaciones se ve afectado por lo siguiente:

1. la varianza de los parámetros estimados no es mínima;
2. esta varianza tiene sesgo; y
3. la varianza residual está mal estimada (Cunia, 1964; Parresol, 1993; Gregoire & Dyer, 1989).

No corregir estos problemas de hetereocedasticidad tiene pocas consecuencias sobre la estimación del valor medio de la biomasa o del volumen. Por el contrario es absolutamente necesario hacer la corrección para obtener los intervalos de confianza correctos para las predicciones. Para corregir esos problemas de heterocedasticidad, se suelen presentar dos métodos: el primero consiste en efectuar una ponderación (por ejemplo, por la inversa del diámetro o del diámetro al cuadrado) pero todo reside en la función de ponderación y en particular en la potencia que se aplicará; el segundo consiste en tomar el logaritmo de los términos de la ecuación pero, en este caso, hace falta corregir los valores simulados para encontrar una distribución normal de los valores estimados (Duan, 1983; Taylor, 1986). Además, es frecuente que la transformación logarítmica no resulte en a un modelo lineal (Návar *et al.*, 2002; Saint-André *et al.*, 2005).

La tercera dificultad está asociada a la aditividad de las ecuaciones. Las mediciones de biomasa y luego los ajustes de las funciones se suelen hacer compartimiento por compartimiento. La aditividad de las relaciones no es inmediata y una propiedad deseable del sistema de ecuaciones es que la suma de las predicciones de biomasa compartimiento por compartimiento sea igual a la predicción de la biomasa total del árbol (véanse Kozak, 1970; Reed & Green, 1985; Návar *et al.*, 2002). En general se proponen tres soluciones (Parresol, 1999):

1. la biomasa total se calcula sumando las biomasas compartimiento por compartimiento, y la varianza de esta estimación utiliza las varianzas calculadas para cada compartimiento y las covarianzas calculadas de dos en dos;

2. se garantiza la aditividad al usar los mismos regresores y los mismos pesos para todas las funciones, siendo los parámetros de la función de biomasa total la suma de los parámetros obtenidos para cada compartimiento;

3. los modelos son diferentes compartimientos por compartimiento pero se ajustan conjuntamente y la aditividad se obtiene mediante las restricciones sobre los parámetros.

Cada método tiene sus ventajas y desventajas. En el marco del presente manual ajustaremos un modelo para cada compartimiento y un modelo para la biomasa total verificando que se respete la aditividad. A lo largo de todo el manual se utilizará un ejemplo concreto "Línea roja") para ilustrar los casos. Se trata de un conjunto de datos obtenido en un experimento realizado en Ghana, en un bosque tropical húmedo natural (Henry *et al.*, 2010).

1.2.3. Medición de la biomasa de un árbol

Los modelos de biomasa vinculan la medición individual de la biomasa y la estimación de la misma en el campo a partir de los datos del inventario. Por tanto, pesar los árboles para medir la biomasa forma parte fundamental del proceso de elaboración de ecuaciones alométricas y a ellas se dedica una parte importante de este manual. Aun cuando los principios generales presentados en el Capítulo 3 (segmentación del árbol en compartimientos con una densidad de materia seca homogénea, medición de las tasas entre materia seca con respecto al volumen fresco para alícuotas y aplicación de la regla de tres) deberían permitir medir la biomasa de cualquier tipo de especie arbórea, en el presente manual no se abordarán todos los casos específicos. Las plantas que no son árboles pero que pueden alcanzar la altura de uno (bambúes, ratán, palmeras, helechos arborescentes, plantas musáceas, *Pandanus sp.*, etc.) constituyen excepciones.

Las plantas que usan los árboles como apoyo para crecer (epifitas, plantas parásitas, plantas rastreras, etc.) son otro caso aparte (Putz, 1983; Gerwing & Farias, 2000; Gerwing *et al.*, 2006; Gehring *et al.*, 2004; Schnitzer *et al.*, 2006, 2008). Su biomasa debería disociarse de la de su hospedero.

Por último, los árboles huecos, aquellos cuyo tronco tiene una forma muy diferente a un cilindro (como *Swartzia polyphylla* DC.), amates (*Ficus spp.*), etc., constituyen las excepciones para las cuáles no podrán usarse los modelos de biomasa sin un ajuste específico (Nogueira *et al.*, 2006).

2

Muestreo y estratificación

El muestreo consiste en predecir las características de un conjunto a partir de una parte (muestra) del mismo. Por ejemplo, se quiere estimar el volumen de madera de un bosque pero no se pueden cubicar todos los árboles uno por uno así que se va a realizar la cubicación de una muestra de árboles del bosque y luego se va a extrapolar la estimación obtenida a todo el bosque (CTFT, 1989, p.252). Como el volumen sólo se mide en una muestra y no en el conjunto de árboles del rodal, la estimación del volumen total así obtenida contiene un *error de muestreo*[1]. El muestreo, en su sentido estricto, consiste en:

1. elegir lo mejor posible los árboles que formarán parte de la muestra que se medirá (se habla más bien de *plan de muestreo*),

2. elegir un método de cálculo (se suele hablar de *estimador*) del volumen total a partir de las mediciones,

para reducir al mínimo el error de muestreo.

En la teoría de muestreo clásica, los volúmenes de N árboles del rodal son datos fijos: la única fuente de variación de las estimaciones es el muestreo, de forma tal que un muestreo exhaustivo dará siempre la misma estimación. Aquí utilizaremos el enfoque llamado de "super-población" que surgió en el decenio de 1970 (Cochran, 1977). Consiste en considerar que los volúmenes de N árboles que componen el rodal son variables aleatorias, de modo que el rodal observado no es más que uno entre otros, sacado de una super-población. Este enfoque permite liberarse de ciertas aproximaciones y definir un plan de muestreo óptimo (lo que no suele ser posible con el enfoque clásico) pero tiene el inconveniente de llevar a malas soluciones si el modelo de super-población adoptada no se ajusta a la realidad.

La elección de un método de muestreo depende del objetivo fijado. En principio hay que empezar por preguntarse para qué van a servir los modelos de volumen o de biomasa que nos proponemos elaborar. ¿Queremos predecir las características de un árbol particular cuyas variables de entrada conocemos? ¿Se trata de predecir las características del árbol promedio para los valores dados de las variables de entrada? ¿Se trata de predecir el volumen total del rodal donde proceden los árboles usados para elaborar el modelo o el volumen total de

[1] Ponemos en itálica la jerga de la teoría de muestreo; en el Anexo 2 de Bellefontaine *et al.* (2001) figura una definición en francés de dichos términos.

otro rodal? En estos dos últimos casos, ¿las variables de entrada de el modelo se miden sobre todos los árboles del rodal o bien nuevamente sobre una muestra de árboles? Etc. De este modo se puede construir una cadena que vaya del rodal estudiado a la magnitud que se trata de predecir (Figura 2.1).

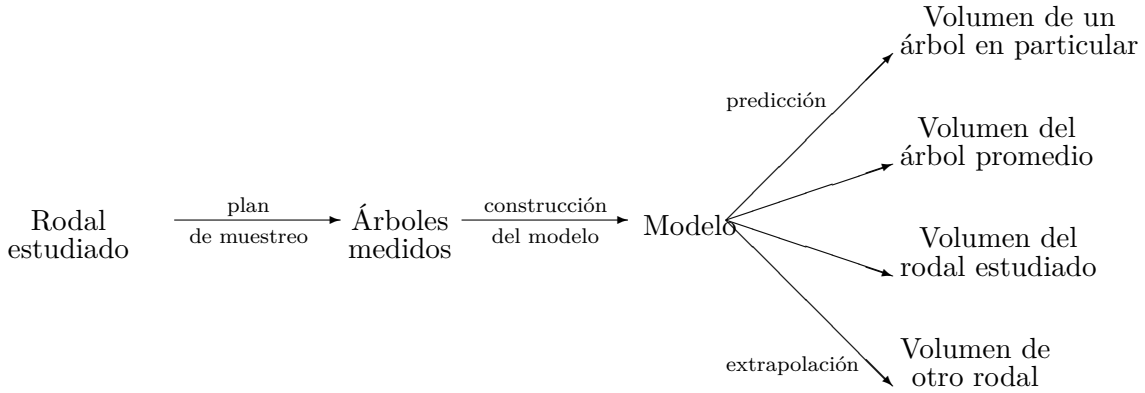


Figura 2.1 – *Cadena que va del rodal estudiado a las magnitudes que se desean predecir.*

Remontando esta cadena de atrás para adelante, la precisión con respecto a la magnitud predicha depende de la precisión de los parámetros del modelo, la que a su vez depende del plan de muestreo (número y elección de los árboles medidos) y de la variabilidad dentro del rodal estudiado (Cunia, 1987b). Se puede determinar un nivel de precisión que se desea alcanzar en las predicciones, lo que, por efecto retroactivo, para un tipo de modelo y un tipo de muestreo dados, implica el número mínimo de árboles que habrá que medir. Asimismo se puede seguir un proceso de optimización para determinar, para una precisión dada que se pretende alcanzar y un tipo de modelo dado, el método de muestreo que minimiza el tipo (o el costo) de las mediciones (Cunia, 1987c,d). En ciertos casos el costo de las mediciones es el factor limitante. Es justamente el caso para medir la biomasa de los sistemas de raíces. En este caso no se trata tanto de alcanzar una precisión dada para las predicciones sino en mantenerse dentro de los límites razonables de costo. De este modo se puede buscar, en función del costo de una medición dada y para un tipo de modelo dado, el método de muestreo que maximiza la precisión de las estimaciones.

Este razonamiento suele ser demasiado complejo para aplicarlo rigurosamente, debe hacerse caso por caso ya que depende (*i*) de lo que se trata de predecir, (*ii*) del tipo de modelos utilizados y (*iii*) del tipo de muestreo adoptado. El hecho de utilizar un modelo ya implica, de por sí, una restriccíon sobre el método de muestreo: el volumen total de un rodal podría estimarse a partir de la cubicación de una muestra de árboles, sin pasar por el modelo de volumen. Al usar un modelo de volumen para estimar el volumen total del rodal, ya nos hemos restringido a un tipo de *estimador* del volumen total.

Más aún, el razonamiento que permite determinar el plan de muestreo en función de la precisión que se desea obtener en las predicciones implica conocer la relación entre la precisión de las predicciones y la precisión de los parámetros del modelo, la relación entre los parámetros usados para construir el modelo y el tamaño de la muestra, etc. En algunos casos simples estas relaciones se conocen explícitamente. Pero, con mucha frecuencia, en cuanto el modelo adquiere una forma un poco complicada, esas relaciones ya no son explícitas. Ya no es posible utilizar simplemente dicho razonamiento.

El golpe de gracia a este razonamiento lo da el darse cuenta de que: (*i*) la finalidad del modelo suele ser múltiple, incluso imprecisa y (*ii*) la forma del modelo no suele conocerse de antemano. En efecto, con mucha frecuencia deseamos poder usar un modelo para distintos fines: para evaluar el volumen de un árbol en particular, de un árbol promedio, de todo

un rodal, etc. La construcción del modelo se convierte en un fin por sí sola, que no guarda relación con una cantidad que debe predecirse. Además, la elección sobre la forma del modelo suele resultar de un análisis exploratorio de los datos que, por ende, no se conoce de antemano. Es cierto que algunas relaciones, como la función de potencia o los polinomios de grado 2, aparecen con frecuencia pero, *a priori*, no es posible determinar una regla. Por lo tanto es inútil intentar optimizar un plan de muestreo.

Al final de cuentas, el muestreo utilizado para construir tablas de volumen o biomasa se basa generalmente en consideraciones empíricas relacionandas al plan de muestreo. La elección del *estimador*, que corresponde en realidad a la elección del modelo, es consecuencia de un razonamiento *a posteriori*, en función de los datos acopiados e independientemente del plan de muestreo.

2.1. Muestreo para una regresión lineal simple

Comencemos con un ejemplo sencillo que permitirá ilustrar las ideas presentadas anteriormente. Supongamos que los árboles del rodal se describen mediante su diámetro D, su altura H y su volumen V. Se usa un modelo de volumen para predecir el volumen V en función de la variable D^2H. El modelo de super-población que se adopta para describir el rodal implica que la relación entre V y D^2H es lineal con un error ε de varianza σ^2:

$$V = \alpha + \beta\, D^2H + \varepsilon \tag{2.1}$$

donde ε a una distribución normal de esperanza nula y de desviación estándar σ. Además, se supone que la cantidad D^2H está distribuida según una distribución normal de promedio μ y de desviación estándar τ. El error ε incorpora todos los factores que hacen que dos árboles del mismo diámetro y la misma altura no tengan obligatoriamente el mismo volumen. Los parámetros α y β son desconocidos. Para estimarlos, se van a medir n árboles; se obtiene así una muestra de n dobletes $(D_1^2H_1,\, V_1), \ldots, (D_n^2H_n,\, V_n)$, luego se hace la regresión lineal siguiente:

$$V_i = a + b\, D_i^2H_i + \varepsilon_i \tag{2.2}$$

En la jerga de la teoría del muestreo, las variables de entrada del modelo (diámetro, altura, etc.) se llaman variables *auxiliares*. Hay que distinguir bien estas variables, que son relativas al árbol, de las otras, como la edad, que son relativas al rodal. Estas últimas son consideradas como parámetros (Pardé & Bouchon, 1988, p.106). Además, la *unidad* de muestreo es el árbol. Veamos ahora cómo definir el plan de muestreo en función del objetivo fijado.

2.1.1. Predicción del volumen de un árbol en particular

Supongamos que el objetivo sea predecir el volumen de un árbol del rodal de diámetro D^* y de altura H^*. El volumen predicho es:

$$V^* = a + b\, D^{*2}H^*$$

El modelo de super-población estipula que, debido al error ε, dos árboles tomados al azar y con el mismo diámetro D y la misma altura H no tienen forzosamente el mismo volumen. De ello resulta una variabilidad intrínseca cuando se mide un árbol en particular, que es igual a σ^2. A esta variabilidad intrínseca se agrega, para el error de predicción del volumen, la variabilidad debida a la imprecisión de las estimaciones de los parámetros α y β del modelo de volumen. Más adelante volveremos a estas nociones (en el Capítulo 7). De este

modo, para una regresión lineal, la semiamplitud del intervalo de confianza en el umbral α (típicamente 5 %) de V^* es igual a (Saporta, 1990, p.374):

$$t_{n-2}\,\hat{\sigma}\,\sqrt{1+\frac{1}{n}+\frac{(D^{*2}H^*-\overline{D^2H}_e)^2}{\sum_{i=1}^{n}(D_i^2H_i-\overline{D^2H}_e)^2}}$$

donde t_{n-2} es el cuantile $1-\alpha/2$ de la distribución de Student a $n-2$ grados de libertad, $\overline{D^2H}_e$ es la media empírica de los valores de D^2H medidos en la muestra:

$$\overline{D^2H}_e=\frac{1}{n}\sum_{i=1}^{n}D_i^2H_i$$

y $\hat{\sigma}$ es la estimación de la desviación estándar de los residuos:

$$\hat{\sigma}^2=\frac{1}{n-2}\sum_{i=1}^{n}[V_i-(a+b\,D_i^2H_i)]^2$$

El valor mínimo de esta semiamplitud (cuando $n\to\infty$) es $1{,}96\,\sigma$. Nos fijamos como objetivo de precisión de la estimación una desviación de $E\,\%$ con respecto a este mínimo incompresible, es decir que, aproximativamente, buscamos el tamaño de la muestra n tal que:

$$1+E\approx\sqrt{1+\frac{1}{n}+\frac{(D^{*2}H^*-\overline{D^2H}_e)^2}{\sum_{i=1}^{n}(D_i^2H_i-\overline{D^2H}_e)^2}} \tag{2.3}$$

Muestreo aleatorio

Ante todo veamos el caso en donde no se procura optimizar el plan de muestreo, por ejemplo, seleccionando al azar los árboles de la muestra. La media empírica de D^2H de la muestra es entonces una estimación de μ, mientras que la varianza empírica de D^2H de la muestra es una estimación de τ^2. Así pues:

$$(1+E)^2-1\approx\frac{1}{n}\left[1+\frac{(D^{*2}H^*-\mu)^2}{\tau^2}\right]$$

Como ejemplo numérico, tomemos $\mu=5$ m^3 para el valor medio de D^2H en el rodal entero y $\tau=1$ m^3 para su desviación estándar. Si queremos predecir el volumen de un árbol cuyo tamaño D^2H es igual a 2 m^3 con una diferencia de precisión de $E=5\,\%$, hace falta entonces medir aproximadamente $n=98$ árboles. Cabe señalar que la expresión de n en función de $D^{*2}H^*$ es simétrica alrededor de μ y pasa por un mínimo para $D^{*2}H^*=\mu$. Como $\mu-2=3$ m^3 y $\mu+3=8$ m^3, hacen falta $n=98$ árboles para predecir el volumen de un árbol cuyo tamaño D^2H es igual a 8 m^3 con una diferencia de precisión de 5 %. Se puede interpretar así el tamaño de la muestra $n=98$ como la que garantiza una diferencia de precisión de, por lo menos, 5 % (en el umbral $\alpha=5\,\%$) para toda predicción en el intervalo 2–8 m^3.

Muestreo optimizado

Veamos ahora el caso en que se procura optimizar el plan de muestreo en función del valor de $D^{*2}H^*$. La ecuación (2.3) muestra que la diferencia de precisión E es mínima cuando $\overline{D^2H}_e=D^{*2}H^*$. Así pues nos conviene escoger los árboles de la muestra de forma tal que la media empírica de su tamaño D^2H sea igual a $D^{*2}H^*$. En la práctica la media empírica

de los D^2H de la muestra no será jamás exactamente igual a $D^{*2}H^*$, así que también nos conviene maximizar el denominador $\sum_i (D_i^2 H_i - \overline{D^2 H_e})^2$, es decir, maximizar la varianza empírica de los valores de D^2H de la muestra. Al final de cuentas, el plan de muestreo que maximiza la precisión de la predicción del volumen de un árbol de D^2H igual a $D^{*2}H^*$ consiste en elegir $n/2$ árboles de D^2H igual a $D^{*2}H^* - \Delta$ y $n/2$ árboles de D^2H igual a $D^{*2}H^* + \Delta$, con Δ tan grande como sea posible (Figura 2.2). Este plan de muestreo permite

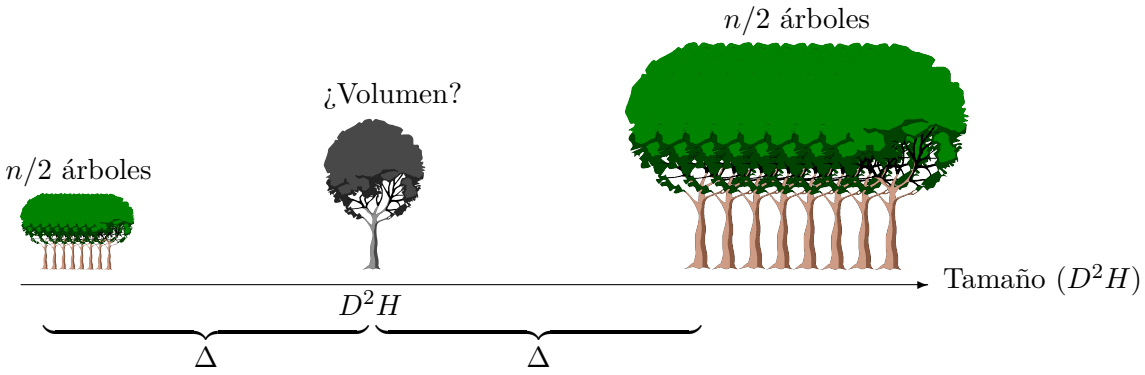


Figura 2.2 – *Plan de muestreo que optimiza la precisión de la predicción del volumen para un árbol en particular. La desviación del tamaño Δ debe ser también la mayor posible.*

omitir el término que depende de $D^{*2}H^*$ en (2.3), de modo que esta relación se simplifica en:

$$(1+E)^2 - 1 \approx \frac{1}{n}$$

Para $E = 5\,\%$, se obtiene entonces $n = 10$ árboles. La optimización del plan de muestreo permitió "ahorrar" 88 árboles en la medición con respecto al plan de muestreo que consistía en tomar árboles al azar. Sin embargo, el plan de muestreo optimizado depende de la estimación del volumen de un árbol de tamaño $D^{*2}H^*$. No está optimizado para estimar el volumen de un árbol de cualquier otro tamaño. Así vemos la limitaciones de este razonamiento porque un modelo de volumen no se elabora habitualmente, por no decir nunca, para predecir el volumen de un solo tamaño de árboles.

Más grave aún, el plan de muestreo optimizado suele depender del modelo de superpoblación y puede llevar a estimaciones erróneas si dicho modelo no corresponde a la realidad. La Figura 2.3 lo muestra bien. El plan de muestreo optimizado para un tamaño $D^{*2}H^*$ dado lleva a elegir puntos extremos (en negro en la Figura 2.3) para la muestra. Esta situación es crítica para una regresión lineal ya que el hecho de tener dos grupos de puntos alejados va a dar una R^2 elevada sin que se sepa lo que ocurre realmente entre las dos. Si la relación lineal supuesta para el modelo de sobrepoblación es exacto (Figura 2.3 izquierda), entonces no hay problemas: el volumen predicho para el modelo (representado por una estrella) será efectivamente próximo al volumen real (punto grisáceo). En cambio, si nos equivocamos para el modelo de sobrepoblación, entonces el volumen predicho será erróneo: es lo que se ve en la Figura 2.3 derecha (en la que los puntos de muestra en negro son exactamente los mismos que los de la Figura 2.3 izquierda), donde la relación tamaño-volumen es en realidad parabólica y no lineal. En la práctica, la forma de la relación tamaño-volumen (y, en consecuencia, del modelo) no se conoce de antemano y, por tanto, conviene hacer un muestreo de los árboles en todo el intervalo de variación del tamaño de forma tal que se vea el carácter de la relación tamaño-volumen.

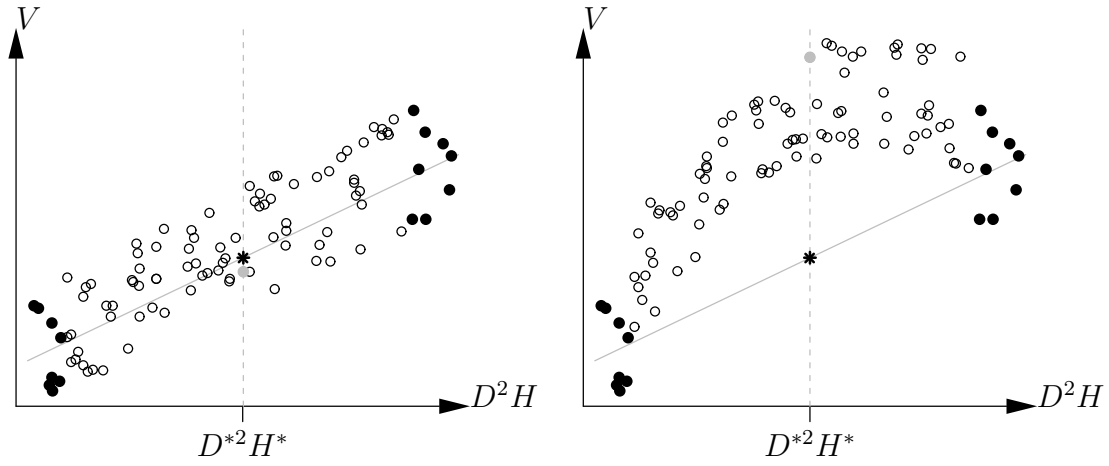


Figura 2.3 – *Predicción del volumen mediante una regresión lineal apoyándose en los puntos extremos (en negro) cuando la relación tamaño-volumen es efectivamente lineal (a la izquierda) y cuando no lo es (a la derecha). Los puntos negros son los mismos en ambos casos. La estrella indica el volumen predicho para la regresión lineal apoyándose en los puntos negros, mientras que el punto gris indica el volumen real correspondiente a* $D^{*2}H^*$.

2.1.2. Predicción del volumen del rodal

Supongamos ahora que el objetivo sea predecir el volumen de todo un rodal. Para hacerlo partimos primero del supuesto de que se miden el diámetro D y la altura H de *todos* los árboles del rodal, siendo N el número total de árboles del rodal (incluidos los n árboles de la muestra). Después de una nueva numeración de los árboles, se dispone de una medición del volumen V para $i = 1, \ldots, n$ y de una medición del tamaño D^2H para $i = 1, \ldots, N$. El estimador del volumen total del rodal deducido del modelo de volumen es:

$$V_{\text{tot}} = \sum_{i=1}^{N} (a + bD_i^2 H_i)$$

lo que también se puede escribir: $V_{\text{tot}} = N\bar{V}$, donde: $\bar{V} = a + b\overline{D^2H}$ representa el volumen medio de los árboles del rodal, y $\overline{D^2H} = (\sum_{i=1}^{N} D_i^2 H_i)/N$ es el diámetro promedio de los árboles del rodal. En la medida en que el modelo de volumen se obtiene por regresión lineal de $(V_1, \ldots, V_n)$ con relación a $(D_1^2H_1, \ldots, D_n^2H_n)$, los valores numéricos de los coeficientes a y b verifican que (Saporta, 1990, p.363): $\bar{V}_e = a + b\overline{D^2H}_e$, donde $\bar{V}_e = (\sum_{i=1}^{n} V_i)/n$ sea el volumen promedio de los árboles de la muestra y $\overline{D^2H}_e = (\sum_{i=1}^{n} D_i^2 H_i)/n$ sea el tamaño promedio de los árboles de la muestra. Por sustracción se llega al siguiente resultado:

$$\bar{V} = \bar{V}_e + b(\overline{D^2H} - \overline{D^2H}_e) \tag{2.4}$$

En esta ecuación se prestará especial atención a que $\bar{V}$ y $\overline{D^2H}$ sean las medias de todo el rodal al tiempo que $\bar{V}_e$ y $\overline{D^2H}_e$ son las medias de la muestra. Además $\overline{D^2H}$, $\overline{D^2H}_e$ y $\bar{V}_e$ son resultado de las mediciones, mientras que $\bar{V}$ es la cantidad que se trata de estimar.

En la fórmula (2.4), se reconoce el tipo de estimadores bien conocido en la teoría de muestreo: los *estimadores de la regresión*. La teoría de los estimadores de la regresión se expone con todo detalle en Cochran (1977, Capítulo 7) o en Thompson (1992, Capítulo 8). En el marco forestal se pueden hallar presentaciones sobre los estimadores de la regresión en de Vries (1986) y en Shiver & Borders (1996, Capítulo 6) (el primero es más bien teórico,

y el segundo más bien práctico). La teoría de los estimadores de la regresión se aplica en el caso de una relación lineal entre la cantidad que se debe predecir ($\bar{V}$ en el ejemplo anterior) y una variable auxiliar ($\overline{D^2H}$ en el ejemplo anterior). Por el contrario esta teoría está menos desarrollada en el caso de las relaciones no lineales o de regresiones múltiples aunque estos casos son frecuentes para los modelos de volumen.

La semiamplitud del intervalo de confianza en el umbral α (típicamente 5 %) de $\bar{V}$ es igual a (Cochran, 1977, p.199; Thompson, 1992, p.83):

$$t_{n-2}\,\hat{\sigma}\,\sqrt{\frac{1}{n}-\frac{1}{N}+\frac{(\overline{D^2H}-\overline{D^2H}_e)^2}{\sum_{i=1}^{n}(D_i^2H_i-\overline{D^2H}_e)^2}} \tag{2.5}$$

Cabe observar que el mínimo de esta amplitud es cero, que se alcanza cuando todo el rodal está incluido en la muestra ($n=N$, lo que conlleva $\overline{D^2H}=\overline{D^2H}_e$). Igual que antes, el plan de muestreo óptimo es tal que $\overline{D^2H}_e$ sea lo más cercano posible de $\overline{D^2H}$, con una varianza empírica máxima de D^2H en la muestra.

En la derivación del estimador de la regresión, partimos del supuesto de que el tamaño D^2H se mide en *todos* los árboles del rodal para llegar a la estimación del volumen total V_{tot}. En la práctica un protocolo de medición más realista es el siguiente: se mide el tamaño de los árboles en la muestra de tamaño $n' < N$; se mide al mismo tiempo el tamaño y el volumen en una submuestra de tamaño $n < n'$ de dicha muestra. La regresión del volumen con respecto al diámetro (es decir la tabla de volumen) se realiza a partir de una submuestra; se deduce una estimación del volumen de la muestra, luego, por extrapolación, de todo el rodal. Esta estrategia de muestreo se llama *muestreo doble*. Su teoría se presenta en Cochran (1977, Section 12.6) o, de manera más pragmática, en Shiver & Borders (1996, Capítulo 7). Su aplicación a la estimación de la biomasa de los rodales fue elaborada por Cunia (1987b,c,d).

Por último, las propiedades del estimador de la regresión (2.4) son conocidas en la teoría clásica del muestreo que no necesita la hipótesis del modelo lineal (2.1) sino que considera que la única fuente de variabilidad es el muestreo. En el caso de un plan de muestreo aleatorio simple y para un tamaño de muestra n suficientemente grande, en la teoría clásica la varianza de $\bar{V}$ es aproximadamente igual a (Cochran, 1977, p.195; Shiver & Borders, 1996, p.181):

$$\widehat{\text{Var}}(\bar{V})=\frac{1-n/N}{n(n-2)}\left\{\sum_{i=1}^{n}(V_i-\bar{V}_e)^2-\frac{\left[\sum_{i=1}^{n}(V_i-\bar{V}_e)(D_i^2H_i-\overline{D^2H}_e)\right]^2}{\sum_{i=1}^{n}(D_i^2H_i-\overline{D^2H}_e)^2}\right\} \tag{2.6}$$

y la semiamplitud del intervalo de confianza en el umbral α (típicamente 5 %) de $\bar{V}$ es aproximadamente igual a (Thompson, 1992, p.80; Shiver & Borders, 1996, p.185):

$$t_{n-2}\,\sqrt{\widehat{\text{Var}}(\bar{V})}$$

Esta última expresión es considerada más adecuada que la expresión (2.5) cuando la realidad se desvía del modelo de super-población (2.1) (Thompson, 1992, p.84).

En conclusión, este ejemplo simple muestra a la vez las ventajas y los limitaciones del muestreo para planificar los modelos de volumen: ventajas porque la teoría de muestreo permite planificar el número mínimo de árboles que deben medirse para alcanzar una precisión dada en las predicciones y permite optimizar el plan de muestreo; limitaciones porque el razonamiento supone conocer de antemano la forma de la tabla de volumen (y el modelo de super-población implícito) y usar dicho modelo para una aplicación dada. Ninguno de estos dos requisitos previos puede verificarse en la práctica. Además, los cálculos que son relativamente simples en el caso del modelo lineal que acabamos de presentar, se vuelven rápidamente inextricables para modelos más realistas.

2.2. Muestreo para la construcción de un modelo

Primero consideremos el problema de la predicción del volumen o de la biomasa de un árbol en particular con la ayuda de un modelo. ¿Cuántos árboles hay que medir para elaborar dicho modelo (§ 2.2.1)? ¿Cómo elegir esos árboles dentro del rodal? Esta segunda pregunta implica: ¿cómo desglosar los árboles de la muestra en función de las variables de entrada del modelo, empezando por su tamaño (§ 2.2.2)? Llegado el caso, ¿cómo estratificar la muestra (§ 2.2.3)? ¿Es mejor seleccionar individuos de la muestra en distintos lugares del bosque o, por el contrario, hacer un inventario de todos los árboles de una parcela dada (§ 2.2.4)?

2.2.1. Número de árboles

Debido a los límites de la teoría de muestreo, el número de árboles cubicados o pesados (en otras palabras, el tamaño de la muestra) se suele elegir de forma empírica, a partir de reglas resultantes de la experiencia. Un principio general es que, a igualdad de precisión, el tamaño de la muestra debe ser mucho mayor cuanto más variable sea el material: hacen falta menos árboles para una plantación de clones que para un bosque tropical natural, para una especie dada que para un grupo de especies, o para una parcela de 10 ha que para una región natural. En ciertos casos, como para la biomasa de las raíces, es el costo de la medición lo que orienta la elección del tamaño de la muestra más que la precisión supuesta de las predicciones: se elegirá un número de árboles que genera una cantidad de trabajo aceptable para la medición. A título indicativo, para la construcción de una tabla de volumen, la guía para agentes forestales *Mémento du forestier* (CTFT, 1989, p.256) recomienda que se midan unos 100 árboles "en caso de uno o varios rodales de plantación reciente en una superficie restringida (tipo parcelas de investigación silvícola)". Pardé & Bouchon (1988, p.108), por su parte, recomiendan los tamaños de muestra dados en el Cuadro 2.1, en función de la superficie de la zona en la cual se quiere usar el modelo. Zianis *et al.* (2005) efectuaron compilaciones del modelo de volumen y de biomasa para Europa y Henry *et al.* (2011) para África subsahariana. Los tamaños de las muestras utilizadas para los modelos mencionados en esas reseñas bibliográficas, permiten hacerse una idea del trabajo de muestreo realizado. Chave *et al.* (2004) demostraron que usando 300 árboles para elaborar un modelo de biomasa, la estimación de ésta en un bosque tropical húmedo (Isla de Barro Colorado en Panamá) daba un coeficiente de variación de apenas 3,1 %. Dicho coeficiente superaba el 10 % cuando el número de árboles usados para elaborar el modelo de biomasa estaba por debajo de 50, con una reducción del coeficiente de variación aproximadamente proporcional a $1/\sqrt{n}$ (Chave *et al.*, 2004, Figura 3). Van Breugel *et al.* (2011) encontraron la misma disminución en la precisión de la estimación con el tamaño de la muestra usada para elaborar el modelo de biomasa para n comprendida entre 49 y 195 árboles.

Cuanto más costosa resulta una observación en términos de tiempo y de esfuerzo de medición, más se tiende a realizar el plan de muestreo en función del trabajo de muestreo que se está dispuesto a realizar en vez de hacerlo en función de la precisión de la estimación esperada. Al ser la biomasa epigea de un árbol más difícil de medir que el volumen de su fuste, los modelos de biomasa tienden a elaborarse a partir de menos observaciones que los modelos de volumen. Algunos modelos de biomasa se elaboran solamente a partir de unas pocas mediciones (8 árboles para Brown *et al.*, 1995 en Brasil, 12 árboles para Ebuy Alipade *et al.*, 2011 en la República Democrática del Congo, 14 árboles para Deans *et al.*, 1996, 15 árboles para Russell, 1983 en Brasil). Los modelos para las raíces, que exigen todavía más trabajo de medición, suelen basarse en tamaños de muestras aún menores. Los modelos

Cuadro 2.1 – *Número de árboles por medir para determinar una tabla de cubicación en función de la superficie sobre la que se la quiere utilizar: recomendaciones de Pardé & Bouchon (1988).*

Zona	n
Rodal único y homogéneo	30
Parcela de 15 ha	100
Bosque de 1000 ha	400
Región natural	800
Área de la especie	2000 à 3000

elaborados con muestras tan pequeñas generalmente son poco fiables y sólo tienen una validez muy local. Sin embargo, estos pequeños conjuntos de datos pueden agruparse luego en conjuntos mayores que, a su vez, tienen ventajas para ajustar modelos (siempre y cuando se sepa controlar mediante covariables como edad, densidad de la madera, o por factores de estratificación como especie, tipo de formación vegetal, etc., la variabilidad inducida al reunir los datos).

2.2.2. Clasificación de los árboles

La clasificación de los árboles de la muestra en función de su tamaño (y, en líneas más generales, en función de las variables utilizadas como entrada del modelo), en principio, puede optimizarse. En el caso de una regresión lineal, por ejemplo, la semiamplitud del intervalo de confianza en el umbral α del gradiente de regresión es (Saporta, 1990, p.367):

$$t_{n-2}\,\frac{\hat{\sigma}}{S_X\sqrt{n}}$$

donde t_{n-2} es el cuantile $1-\alpha/2$ de una distribución de Student con $n-2$ grados de libertad, $\hat{\sigma}$ es la desviación estándar empírica de los residuos del modelo, n es el tamaño de la muestra y S_X es la desviación estándar empírica de la variable de entrada X dentro de la muestra:

$$S_X^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2 \quad \text{donde} \quad \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Cuando mayor sea S_X más precisa será la estimación de la pendiente, lo que, para un tamaño de muestra fijo, nos da un desglose de los árboles parecido al de la Figura 2.2. Ya vimos los límites de este razonamiento: aunque el plan de muestreo que consiste en tomar árboles en los dos extremos del gradiente de tamaño resulta óptimo cuando se ha verificado bien la hipótesis de una relación lineal, lleva a estimaciones erróneas cuando la relación no es lineal (Figura 2.3). Por tanto, en la práctica, conviene muestrear los árboles en todo el gradiente de tamaño de modo para garantizar la forma de la relación entre su volumen (o su masa) y su tamaño.

La teoría de las superficies de respuesta (Box & Draper, 1987; Goupy, 1999; Myers & Montgomery, 2002) permite optimizar la clasificación de los árboles en función de su diámetro a la altura del pecho (y, en líneas más generales, en función de las variables usadas como entrada en el modelo). No vamos a entrar en los detalles de esta teoría sino que nos contentaremos con algunos principios generales. El primero es extender al máximo el gradiente de tamaño de los árboles de la muestra.

Si la varianza del volumen (o de la masa) es constante, cualquiera que sea el tamaño del árbol, la regla es medir la misma cantidad de árboles en cada clase de tamaño (Pardé & Bouchon, 1988, p.108; CTFT, 1989, p.256). Para la muestra, tomar un número de árboles por clase de tamaño proporcional a la magnitud de esa clase en el rodal (en otras palabras, escoger los árboles al azar) sería un error. Sin embargo, la varianza del volumen raramente es constante; generalmente aumenta con el tamaño (heterocedasticidad de los residuos). En este caso la regla es aumentar la intensidad del muestreo de las clases más variables de forma que se garantice mayor precisión. En teoría, dentro de una clase de tamaño dada, lo ideal es medir una cantidad de árboles proporcional a la desviación estándar del volumen de los árboles de esa clase (CTFT, 1989, p.256). En la práctica, cuando la variable de entrada es el diámetro a la altura del pecho, una regla empírica consiste en tomar un número de árboles, constante por clase de área basal, lo que garantiza una mejor representación de los árboles de gran diámetro (CTFT, 1989, p.256–257).

El razonamiento se aplica también a otras variables explicativas. Si la variable de entrada del modelo es D^2H, se desglosarán los árboles según las clases de D^2H. Para los modelos de biomasa pluriespecíficos la densidad de la madera ρ se suele usar como variable de entrada (con la especificidad que se da a nivel de la especie y no a nivel del árbol). Para un modelo pluriespecífico que usa el diámetro D y la gravedad específica de la madera ρ como variable de entrada, una classificacíon adecuada de los árboles de la muestra consistiría en distribuirlos de forma uniforme por clase de diámetro y por clase de densidad de la madera.

2.2.3. Estratificación

Mostramos que escoger los árboles al azar al clasificarlos por tamaño, dando una *probabilidad de inclusión* igual a todos los árboles, no es un plan de muestreo óptimo. La estratificación también pretende tener en cuenta informaciones exógenas para definir *estratos* de muestreo homogéneos y así mejorar la precisión de las estimaciones. Al igual que antes, el principio es aumentar la intensidad de muestreo de los estratos más variables (con respecto a los otros estratos). Para retomar el ejemplo del párrafo 2.1, la varianza del estimador de regresión $\bar{V}$ en el caso de un muestreo estratificado (Cochran, 1977, p.202):

$$\begin{aligned}\widehat{\mathrm{Var}}(\bar{V}) \quad = \quad & \sum_h \left(\frac{N_h}{N}\right)^2 \frac{1-n_h/N_h}{n_h(n_h-2)} \left\{ \sum_{i=1}^{n_h} (V_{hi}-\bar{V}_{eh})^2 \right. \\ & \left. - \frac{\left[\sum_{i=1}^{n_h}(V_{hi}-\bar{V}_{eh})(D_{hi}^2 H_{hi} - \overline{D^2H}_{eh})\right]^2}{\sum_{i=1}^{n_h}(D_{hi}^2 H_{hi} - \overline{D^2H}_{eh})^2} \right\}\end{aligned} \tag{2.7}$$

donde h designa el estrato, N_h es el número de individuos del rodal que pertenecen al estrato h, n_h es el número de individuos de la muestra que pertenecen al estrato h, V_{ih} es el volumen del i-ésimo individuo del estrato h dentro de la muestra, $\bar{V}_{eh}$ es la media empírica del volumen promedio en el estrato h de la muestra, etc. Esta fórmula remplaza a (2.6). Para ilustrar el aumento de precisión aportado por la estratificación demos un pequeño ejemplo numérico. Para simplificar, supongamos que hay dos estratos y que cada uno corresponde al 50 % del rodal (de forma tal que $N_1/N = N_2/N = 0{,}5$), y que se muestrea dentro de cada estrato de forma tal que el segundo término entre corchetes de (2.7) sea despreciable. Además se supone que $n_1 \ll N_1$ y $n_2 \ll N_2$. La varianza del estimador de la regresión es entonces aproximadamente proporcional a:

$$\widehat{\mathrm{Var}}(\bar{V}) \propto \frac{1}{n_1-2}\left\{\frac{1}{n_1}\sum_{i=1}^{n_1}(V_{1i}-\bar{V}_{e1})^2\right\} + \frac{1}{n_2-2}\left\{\frac{1}{n_2}\sum_{i=1}^{n_2}(V_{2i}-\bar{V}_{e2})^2\right\}$$

Los términos entre corchetes representan las varianzas del volumen entre estratos. Supongamos que la desviación estándar del volumen sea de 4 m^3 en el primer estrato y de 2 m^3 en el segundo. El tamaño total de la muestra se fija en $n_1 + n_2 = 60$ individuos. Si se tiene en cuenta la estratificación, es decir, si se elige el número de árboles en cada estrato proporcionalmente a la frecuencia N_h/N del estrato en el rodal, entonces tenemos en este caso la misma cantidad de árboles en cada estrato de la muestra: $n_1 = n_2 = 30$ individuos. La varianza del estimador de la regresión es entonces aproximadamente:

$$\frac{4^2}{30-2} + \frac{2^2}{30-2} = 0{,}71 \text{ m}^6$$

Por el contrario, si determinamos el número de árboles en cada estrato proporcionalmente a la desviación estándar del volumen en el estrato, entonces: $n_1 = 2n_2$, donde $n_1 = 40$ individuos y $n_2 = 20$ individuos. La varianza del estimador por regresión es entonces aproximadamente:

$$\frac{4^2}{40-2} + \frac{2^2}{20-2} = 0{,}64 \text{ m}^6$$

Así pues vemos que, desde el punto de vista de la varianza del estimador, $30+30$ no es igual a $40+20$. Por otro lado, se podrá verificar que el mínimo de la función que en n_1 asocia $16/(n_1-2) + 4/(58-n_1)$ se obtiene para $n_1 = 39{,}333$.

Desde el punto de vista de la teoría de muestreo, la estratificación tiene por objeto aumentar la precisión de la estimación ajustando el plan de muestreo a la variabilidad dentro de cada estrato. Pero, desde el punto de vista de la elaboración de unmodelo de volumen, la estratificación tiene un segundo objetivo tan importante como el primero: comprobar que la relación entre el volumen (o la biomasa) y el tamaño de los árboles sea la misma dentro de cada estrato y, llegado el caso, elaborar el modelo para tantas relaciones como sea necesario. Este segundo punto está implícito en la fórmula (2.7) que se basa en un ajuste una pendiente b diferente (cf. ecuación 2.2) para cada estrato.

En resumen, la estratificación tiene por objeto explorar la variabilidad dentro de la zona de estudio para (*i*) hacer variar, llegado el caso, la forma del modelo en función de los estratos y (*ii*) adaptar el plan de muestreo a la variabilidad dentro de los estratos. Con frecuencia, para la elaboración de un modelo de volumen, el punto (*i*) prima sobre el punto (*ii*), mientras que ocurre lo contrario en la teoría de muestreo. La Figura 2.4 presenta estos dos objetivos.

Factores de estratificación

Todo factor capaz de explicar la variabilidad dentro de la zona de estudio puede considerarse: edad del rodal (sobre todo en el caso de plantaciones), fertilidad, sitio ecológico, tratamiento silvícola, variedad o especie, elevación, profundidad de nivel freática, etc. (Pardé & Bouchon, 1988, p.106; CTFT, 1989, p.255). Los factores de estratificación pueden ser anidados: estratificación según la región morfopedológica, luego según la fertilidad dentro de cada región, después según la edad dentro de cada clase de fertilidad, a continuación la densidad dentro de cada clase de edad. La "finura" de los factores de estratificación debe adaptarse también al contexto. Los factores de estratificación no serán los mismos según se razone a escala global como Brown (1997), a escala de un paisaje como van Breugel *et al.* (2011), o a escala de una plantación de clones como Saint-André *et al.* (2005). Brown (1997) propone modelos para todas las especies para las zonas climáticas (bosque seco, bosque muy húmedo). En el otro extremo, en una parcela de medición de intercambio gaseoso (eddy-correlation), con objeto de comparar las estimaciones de la producción neta de un

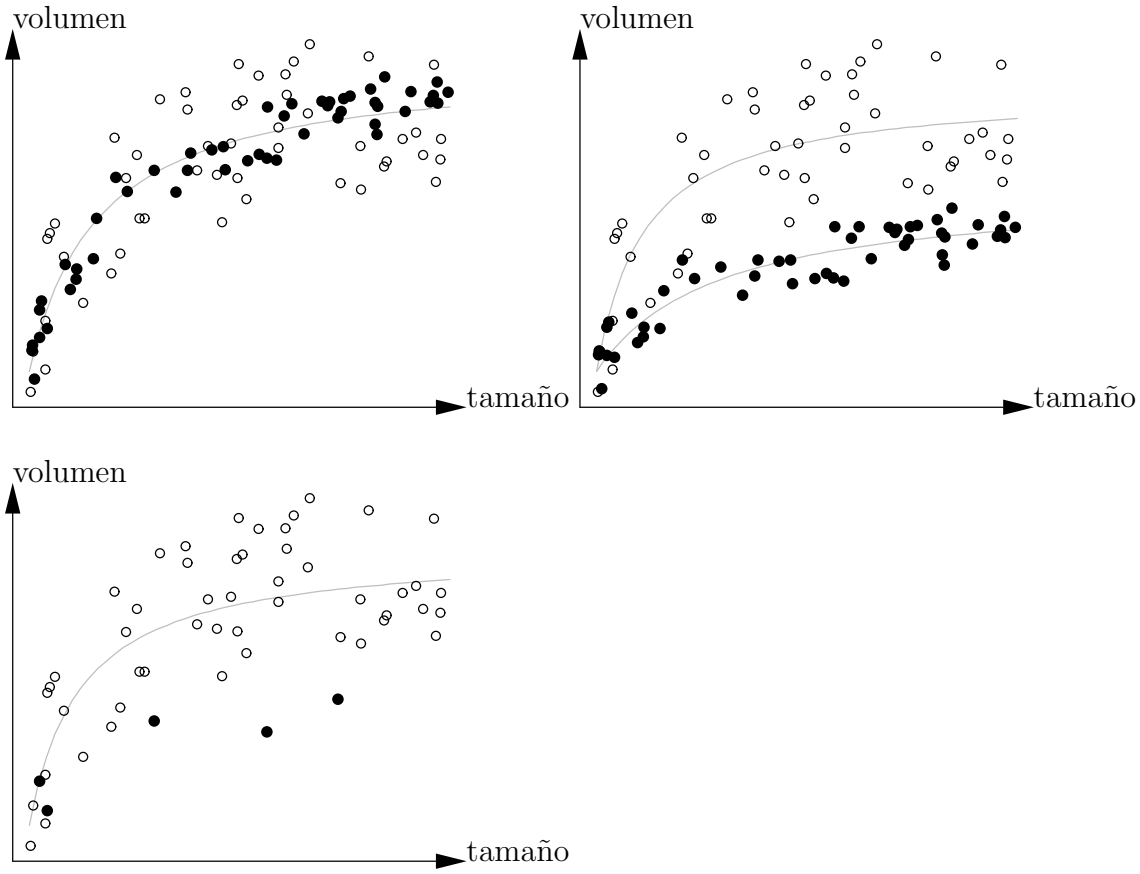


Figura 2.4 – *Predicción del volumen en función del tamaño para dos estratos (correspondientes a los puntos negros y blancos): arriba a la izquierda, los dos estratos corresponden a dos varianzas de los residuos (varianza más elevada para los puntos blancos que para los negros) pero la relación es la misma; arriba a la derecha, tanto la varianza como la relación varían entre los estratos; abajo, la situación es la misma que arriba a la derecha pero el segundo estrato fue objeto de un submuestreo de forma tal que se puede pensar que se trata de la misma relación entre los dos estratos.*

ecosistema (NEP), la estratificación podrá hacerse en función de la edad de la parcela, de la estación y de la huella de la torre de flujo.

Las especies como factor de estratificación

Para las formaciones naturales que contienen diversas especies, estas pueden considerarse también como un factor de estratificación. Para las formaciones pluriespecíficas es habitual elaborar un modelo de volumen para cada especie (o, al menos, para las más abundantes), y luego intentar reagruparlas sea por género o reuniendo todas las especies (modelo de "todas las especies"). Al fusionar los datos se aumenta el tamaño de la muestra, lo que es interesante si compensa el aumento de la variabilidad asociado a la mezcla de diferentes especies. Comparado a un modelo monoespecífico, el uso de un modelo pluriespecífico equivale a introducir un sesgo de predicción, que se puede considerar como la variabilidad entre especies. Van Breugel *et al.* (2011) cuantificaron de este modo el sesgo de predicción resultante de la suma de diversas especies. Así pues, fusionar los datos relativos a varias especies es ventajoso si el aumento de variabilidad dentro de una misma especie, aportada por esta fusión, compensa la variabilidad interespecífica introducida. No obstante hay que garantizar que (*i*) esta fusíon tenga sentido y que (*ii*) los tamaños de la muestra relativos a las diferentes especies sean comparables (Figura 2.4). Cuando de entrada procuramos elaborar un modelo "de todas las especies" (lo que suele ser el caso para los modelos de rodales naturales), hay que tener cuidado que la elección de los individuos que formarán parte de la muestra sea independiente de su especie, para no sesgar el modelo en favor de una especie en particular.

Asignación entre estratos

Una vez identificados los estratos, se adaptará el plan de muestreo en función de reglas empíricas. Si se dispone de una estimación *a priori* de la variabilidad del volumen (para modelo de volumen) o de la biomasa (para un modelo de biomasa) dentro de cada estrato, una regla empírica es tomar una intensidad de muestreo proporcional a la desviación estándar en cada estrato. Si no se dispone de una estimación *a priori* de la variabilidad, se tratará de tomar una intensidad de muestreo constante dentro de cada estrato (lo que no corresponde a un muestreo al azar puesto que los estratos no tienen la misma frecuencia en el rodal).

Modelos parametrizados

La información relativa a los estratos se incorpora luego al modelo de volumen creando un modelo diferente para cada estrato. Se puede probar si los modelos adaptados a los dos estratos son significativamente diferentes y, llegado el caso, fusionar ambos conjuntos de datos para elaborar un modelo único. También se podría hacer un modelo *parametrizado* a partir de los modelos para los distintos estratos siguiendo el principio del modelo mixto: los propios parámetros del modelo se convierten en funciones de variables que definen los estratos. Estos distintos puntos se desarrollarán en las Secciones siguientes dedicadas a la elaboración propiamente dicha de los modelos de biomasa o de volumen. Como ejemplo, Ketterings *et al.* (2001) elaboraron modelos de biomasa individuales con una entrada en forma de potencia:

$$B = aD^b$$

donde D es el diámetro a la altura del pecho y B su biomasa, para los distintos árboles de las diferentes especies en distintos lugares de la provincia de Jambi en Sumatra, Indonesia. El factor de sitio también se tuvo en cuenta en el parámetro b que se escribió como $b = 2+c$, donde c es el parámetro de la ecuación alométrica que asocia la altura al diámetro en cada

lugar: $H = kD^c$. El factor especie, por su parte, se tomó en cuenta en el parámetro a que se describió como $a = r\rho$, donde ρ es la densidad de la madera de la especie y r un parámetro constante. El modelo final, válido para todas las especies en todos los lugares, es un modelo parametrizado:

$$B = r\rho D^{2+c}$$

2.2.4. Selección de los árboles

Una vez definida la composición de la muestra, hay que determinar los árboles que se medirán en el campo. Dado que se trata de mediciones que demandan mucho tiempo y energía y, en el caso de la biomasa, son destructivas, la elección de los árboles debe efectuarse cuidadosamente. Una de las estrategias adoptadas por algunos para elaborar modelos de biomasa consiste en cortar todos los árboles en un área determinada (por ejemplo, en media hectárea). Esto tiene la ventaja de matar dos pájaros de un tiro, ya que aporta al mismo tiempo una estimación de la biomasa del rodal y observaciones individuales para la elaboración del modelo. En términos prácticos, el espacio liberado por la corta de los primeros árboles facilita luego la corta de los siguientes. Pero esta estrategia presenta un gran inconveniente: la distribución de los tamaños de los árboles en el rodal tiene muy pocas posibilidades de coincidir con la clasificación deseada de los árboles de la muestra por clase de tamaño, por lo que llevará a una distribución de tamaños de árboles de la muestra que no sea óptima. Lo mismo ocurrirá con todo factor usado para estructurar la muestra (clases de densidad de la madera, estratos, etc.). Además, la perturbación del rodal a esta escala tiene a veces consecuencias inesperadas. Djomo *et al.* (2010) mencionan una parcela que fue invadida por hormigas después de la corta de árboles, en tal magnitud que no fue posible mdir la biomasa de los árboles. Esta estrategia de elegir árboles tendrá que evitarse en las zonas infestadas por hormigas *Wasmannia*, porque sus ataques son altamente peligrosos.

En vez de elegir todos los árboles dentro de un área determinada, conviene seleccionar árboles uno por uno en función de las necesidades identificadas para construir la muestra. Esta estrategia puede demorar más para ponerse en práctica ya que necesita que se identifiquen individualmente los árboles. Teniendo en cuenta las dificultades impuestas por la medición de la biomasa de los árboles (cf. Capítulo 3), entre todos los árboles que satisfacen los criterios del plan de muestreo, se elegirán los que sean de más fácil acceso.

2.3. Muestreo para estimar un rodal

Consideremos ahora el problema de la predicción del volumen o de la biomasa de un rodal. Desde un punto de vista estadísticamente riguroso, habría que considerar toda la cadena de propagación de errores como descrita en la Figura 2.1 (Parresol, 1999). Esto nos lleva a plantearnos las cuestiones del doble muestreo y de estimadores de la regresión, como las definidas en la Sección 2.1.2. Cunia (1987b,c,d), Chave *et al.* (2004) y van Breugel *et al.* (2011) son raros ejemplos donde se tuvo en cuenta efectivamente la cadena de propagación de errores completa y donde el error de estimación de la biomasa de un rodal fue vinculado al tamaño de la muestra de árboles utilizada para elaborar el modelo de biomasa necesario para esta estimación. En la práctica, generalmente se simplifica el problema considerando el modelo como exacto y sin ningún error de predicción. Esta aproximación que equivale a desvincular el muestreo del rodal usado para predecir su volumen o su biomasa, del muestreo de los árboles para elaborar el modelo, reduce el primero a un problema clásico de inventario forestal.

No nos detendremos en esta cuestión del inventario forestal porque, por un lado, sigue siendo marginal con respecto al objetivo central del presente manual y, por otro, porque ya ha sido objeto de numerosas obras (Loetsch & Haller, 1973; Lanly, 1981; de Vries, 1986; Schreuder *et al.*, 1993; Shiver & Borders, 1996; West, 2009). No obstante, presentaremos algunos aspectos relativos a la estimación de la biomasa de un rodal.

2.3.1. Unidad de muestreo

Mientras que para la construcción de un modelo de biomasa es possible seleccionar los árboles que se incluirán en la muestra en forma individual, esta estrategia de muestreo no es realista cuando se trata de estimar la biomasa de un rodal. En este caso, optar más bien por medir todos los árboles dentro una área dada, incluso repitiendo esta técnica en otra área para ampliar el tamaño de la muestra. Esta área o parcela se convierte entonces en la unidad de muestreo. Asumimos que n es el número de parcelas inventariadas, N_i es el número de árboles encontrados en la i-ésima parcela ($i = 1, \ldots, n$), y B_{ij} es la biomasa del j-ésimo árbol de la i-ésima parcela ($j = 1, \ldots, N_i$), calculada con el modelo de biomasa y de las características medidas del árbol. El número N_i es aleatorio pero, para un árbol dado, la predicción de B_{ij} es determinista. se considera como determinista. La biomasa de la i-ésima parcela es entonces: $B_i = \sum_{j=1}^{N_i} B_{ij}$.

Siendo A la superficie de una parcela de muestreo y $\mathcal{A}$ a superficie del rodal. En un modelo de super-población, la biomasa del rodal se estima entonces mediante: $(\mathcal{A}/A)\,\bar{B}$, donde $\bar{B} = (\sum_{i=1}^{n} B_i)/n$ es la biomasa promedio de una parcela. Generalmente se considera que se conocen exactamente A y $\mathcal{A}$. El error de estimación de la biomasa del rodal se desprende entonces del de la biomasa media $\bar{B}$.

2.3.2. Relación entre el coeficiente de variación y el tamaño de las parcelas

Según el teorema central del límite, el intervalo de confianza en el umbral α de la esperanza de la biomasa de una parcela corresponde a la ecuación siguiente (Saporta, 1990, p.304). Esta expresión es exacta cuando la biomasa sigue una distribución normal o en el límite en el cual el número de parcelas tiende al infinito.

$$\bar{B} \pm t_{n-1} \frac{S_B}{\sqrt{n-1}}$$

donde t_{n-1} es el cuantile $1 - \alpha/2$ de una distribución t de Student con $n - 1$ grados de libertad, y S_B es la desviación estándar empírica de la biomasa de una parcela:

$$S_B^2 = \frac{1}{n-1} \sum_{i=1}^{n} (B_i - \bar{B})^2$$

Por definición, la precisión de la estimación E en el umbral α es la razón entre la semiamplitud del intervalo de confianza en el umbral α y la biomasa promedio:

$$E = t_{n-1} \frac{S_B}{\bar{B}\sqrt{n-1}} = t_{n-1} \frac{\mathrm{CV}_B}{\sqrt{n-1}} \tag{2.8}$$

donde $\mathrm{CV}_B = S_B/\bar{B}$ es el coeficiente de variación de la biomasa. Al redondear t_{n-1} a 2, el tamaño de la muestra n necesaria para alcanzar una precisión de estimación dada de E es:

$$n \simeq \left(\frac{2\mathrm{CV}_B}{E} \right)^2 + 1$$

El coeficiente de variación de la biomasa de una parcela de superficie A es por tanto el elemento esencial para construir el plan de muestreo. Además, como se desconoce a priori la superficie A de las parcelas, en realidad hay que conocer la relación entre el coeficiente de variación de la biomasa y la superficie A de las parcelas.

La derivación exacta de la relación entre A y CV_B exige especificar un modelo capaz de describir la distribución espacial de los árboles. La teoría de procesos puntuales responde a dicha necesidad (Cressie, 1993; Stoyan & Stoyan, 1994). El cálculo exacto de la relación entre A y CV_B en el marco de un proceso puntual es viable pero complicado (Picard *et al.*, 2004; Picard & Bar-Hen, 2007). El cálculo exacto permite darse cuenta de dos cosas:

1. aunque la *forma* de las parcelas tenga un efecto sobre el coeficiente de variación (como se ha demostrado empíricamente, cf. Johnson & Hixon, 1952; Bormann, 1953), tiene un efecto despreciable comparado al tamaño de las parcelas;

2. la relación entre A y CV_B puede aproximarse por una relación de potencia (Fairfield Smith, 1938; Picard & Favier, 2011):

$$\mathrm{CV}_B = kA^{-c}$$

En la práctica es esta relación de potencia la que suele especificarse. Intuitivamente, el valor $c = 0{,}5$ corresponde a una distribución espacial aleatoria de la biomasa dentro del rodal; un valor $0 < c < 0{,}5$ corresponde a una distribución espacial agregada de la biomasa; y un valor $c > 0{,}5$ corresponde a una distribución espacial regular de la biomasa (CTFT, 1989, p.284). Usando los datos de la biomasa de una parcela de gran tamaño en Paracou, en la Guayana Francesa, Wagner *et al.* (2010) descubrieron que:

$$\mathrm{CV}_B = 557 \times A^{-0{,}430} \qquad (A \text{ en m}^2,\ \mathrm{CV}_B \text{ en } \%)$$

Según la interpretación anterior, esto corresponde a una distribución espacial ligeramente agregada de la biomasa. En la Amazonia brasileña, Keller *et al.* (2001) encontraron la relación siguiente (ajustada a los datos de su Figura 4 con $R^2 = 0{,}993$ con datos transformados logarítmicamente):

$$\mathrm{CV}_B = 706 \times A^{-0{,}350} \qquad (A \text{ en m}^2,\ \mathrm{CV}_B \text{ en } \%)$$

El valor menor (en valor absoluto) del exponente refleja una distribución espacial de la biomasa mucho más agregada que en la Guayana Francesa. Un estudio parecido fue realizado por Chave *et al.* (2003) usando los datos de una parcela de 50 ha en la Isla de Barro Colorado en Panamá. Chave *et al.* (2003) reportaron en su Cuadro 5 los valores de la amplitud del intervalo de confianza al 95 % no para la esperanza de la biomasa de una parcela sino para la esperanza de la biomasa por unidad de superficie. La amplitud del intervalo de confianza al 95 % de la esperanza de la biomasa de una parcela corresponde entonces a la amplitud reportada por Chave *et al.* (2003) multiplicada por la superficie de la parcela, o sea:

$$2t_{n-1}\frac{S_B}{\sqrt{n-1}} = \Delta \times A$$

donde Δ es la amplitud del intervalo de confianza al 95 % mencionado por Chave *et al.* (2003) en su Cuadro 5. De eso se deduce:

$$\mathrm{CV}_B = \frac{S_B}{\bar{B}} = \frac{\Delta A\sqrt{n-1}}{2t_{n-1}\bar{B}} = \frac{\Delta\sqrt{n-1}}{2t_{n-1}\mu}$$

Cuadro 2.2 – *Coeficiente de variación de la biomasa de una parcela en función de su tamaño: datos tomados del Cuadro 5 de Chave et al. (2003) para la Isla de Barro Colorado en Panamá.*

A (m^2)	n	Δ ($Mg\,ha^{-1}$)	CV_B (%)
100	5000	17,4	114,5
200	2500	18,7	87,0
400	1250	20,0	65,7
1000	500	21,4	44,4
2500	200	20,1	26,2
5000	100	22,4	20,5
10000	50	23,5	14,9

donde μ es la biomasa promedio por unidad de superficie, igual a 274 $Mg\,ha^{-1}$ en el estudio de Chave *et al.* (2003). El Cuadro 2.2 completa el Cuadro 5 de Chave *et al.* (2003) con el valor de CV_B calculado de ese modo. Los valores de CV_B dados en el Cuadro 2.2 se ajustan muy bien ($R^2 = 0{,}998$ con datos transformados logarítmicamente) a la relación de potencia siguiente que toma en cuenta el tamaño de las parcelas:

$$CV_B = 942 \times A^{-0{,}450} \qquad (A \text{ en } m^2,\ CV_B \text{ en } \%)$$

La variabilidad de la biomasa (representada por el valor del coeficiente multiplicador $k = 942$) es mayor que en Paracou pero la estructuración espacial de la biomasa (representada por el exponente $c = 0{,}45$) es bastante parecida a la observada en Paracou por Wagner *et al.* (2010). Además, el hecho de que c se aproxime al valor 0,5 representa una pequeña agregación espacial de la biomasa. Chave *et al.* (2003) por otro lado, subrayan que no hay una autocorrelación espacial significativa de la biomasa (lo que correspondería a $c = 0{,}5$, o a un valor constante de Δ).

2.3.3. Elección del tamaño de las parcelas

La elección del tamaño de las parcelas de muestreo puede hacerse de forma tal que optimice la precisión de la estimación en función de un esfuerzo de muestreo (Bormann, 1953; Schreuder *et al.*, 1987; Hebert *et al.*, 1988), o de manera tal que minimice el esfuerzo de muestreo para una precisión dada de la estimación (Zeide, 1980; Gambill *et al.*, 1985; Cunia, 1987c,d). Estos dos puntos de vista son duales uno con respecto al otro y llevan al mismo óptimo. El trabajo de muestreo puede cuantificarse sencillamente mediante la tasa de muestreo $n \times A/\mathcal{A}$ o, en forma más realista, mediante el costo cuya expresión es más compleja. Examinemos ambas opciones.

Tasa de muestreo fijo

A una tasa de muestreo constante, el área A y el número n de parcelas de muestreo están unidas por una relación inversamente proporcional: $n \propto 1/A$. La elección del tamaño de las parcelas se reduce a la pregunta siguiente: "¿conviene más tener pocas parcelas grandes o muchas parcelas pequeñas?", lo que también se llama el dilema SLOSS (del inglés "single large or several small"; Lahti & Ranta, 1985). Si calculamos la relación $n \propto 1/A$ en (2.8) (y considerando que $t_{n-1}/\sqrt{n-1}$ es ligeramente diferente a $2/\sqrt{n}$):

$$E \propto 2\, CV_B \sqrt{A}$$

Si la distribución espacial de la biomasa es aleatoria, entonces $\mathrm{CV}_B \propto A^{-0,5}$ y, en consecuencia, la precisión de la estimación E es independiente del tamaño A de las parcelas. Si la distribución espacial de la biomasa es agregada, entonces $\mathrm{CV}_B \propto A^{-c}$ con $c < 0,5$ y, por ende, $E \propto A^{0,5-c}$ con $0,5 - c > 0$: la precisión de la estimación es mucho mejor (valor de E pequeño) cuanto menor es el área A de las parcelas. En este caso, con una tasa de muestreo fija, conviene más tener muchas parcelas pequeñas que pocas y grandes. Es lo que se observa en el Cuadro 2.2, donde el valor de Δ disminuye cuando disminuye A (esta disminución sigue siendo pequeña porque c es cercano a 0,5). Si la distribución espacial de la biomasa es regular, $\mathrm{CV}_B \propto A^{-c}$ con $c > 0,5$ por ende, $E \propto A^{0,5-c}$ con $0,5 - c < 0$, entonces la precisión de la estimación es mucho mejor (valor de E pequeño) cuanto mayor es el tamaño A de las parcelas. En este caso, con una tasa de muestreo fija, conviene más tener pocas parcelas grandes que muchas pequeñas.

Las magnitudes medidas en biología suelen tener una distribución espacial agregada ($c < 0,5$), a veces aleatoria ($c = 0,5$), raras veces regular ($c > 0,5$) (Fairfield Smith, 1938). En otras palabras, el dilema SLOSS se resolverá frecuentemente a favor una multitud de pequeñas parcelas. Si llevamos este razonamiento hasta sus últimas consecuencias, vemos que la precisión de la estimación será óptima (valor E mínimo) para $A = 0$, es decir, ¡creando una infinidad de parcelas de tamaño nulo! Aquí se ven los límites de este razonamiento. Cuando se cuantifica el trabajo de muestreo en función de la tasa de muestreo $nA/\mathcal{A}$, se supone implícitamente que el costo de muestreo, es decir el tiempo o el dinero necesario para dicho muestreo, es proporcional a nA. Esto equivale a considerar solamente un costo por superficie, es decir, un costo de muestreo que sea proporcional a la superficie inventariada.

Costo de muestreo

En realidad el costo relativo al área no es más que un componente del costo de muestreo. El inventario propiamente dicho de las parcelas de muestreo, cuya duración es proporcional a la superficie inventariada, no es la única tarea que lleva tiempo. Delimitar las parcelas de muestreo también toma tiempo. Esta delimitación es proporcional a su perímetro acumulado: se trata de un costo lineal. Ir de una parcela a otra también consume tiempo. Es más realista entonces medir el esfuerzo de muestreo por un costo que tenga en cuenta todas esas tareas en vez de hacerlo simplemente mediante la intensidad de muestreo. Si medimos este costo en términos de tiempo y si las parcelas de muestreo tienen forma cuadrangular, dicho costo será, por ejemplo (Zeide, 1980; Gambill *et al.*, 1985):

$$C = \alpha\, nA + \beta \times 4n\sqrt{A} + \gamma\, d(n, A)$$

donde α es el tiempo de inventario por unidad de superficie, β es el tiempo de delimitación por unidad de longitud ($4\sqrt{A}$ representa el perímetro de una parcela cuadrada de superficie A), γ es la velocidad de desplazamiento y $d(n, A)$ es la longitud del camino que une las n parcelas de muestreo. Se puede completar esta expresión del costo de muestreo para tener en cuenta otras tareas. El razonamiento utilizado en el párrafo precedente equivalía a plantear $\beta = \gamma = 0$. Con $\beta > 0$ y $\gamma > 0$, la solución al dilema SLOSS ya no será $A = 0$ en el caso de una distribución espacial agregada de la biomasa ($c < 0,5$).

Otras restricciones

Limitar la cuestión del muestreo de la biomasa de un rodal a una cuestión de precisión de la estimación es demasiado restrictivo. Con frecuencia, la cuestión no se limita a la estimación de la biomasa del rodal sino que se persiguen múltiples objetivos simultáneamente. Por ejemplo, se tratará de estimar no sólo la biomasa del rodal sino también sus variaciones a lo

largo del tiempo. En este caso, teniendo en cuenta la mortalidad, las superficies que habrá que inventariar pueden ser entonces muy superiores (Chave *et al.*, 2003, 2004; Rutishauser *et al.*, 2010; Wagner *et al.*, 2010). Se tratará tal vez de estimar la biomasa de las parcelas de forma que permitan establecer una relación con los índices resultantes de imágenes satelitales, para extrapolar la estimación de la biomasa a escala del paisaje. En este caso, la superficie de las parcelas de muestreo esta restringida por la resolución de las imágenes del satélite y por la superficie mínima necesaria para calcular los índices satelitales.

Además, el tipo de plan de muestreo considerado implícitamente aquí, a saber, un plan aleatorio simple mediante parcelas de tamaño fijo, no suele ser el más eficiente (por. ej., con la mejor precisión de la estimación a un costo de muestreo dado). A escala de paisaje con distintos tipos de bosques, otros tipos de muestreo pueden resultar más eficientes (Whraton & Cunia, 1987; van Breugel *et al.*, 2011). Comparado con un plan aleatorio simple con tamaño de muestra único, un plan de muestreo estratificado será más caro (puesto que la estratificación implica un costo) pero dará una mejor precisión de la estimación. Un plan por conglomerados implicaría un costo menor (porque habría que desplazarse menos) pero daría una menor precisión de la estimación. Las técnicas específicas de inventario forestal, tales como el inventario por distancias (Magnussen *et al.*, 2008a,b; Picard & Bar-Hen, 2007; Picard *et al.*, 2005) o el inventario con el relascopio de Bitterlich (Schreuder *et al.*, 1993; West, 2009), se basan en parcelas de tamaño variable y también pueden ser alternativas más eficientes a los enfoques que usan parcelas de tamaño fijo.

3
Fase de campo

La fase de campo es la más crucial porque puede generar errores de medición que no pueden corregirse posteriormente. Esta fase debe regirse por tres principios clave: (*i*) es preferible pesar todo el material en el campo que calcular un volumen y multiplicarlo luego por una medida de densidad (cf. Capítulo 1, y las variaciones de la forma de los fustes y de la densidad de la madera en los árboles); (*ii*) si se toma una alícuota, hay que pesar el total y luego la alícuota para garantizar el seguimiento de la pérdida de humedad; por último, (*iii*) es muy difícil y también muy caro realizar una campaña de biomasa así que pueden hacerse otras mediciones al mismo tiempo para evitar tener que volver luego en el campo (por ejemplo, perfil de los fustes, muestreo para la mineralomasa).

La selección de los árboles que se miden en el campo (véase el Capítulo 2), ya se haga por individuo o sea exhaustiva en una superficie dada, requiere que se marquen los árboles con pintura, que se mida la circunferencia, de ser posible a 1,30 m (haciendo un círculo de pintura a esa altura), y también la altura. Por un lado, éstos procedimientos permiten verificar que el árbol seleccionado corresponda al plan de muestreo elegido (en caso de una selección por individuo) y, por otro, realizar mediciones de control una vez que se ha derribado el árbol. También resulta práctico tomar una foto del individuo seleccionado y hacer un esquema sintético en la ficha de campo. Esto facilita la interpretación de los datos y la verificación de los resultados obtenidos. En general no se seleccionan los árboles demasiado particulares (copa rota, fuste nudoso o sinuoso) a menos que representen una proporción significativa del rodal o si el objetivo es cuantificar un accidente (por ejemplo, la rotura de la copa como consecuencia de una helada). Asimismo cabe excluir los árboles situados en un entorno no representativo (bordes del bosque, claros, bosque degradado, etc.). En efecto, su arquitectura suele ser diferente de los otros árboles del rodal. Por último, no es raro que los obstáculos del terreno (pendiente, acceso, rodal no conforme al estrato, etc.) pongan en tela de juicio la muestra inicial.

La base general de las mediciones de biomasa y, mucho más aún de mineralomasa, reside en una regla de tres entre la biomasa fresca medida en el campo, la biomasa fresca de la alícuota y la biomasa seca de la alícuota. Como los distintos órganos de un árbol no tienen el mismo porcentaje de humedad ni la misma densidad, es preferible proceder por partes para tener en cuenta las variaciones de densidad y de humedad en el árbol (y de concentración en elementos minerales para la mineralomasa). La estimación de la biomasa será mucho más

precisa cuanto más fina sea la estratificación pero eso exige más trabajo.

Hay que encontrar una solución intermedia entre la precisión de la medición y la rapidez del trabajo en el campo. Tradicionalmente las partes del árbol se definen del modo siguiente: el tronco, diferenciando la madera de la corteza, que conviene cortar en secciones para tomar en cuenta las variaciones de densidad y de humedad en función del diámetro de las secciones; las ramas, tomando las muestras generalmente por clases de diámetro, diferenciando o no la madera de la corteza; las ramas pequeñas suelen incluir las yemas; las hojas; los frutos; las flores; y, por último, las raíces por clases de diámetro. En la Figura 3.1 se da un ejemplo de esa división por partes para el haya.

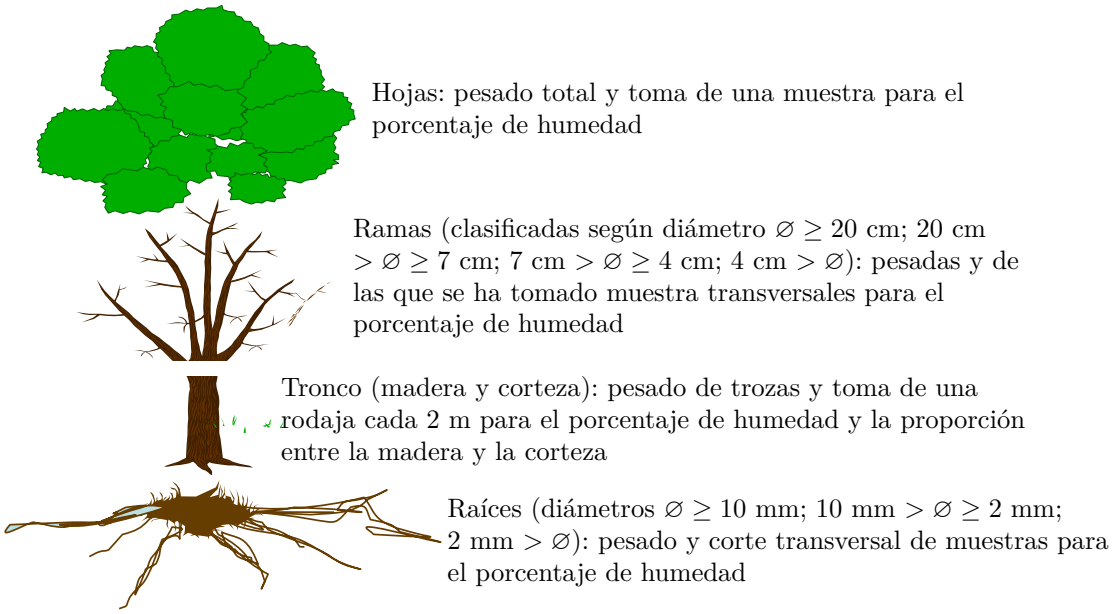


Figura 3.1 – *Ejemplo de las secciones de los árboles para una campaña de biomasa y de mineralomasa en el haya en Francia.*

Es posible aprovechar las áreas de corta existentes para cosechar los árboles necesarios para elaborar el modelo. En efecto, el acceso y la corta de árboles suelen estar reglamentados en los bosques y la explotación silvícola brinda uno de los únicos medios de tener acceso a los árboles deseados. Sin embargo, este método corre el riesgo de introducir un sesgo en la selección de los árboles ya que las especies cortadas serán principalmente comerciales. No se cortarán las otras a menos que estorben el derribo de un árbol seleccionado por la compañía forestal o bien si se encuentran en los caminos de acopio y arrastre o el área de almacenamiento. Además, los árboles derribados por motivos comerciales no siempre pueden ser cortados en trozas de tamaño razonable para poder pesarlos en el campo. Todo depende de la capacidad de las básculas de que se dispone y de la longitud de las trozas. Estas restricciones implican efectuar una cuidadosa selección de los individuos y combinar dos métodos: (1) pesado integral de las secciones no comerciales de los árboles, en especial las ramas; (2) mediciones de volumen y de densidad de la madera para el tronco.

Por estas razones no existe un método de campo estándar ya que cada uno tendrá que adaptarse al contexto. Por el contrario, en el marco del presente manual presentamos tres casos típicos que sientan las bases para realizar luego cualquier campaña de campo. El primero se refiere a los bosques regulares (resultantes de la regeneración o plantados), el segundo a un bosque seco y el tercero a un bosque tropical muy húmedo. En el primer caso,

todos los compartimentos se pesan directamente en el campo. En el segundo, no pueden cortarse los árboles y las mediciones son semidestructivas. El tercero se refiere a árboles de dimensiones demasiado grandes para un pesado integral en el campo. Las mediciones se obtienen a partir de las tres fases que se describen a continuación: el campo, el laboratorio y el cálculo informático. Como el trabajo de campo y el cálculo informático son específicos a cada método, se los presenta para cada uno de los casos. Los procedimientos de laboratorio son generalmente los mismos para todos los casos.

3.1. Pesado directo de todos los compartimientos en el campo

El primer caso que consideraremos es el más frecuente. Se trata de pesar directamente en el campo todos los compartimientos. El sistema propuesto es el resultado de varias campañas de campo efectuadas en rodales tanto de clima templado como tropical. Presentamos ejemplos tomados en distintos rodales regulares: plantaciones de eucalipto en el Congo (Saint-André *et al.*, 2005), de caucho de Tailandia, bosques procedentes de semillas de haya y roble en Francia (Genet *et al.*, 2011). Rivoire *et al.* (2009) dan un ejemplo de aplicación de este método, con un complemento de mediciones sobre ahusamiento acentuado de ramas grandes y toma de muestras para la mineralomasa.

3.1.1. En el campo

El aprovechamiento forestal es una actividad compleja cuya organización debe ser fluida para que todos los equipos puedan trabajar sin perder tiempo (véase el detalle de estos equipos en 3.6). El responsable del área de corta prepara la operación de antemano, haciendo una preselección de los árboles con su localización en el campo. A continuación hay un trabajo de laboratorio para (*i*) preparar el material necesario (véanse los detalles en 3.5), (*ii*) preparar los formularios de campo (pesaje de las diferentes partes del árbol, mediciones conexas), (*iii*) preparar las bolsas donde se pondrán las distintas alícuotas tomadas de los árboles (véase la Figura 3.1), (*iv*) explicar a los distintos participantes cómo se organiza el trabajo en el campo para que sepan qué hacer en el campo. La Figura 3.2 propone una organización eficaz para una campaña de biomasa, con siete pasos que trabajan simultáneamente.

Teniendo en cuenta que el desrame demora más, conviene comenzar el trabajo con un árbol de gran tamaño (Foto 3.3). El responsable del área de corta acompaña a los leñadores y coloca al pie del árbol las bolsas destinadas a recoger la muestras (paso 1). El tamaño de las bolsas debe adaptarse al de las muestras que se tomarán. Las bolsas deben llevar sistemáticamente la referencia del compartimiento, del árbol y de la parcela. Después la corta del árbol, el primer equipo que interviene es el que mide los perfiles del tronco (paso 2). Cuando este equipo ha terminado, y mientras los equipos de cortadores de ramas comienzan a trabajar en el primer árbol, pasa al segundo árbol que, entre tanto, han cortado los leñadores (paso 3). Hay que calcular aproximadamente media jornada para un árbol de 12 toneladas (entre 90–100 cm de diámetro). Cuando los cortadores de ramas terminan con el primer árbol, los leñadores ya han tenido tiempo de cortar bastantes árboles para que el equipo de perfiles tenga los suficientes para medir durante todo el día. Los leñadores pueden volver luego al primer árbol para segmentarlo y tomar las rodajas de muestra (paso 4). Una vez efectuadas ambas tareas en el primer árbol, los leñadores pasan al segundo que, entre tanto, ya fue desramado. En este punto se pesan las hojas, las trozas y las ramas del primer árbol (paso 5) mientras el responsable del área de corta toma muestras de las hojas y de las ramas (paso 6). El conjunto de muestras, incluidas las rodajas de muestra, se lleva al área de pesaje de dichas muestras (paso 7). Cuando el equipo de perfiles del tronco termina con

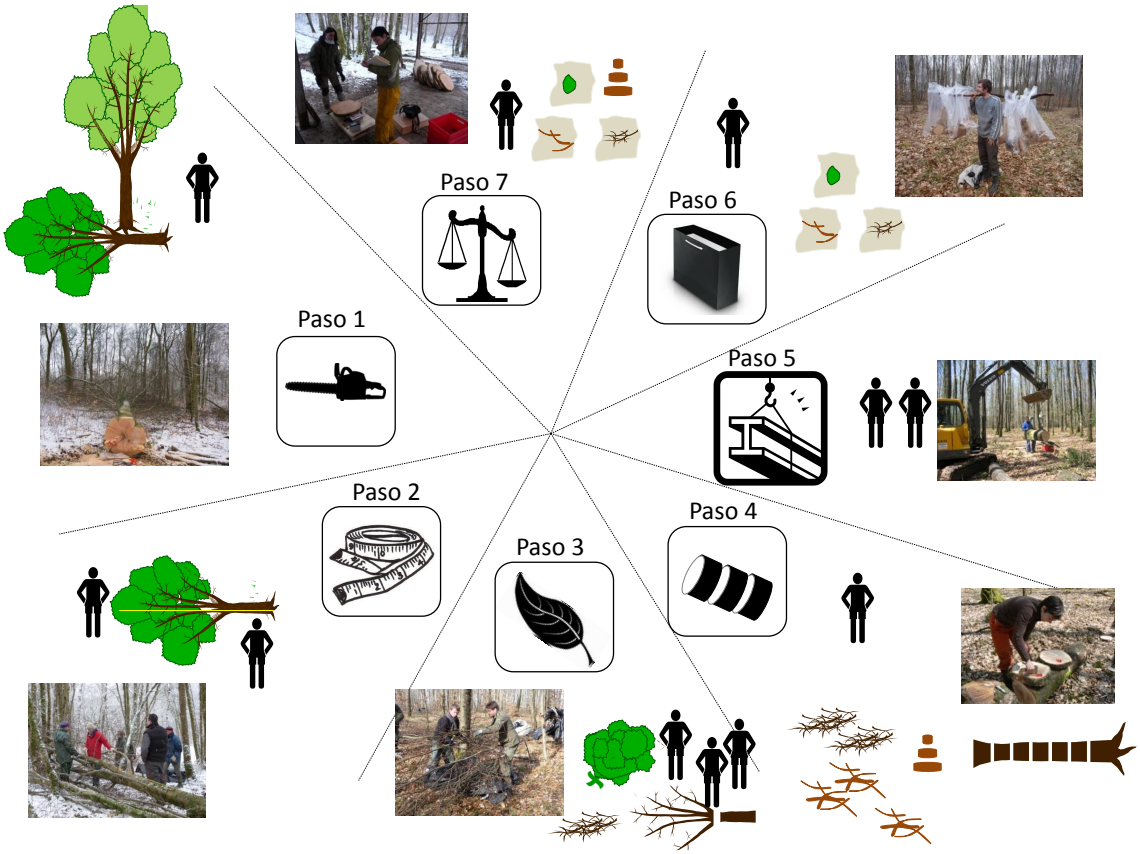


Figura 3.2 – *Organización de un área de medición de biomasa con 7 pasos. Paso 1, preparación del terreno y corta de los árboles (Foto: L. Saint-Andre); paso 2, medición de los árboles cortados: perfiles de pies, posición de las trozas (Foto: M. Rivoire); paso 3, deshoje y desrame (Foto: R. D'Annunzio y M. Rivoire); paso 4, troza y etiquetado de rodajas (Foto: C. Nys); paso 5, pesaje de trozas y de leñas de ramaje (Foto: J.-F. Picard); paso 6, muestreo de ramas (Foto: M. Rivoire); paso 7, área de pesaje de las muestras (Foto: M. Rivoire).*

todos los árboles del día, puede acercarse a esta área también para terminar el pesaje.

Foto 3.3 – *Campaña de medición en un monte medio en Francia. A la izquierda, llegada al área de corte e instalación del material (Foto: M. Rivoire); a la derecha, apeo del primer árbol (Foto: L. Saint-André).*

Este esquema cronológico es válido cuando las condiciones climáticas son templadas. En climas tropicales no es posible esperar al final del día para pesar las muestras. Por eso la medición de las muestras debe hacerse al mismo tiempo que las trozas y las ramas. Si no es posible pesarlas *in situ*, habrá que hacerlo en el laboratorio pero después de transportar las muestras en una caja hermética para limitar al máximo la evaporación del agua contenida en ellas. Esta debe ser la última solución ya que los pesos tomados en el campo son mucho más fiables.

La corta del árbol (paso 1)

El leñador prepara el árbol seleccionado mientas que los técnicos cortan los pequeños árboles o tallos que puedan perturbar la caída del árbol y limpian el lugar antes de la corta. Se puede poner una lona en el suelo para no perder las hojas durante la corta (Foto 3.4). La caída del árbol puede arrastrar otras copas por lo que los técnicos separarán las ramas que pertenezcan al árbol seleccionado de aquellas correspondientes a otros árboles.

Las mediciones en el árbol (paso 2)

Luego de la corta se miden los perfiles del tronco (Foto 3.5). La corta no se hace nunca a nivel del suelo, por lo que es indispensable marcar la altura a 1,30 m con pintura en el tronco antes de cortarlo y poner la cinta métrica con la graduación a 1,30 m en la marca de pintura, una vez cortado el árbol. Esto permite evitar un sesgo en la localización de las secciones (el desfase inducido por la altura de corte). Las circunferencias suelen medirse cada metro o, algo más útil aún para elaborar modelos de perfiles del tronco, como porcentaje de la altura total. No obstante, este método es mucho más difícil de aplicar en el campo. Cuando no es posible medir la circunferencia porque el tronco está totalmente apoyado sobre el suelo, hay que hacerla con una forcípula tomando dos diámetros perpendiculares uno al otro. A lo largo del tronco se marcan con pintura o cinta los puntos donde se acordó con la autoridad forestal (o el que ha comprado la madera), se seccionaría el fuste en trozas.

En el caso de árboles rectos con un tronco principal claramente identificado, no hace falta elegir el eje principal. Por el contrario, en el caso de los fustes muy sinuosos o ramosos (copa de las frondosas), hay que identificar bien el eje principal. Éste puede distinguirse, por

FOTO 3.4 – *Campaña de biomasa en el Congo en una plantación de eucaliptos. A la izquierda, deshoje de un árbol encima de una lona (Foto: R. D'Annunzio). A la derecha, fin del proceso para un árbol mostrando, en el área de pesaje, las bolsas que contienen las hojas, las trozas y las ramas (Foto: L. Saint-André).*

ejemplo, aplicando una marca de pintura. El eje principal se diferencia de los otros por ser su diámetro el mayor en cada bifurcación del tronco. Todos los ejes que parten del tronco principal se consideran como ramas. En el caso de los árboles multicaules, es posible incluir cada eje en el tronco principal (Foto 3.6), o bien considerar cada rama como un individuo. En este último caso habrá que identificar el eje principal en cada uno de ellos.

FOTO 3.5 – *A la izquierda, campaña de biomasa en Ghana en un bosque de teca: medición del ramaje (Foto: S. Adu-Bredu). A la derecha, campaña de biomasa en Francia en un bosque regenerado: medición de perfiles del tronco (Foto: M. Rivoire).*

A continuación se determina la longitud del tronco al igual que la posición de la primera rama viva y de las grandes horquetas. Se pueden varias mediciones de la altura en el árbol cortado, por ejemplo, altura del extremo fino < 1 cm, altura donde el diámetro de corte es 4 cm y altura donde el diámetro de corte es 7 cm. Las mediciones efectuadas en el árbol cortado pueden compararse luego con las mediciones efectuadas durante el inventario forestal de los árboles en pie. Esto permite verificar la coherencia de los conjuntos de datos y eventualmente corregir los datos aberrantes a sabiendas que puede haber diferencias debido

Foto 3.6 – *Campaña de biomasa en las plantaciones de caucho en Tailandia. A la izquierda, un árbol cortado multicaule (3 fustes en el mismo tocón): desrame y deshoje. A la derecha, mezcla de hojas antes de tomar una alícuota (Fotos: L. Saint-André).*

a la imprecisión de la medición de la altura antes de la corta (en general 1 m), o de la sinuosidad del fuste o de las interrupciones durante las mediciones de longitud después de la corta.

Troceo (pasos 3 y 4)

Lo ideal es poder segmentar el árbol en trozas de 2 m de largo para poder tener en cuenta las variaciones de densidad de la madera y de la humedad del fuste. Una vez preparado el árbol, se separan las ramas del tronco (al igual que las hojas, si es necesario). A continuación se vuelven a cortar las ramas para clasificarlas según el diámetro del extremo fino. Si se trata de un rodal de latifoliadas templado, los cortes se hacen en general por clase de diámetro > 20 cm, 20–7 cm, 7–4 cm, < 4 cm. En el caso del eucalipto la República del Congo, las ramas se han dividido en dos grupos: < 2 cm y > 2 cm. Se arman haces de ramillas con marcos de hierro y dos sogas sólidas (véanse la Sección 3.5 y la Foto 3.14). Cuando las ramas tienen hojas, conviene separarlas de las ramillas. Para hacerlo, hay que usar lonas para no perder las hojas. Si las hojas no se desprenden bien de los ejes leñosos (por ejemplo, encina o resinosa), conviene adoptar entonces una estrategia de submuestreo (véase el ejemplo siguiente en Camerún). Las hojas se colocan en grandes bolsas de plástico para pesarlas. El desrame y deshoje son actividades que demoran y a las que hay que asignar los recursos humanos adecuados (número de equipos suficientes) para no demorar el trabajo de los leñadores. Para las ramas principales de un árbol que suelen tener un diámetro considerable (> 20 cm), conviene proceder del mismo modo que para el tronco, mediante el troceo y la extracción de rodajas.

El troceo se realiza una vez que se han separado las ramas del tronco principal. Se toma una rodaja de unos 3–5 cm de espesor a nivel del tocón y luego cada x metros (Foto 3.7). El largo x las trozas depende de la dimensión del árbol y de las disposiciones tomadas con la administración forestal o el rematante. Ya que este trabajo de campo es fastidioso y largo, hay que aprovecharlo bien para tomar múltiples muestras (por ejemplo, sacar una rodaja adicional para mediciones más detalladas de densidad de la madera o de mineralomasa —

véase, por ejemplo Saint-André *et al.*, 2002b, para las concentraciones de elementos minerales en los fustes de eucalipto). Es importante indicar la posición de cada rodaja muestreada. Hay que pesarlas *in situ* el mismo día en que se procesa el árbol para minimizar las pérdidas de humedad (para lo cual hacen falta dos personas — en general es el equipo del perfil de tronco el que se encarga de esta tarea interrumpiendo su trabajo un poco antes para efectuar el pesaje de las rodajas, véase la Figura 3.2).

Pesaje de las trozas y de los haces de ramillas (paso 5)

Los pesajes de las trozas y de los haces de ramillas se realizan en el terreno (Foto 3.7) y al mismo tiempo, para asegurarse de que las mediciones para un árbol dado se efectuaron con la misma tasa de humedad. Resulta muy práctico usar una balanza de colgar amarrada a una pala cargadora. Los haces se colocan en la balanza y se mide la masa fresca. Las sogas y la lona de los haces se recuperan para volverlas a usar.

FOTO 3.7 – *Campaña de biomasa en un robledal. A la izquierda, las rodajas de muestra de un árbol, colocadas en una bolsa grande antes del transporte al área de pesaje de las muestras; centro: área de pesaje de las muestras; derecha, posicionamiento de la pala cargadora para pesar las trozas (Fotos: C. Nys).*

Toma de alícuotas (pasos 6 y 7)

Cuando se miden los haces de ramillas, se toman alícuotas de cada uno para estimar la tasa de humedad de las ramas. Es preferible tomar muestras de diferentes diámetros en distintas ramas para disponer de muestras representativas de la arquitectura de una rama estándar. En efecto, la muestra en una sola rama puede inducir un sesgo si estaba más húmeda o más seca que las otras. Las ramas se diferencian en cuatro grupos en función de su diámetro (clase 1: $0 < \varnothing \leq 4$ cm, clase 2: $4 < \varnothing \leq 7$ cm, clase 3: $7 < \varnothing \leq 20$ cm , y clase 4: $\varnothing > 20$ cm). Para las ramas de la clase 1, se toman muestras de aproximadamente 10 cm de largo. Para las otras clases, el principio es similar pero, al ser su diámetro mayor, se cortan rodajas en vez de pedazos de 10 cm de largo. Se toman aproximadamente 9, 6 y 3 rodajas para las clases 2, 3 y 4. Se trata de cifras indicativas pero que son el resultado de una síntesis de diferentes campañas realizadas en distintos ecosistemas. Las alícuotas se ponen en bolsas de papel preparadas para tal fin (y que antes se habían colocado al pie del árbol, véase el primer paso). Luego, se ponen esas bolsas de papel en una de plástico para un árbol dado para garantizar que las muestras no se mezclen con las de otros árboles.

Para evitar el sesgo del muestreo, es importante que sea siempre la misma persona la que tome las muestras y que lo haga en forma sistemática y representativa de la variabilidad de cada clase de tamaño de ramas. Para minimizar el sesgo asociado a la medición de la tasa de humedad, se transportan las muestras al área de pesaje (el mismo lugar que para las rodajas) y se las pesa en su bolsa de papel antes de tratarlas en el laboratorio. Si no

es posible pesar las muestras en el campo (lo que no se recomienda), habrá que limitar al máximo las pérdidas de humedad y por ello se recomienda mucho el uso de una hielera. Para la toma de muestras de las hojas convendrá mezclar bien todas las hojas y tomar la muestra al azar del medio del montículo así formado. Se recomienda efectuar esta operación de mezcla y muestreo cinco o seis veces para cada árbol (Foto 3.6). Las muestras de cada árbol se ponen en la misma bolsa (habrá que adaptar la cantidad en función del tamaño de las hojas y de su heterogeneidad, en especial la proporción de hojas verdes y de hojas senescentes — en general, una bolsa de plástico de tamaño regular es adecauada).

3.1.2. En el laboratorio

Si las rodajas del tronco no pueden pesarse inmediatamente, habrá que almacenarlas al aire libre y colocarlas sobre unos listones para que el aire circule entre ellas (para evitar el enmohecimiento). Si el pesaje de la biomasa fresca se efectuó en el campo, se las puede dejar secar libremente. Por el contrario, si no ha podido efectuarse dicho pesaje, conviene pesarlas de inmediato, apenas lleguen al laboratorio.

Para las alícuotas pesadas dentro de una bolsa en el campo, será necesario calcular la tara con una bolsa vacía (si fuera posible, habría que medir cada bolsa o, si estuviera demasiado deteriorada, reunir 10 o 20 en un mismo lote y contabilizar un peso correctivo promedio). Esta medición debe deducirse de los valores medidos en el campo. En caso de reemplazo de la bolsa para hacer secar las alícuotas, es indispensable registrar toda la información necesaria.

La temperatura de la cámara de secado debe fijarse en 70°C para secar las hojas, las flores y los frutos, o a 65°C si hay que efectuar análisis químicos sobre las alícuotas. Para las operaciones de biomasa y para la madera solamente, la temperatura será de 105°C. Para todas las categorías de muestras, se pesarán todos los días un mínimo de tres testigos hasta que se estabilice el peso. La estabilización demora en general dos días para las hojas y cerca de una semana para los elementos leñosos en función del tamaño de las muestras.

La Figura 3.3 representa el procedimiento que hay que utilizar para medir las muestras. Las mediciones en laboratorio comienzan pesando las muestras húmedas con su bolsa (medición de control con respecto al pesaje en el campo). En el caso de las rodajas de madera, si son demasiado grandes, es posible tomar submuestras. En ese caso es imperativo volver a pesar la rodaja entera y luego el trozo de la muestra. La pérdida de humedad entre el campo y la medición de la rodaja en el laboratorio se agrega a aquella medida en el laboratorio después de secada completamente la muestra. Si el intervalo de tiempo entre la fase de campo y la fase de laboratorio es considerable, olvidarse de esta etapa del protocolo puede originar errores muy grandes — hasta del 60–70 % — en la biomasa seca. El descortezado suele realizarse con una cuchilla especial para retirar la corteza o con un formón (Foto 3.8). Poner las rodajas en el congelador cuando están todavía húmedas puede facilitar a veces esta operación (por ejemplo en los robles). Luego se pesan las muestras de corteza y de madera y se ponen a secar en la cámara de secado (conviene evitar colocar demasiadas bolsas dentro de dicha cámara).

3.1.3. Los cálculos

Cálculo de la biomasa del tronco

Para cada troza i, se efectuó la medición de la circunferencia en ambos extremios: la circunferencia C_{1i} en el extremo delgado es la circunferencia de la rodaja que se cortó en el extremo más delgado y la circunferencia C_{2i} en el extremo grueso es la circunferencia de la

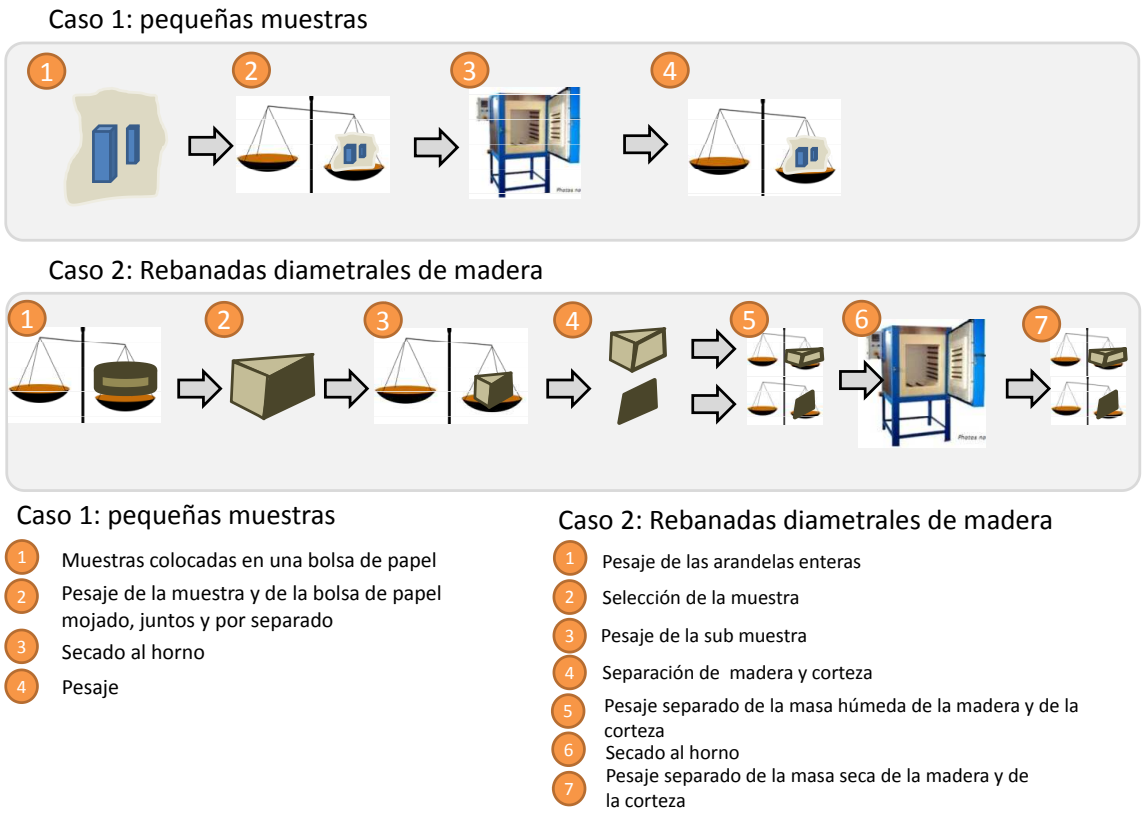


Figura 3.3 – *Procedimiento para pesar las muestras en el laboratorio.*

Foto 3.8 – *Mediciones en el laboratorio: (A) descortezado de las rodajas, (B) pesado de la madera, (C) pesado de la corteza (Fotos: L. Saint-André), (D) secado de las muestras, (E) pesado periódico hasta la estabilización del peso (Fotos: M. Henry).*

rodaja que se cortó en el extremo más grueso. Esto permite calcular el volumen de la troza fresca según la fórmula del cono truncado (o fórmula de Newton):

$$V_{\text{fresco},i} = L_i \times \frac{\pi}{3} \times (R_{1i}^2 + R_{1i}R_{2i} + R_{2i}^2) \tag{3.1}$$

donde L_i es la longitud de la troza i, y $R_{1i} = C_{1i}/(2\pi)$ y $R_{2i} = C_{2i}/(2\pi)$ son los radios de la troza i en sus dos extremos. Este volumen puede calcularse sobre la corteza (con las circunferencias medidas en el campo) o por debajo de ella (con las circunferencias medidas en las rodajas después del descortezado en el laboratorio). El volumen fresco debajo de la corteza es muy utilizado en la venta de madera mientras que la segunda medición permite controlar la coherencia de los datos al posibilitar el cálculo de la densidad de la madera en el árbol.

Cabe señalar que existen otras fórmulas para calcular el volumen de una troza. Las más usadas son la fórmula de Huber (basada en la circunferencia medida en el medio de la troza) y la de Smalian (basada en la media cuadrática de las circunferencias medidas en las partes superior e inferior de la troza). Pero en el caso en que la longitud de las trozas es escasa (1 o 2 m), la forma del tronco no se asemeja a la de un cono con ahusamiento muy poco acentuado y la diferencia entre ambas fórmulas es pequeña.

Además, para cada muestra tomada en la troza i se calcula:

- la proporción en biomasa fresca de la madera (sin corteza):

$$\omega_{\text{madera fresca},i} = \frac{B^{\text{alícuota}}_{\text{madera fresca},i}}{B^{\text{alícuota}}_{\text{madera fresca},i} + B^{\text{alícuota}}_{\text{corteza fresca},i}}$$

donde $B^{\text{alícuota}}_{\text{madera fresca},i}$ es la biomasa fresca de la madera (sin corteza) de la muestra de la troza i, y $B^{\text{alícuota}}_{\text{corteza fresca},i}$ es la biomasa fresca de la corteza de la muestra de la troza i;

- la tasa de humedad de la madera (sin corteza):

$$\chi_{\text{madera},i} = \frac{B^{\text{alícuota}}_{\text{madera seca},i}}{B^{\text{alícuota}}_{\text{madera fresca},i}} \tag{3.2}$$

donde $B^{\text{alícuota}}_{\text{madera seca},i}$ es la biomasa seca de la madera (sin corteza) de la muestra de la troza i;

- la proporción en biomasa fresca en la corteza:

$$\omega_{\text{corteza fresca},i} = 1 - \omega_{\text{madera fresca},i}$$

- la tasa de humedad de la corteza:

$$\chi_{\text{corteza},i} = \frac{B^{\text{alícuota}}_{\text{corteza seca},i}}{B^{\text{alícuota}}_{\text{corteza fresca},i}}$$

donde $B^{\text{alícuota}}_{\text{corteza seca},i}$ es la biomasa seca de la corteza de la muestra de la troza i.

A continuación se extrapolan las mediciones hechas en la muestra de la troza i a la troza i entera por regla de tres:

- la biomasa seca de la madera (sin corteza) de la troza i es:

$$B_{\text{madera seca},i} = B_{\text{fresca},i} \times \omega_{\text{madera fresca},i} \times \chi_{\text{madera},i}$$

donde $B_{\text{frais},i}$ es la biomasa fresca (con la corteza) de la troza i;

- la biomasa seca de la corteza de la troza i es:

$$B_{\text{corteza seca},i} = B_{\text{fresca},i} \times \omega_{\text{corteza fresca},i} \times \chi_{\text{corteza},i}$$

- la densidad de la madera de la troza i es:

$$\rho_i = \frac{B_{\text{madera seca},i}}{V_{\text{fresca},i}}$$

donde $V_{\text{fresca},i}$ es el volumen fresco sin corteza dado por la ecuación (3.1).

A continuación se suman los pesos secos de todas las trozas para obtener el peso seco del tronco:

- la biomasa seca de la madera (sin corteza) del tronco es:

$$B_{\text{madera seca tronco}} = \sum_i B_{\text{madera seca},i}$$

donde la suma se refiere a todas las trozas i que forman el tronco;

- la biomasa seca de la corteza del tronco es:

$$B_{\text{corteza seca tronco}} = \sum_i B_{\text{corteza seca},i}$$

La densidad de la madera ρ_i que interviene en el cálculo de la biomasa seca debe ser la densidad específica de la madera seca en cámara (en inglés: "ovendry wood density"), es decir la relación de la biomasa *seca* (secado en cámara hasta la estabilización del peso seco) sobre el volumen *fresco* de la madera. Hay que tener cuidado en no confundir esta densidad con la densidad volumétrica de la madera, que es la relación de masa sobre volumen, a igual tenor en humedad para la masa y el volumen (es decir, masa seca sobre volumen seco, o masa fresca sobre volumen fresco). Sin embargo, la norma AFNOR (1985) define de un modo diferente la densidad de la madera, como la relación de la biomasa secada al aire libre sobre el volumen de madera con un 12 % de humedad (Fournier-Djimbi, 1998). La densidad específica de la madera seca en cámara puede calcularse a partir de la densidad de la madera con un 12 % de humedad por la relación (Gourlet-Fleury *et al.*, 2011):

$$\rho_\chi = \frac{\rho\,(1+\chi)}{1-\eta\,(\chi_0-\chi)}$$

donde ρ_χ es la relación de la biomasa secada al aire libre sobre el volumen de la madera con χ % de humedad (en $\mathrm{g\,cm^{-3}}$), ρ es la relación de la biomasa secada en la cámara de secado sobre el volumen fresco de la madera (en $\mathrm{g\,cm^{-3}}$), η es el coeficiente de contraccíon volumétrica (número adimensional) y χ_0 es el punto de saturación de las fibras. Los coeficientes η y χ_0 varían de una especie a otra y obligan a conocer las propiedades tecnológicas de la madera de las especies. Al utilizar los datos de ρ y $\rho_{12\,\%}$ de 379 árboles, Reyes *et al.* (1992) determinaron además una relación empírica entre la densidad específica de la madera seca en cámara ρ y la densidad al 12 % de humedad $\rho_{12\,\%}$: $\rho = 0{,}0134 + 0{,}800\rho_{12\,\%}$ con un coeficiente de determinación $R^2 = 0{,}988$.

Cálculo de la biomasa de las hojas

Para cada muestra i de follaje tomada, se calcula la tasa de humedad del follaje:

$$\chi_{\text{hoja},i} = \frac{B^{\text{alícuota}}_{\text{hoja seca},i}}{B^{\text{alícuota}}_{\text{hoja fresca},i}}$$

donde $B^{\text{alícuota}}_{\text{hoja seca},i}$ es la biomasa seca del follaje de la muestra i, y $B^{\text{alícuota}}_{\text{hoja resca},i}$ es la biomasa fresca del follaje de la muestra i. Luego extrapolamos por regla de tres la muestra i al compartimiento i del que se ha extraído dicha muestra:

$$B_{\text{hoja seca},i} = B_{\text{hoja fresca},i} \times \chi_{\text{hoja},i}$$

donde $B_{\text{hoja seca},i}$ es la biomasa seca (calculada) del follaje del compartimiento i, y $B_{\text{hoja fresca},i}$ es la biomasa fresca (medida) del follaje del compartimiento i. Con frecuencia a la copa corresponde a un solo compartimiento. Pero, cuando la copa está dividida por partes, el peso seco total de las hojas se obtiene sumando todas las partes i:

$$B_{\text{hoja seca}} = \sum_i B_{\text{hoja seca},i}$$

Cálculo de la biomasa de las ramas

Cuando hay ramas muy grandes (por ejemplo > 20 cm de diámetro), hay que proceder como se hace con el tronco mientras que para las ramas hay que hacerlo del mismo modo que con las hojas.

Cálculo de la biomasa de frutos y flores

El método es idéntico al que se hace para las hojas.

3.2. Pesado directo para ciertos compartimientos y mediciones de volumen y de densidad para otros

El segundo caso que tomamos en cuenta es el que, debido a dificultades de corta, obliga a efectuar mediciones semidestructivas que combinan el pesado directo de ciertas partes del árbol y mediciones de volumen y de densidad para otras. Ilustraremos este caso construyendo una ecuación alométrica para bosques secos en el norte de Camerún. La evaluación de la biomasa de estos bosques resulta especialmente difícil debido a la complejidad de la arquitectura de los árboles. En las zonas secas, la intervención humana es particularmente significativa debido a la escasez de recursos forestales y a la importancia de la demanda bioenergética. Ésta se refleja en las prácticas de poda y mantenimiento de árboles con frecuencia situados en bosques abiertos, parques agroforestales o setos (Foto 3.9).

FOTO 3.9 – *Poda de árboles de butirospermos (Vitellaria paradoxa) en el norte de Camerún (Foto: R. Peltier).*

En la mayoría de las zonas secas los árboles están protegidos porque la regeneración de los recursos madereros es especialmente lenta y porque las actividades humanas la ponen en peligro. Deben preferirse las mediciones de biomasa no destructivas y hay que aprovechar las podas para medir la biomasa de las partes podadas del árbol. Las actividades de pastoreo limitan la regeneración y los árboles pequeños suelen estar poco representados. Así pues esta parte del manual considera sólo los árboles maduros.

3.2.1. En el campo: caso de las mediciones semidestructivas

Generalmente el tronco y las ramas grandes no se podan, sólo las pequeñas. La medición de la biomasa fresca (en kg) puede dividirse en dos partes: medición de la biomasa fresca

podada y medición de la biomasa fresca no podada (Figura 3.4A).

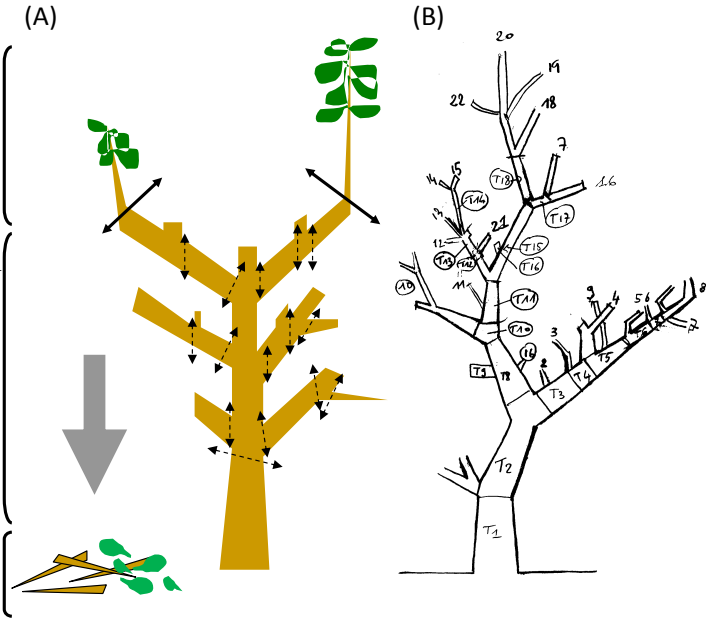


Figura 3.4 – *Determinación de la biomasa fresca total. (A) Separación y medición de la biomasa podada y sin podar, (B) numeración de los segmentos y de las ramas medidas en el árbol podado.*

Biomasa fresca podada

Las ramas pueden podarse siguiendo las prácticas locales (con frecuencia usando un machete). El diámetro de la base de cada rama podada se mide con una cinta métrica. Luego se separan las hojas y la madera de las ramas mondadas. La biomasa fresca de las hojas de las ramas mondadas ($B_{\text{hoja fresca podada}}$) y la biomasa fresca de las ramas podadas ($B_{\text{madera fresca podada}}$) se pesan por separado. La medición del peso se realiza con la ayuda de balanzas adecuadas. Si la masa de las hojas es inferior a dos kilogramos, es posible medir su peso con una balanza electrónica de campo.

Se toma una alícuota de hojas al azar de las ramas podadas. En general, hace falta un mínimo de tres muestras de hojas procedentes de tres ramas diferentes para formar la alícuota. Se mide su masa ($B^{\text{alícuota}}_{\text{hoja fresca}}$ en g). También se toma una alícuota al azar de la madera de las ramas podadas sin retirar la corteza. Su masa fresca ($B^{\text{alícuota}}_{\text{madera fresca}}$ en g) se mide en el campo justo después del corte. Las alícuotas se colocan en bolsas plásticas numeradas y se llevan al laboratorio. El volumen fresco de la alícuota de madera será medido posteriormente en el laboratorio (cf. § 3.2.2), lo que permitirá determinar la densidad media de la madera $\bar{\rho}$.

Biomasa fresca sin podar

La medición de la biomasa sin podar es indirecta dado que no es destructiva. Se determinan las diferentes ramificaciones del árbol podado y se numeran las ramas (Figura 3.4B). Las ramas pequeñas sin podar se tratan en forma diferente de las ramas grandes y del tronco (Figura 3.4A). Para las ramas pequeñas sin podar sólo se mide el diámetro en la base. La biomasa de las ramas pequeñas sin podar se estima a partir de la relación existente entre su diámetro en la base y su masa, como se explica en la Sección 3.2.3.

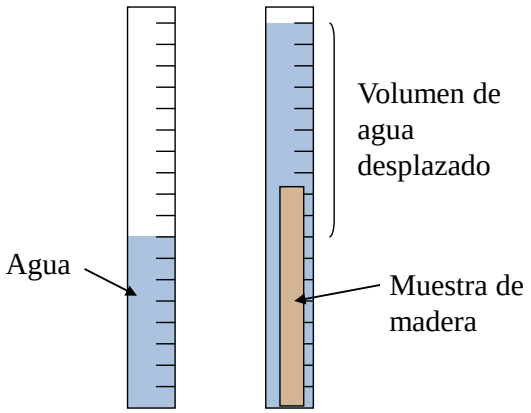


Figura 3.5 – *Medición del volumen de las muestras mediante el desplazamiento del volumen de agua.*

La biomasa del tronco y de las ramas grandes se estima a partir de las mediciones de volúmenes (V_i en cm^3) y de la densidad promedio de la madera ($\bar{\rho}$ en $\mathrm{g\,cm}^{-3}$). Las ramas grandes y el tronco del árbol podado se dividen virtualmente en segmentos mediante marcas hechas en el árbol. El volumen V_i de cada segmento i se obtiene a partir de la medición de su diámetro (o de su circunferencia) y de su longitud. Conviene tener una longitud para el segmento de aproximadamente un metro para poder considerar mejor las variaciones de diámetro a lo largo del tronco y de las ramas.

3.2.2. En el laboratorio

Primero se mide el volumen ($V^{\text{alícuota}}_{\text{madera fresca}}$) de la alícuota de madera extraída de los compartimientos podados. El volumen de madera puede medirse de diferentes formas (Maniatis *et al.*, 2011). El método más corriente usa el desplazamiento del volumen de agua provocado por la inmersión de la muestra en el agua. La medición del volumen de agua puede hacerse con una probeta adaptada al tamaño de la muestra (Figura 3.5). Otro método consiste en cortar las muestras para darles una forma cuyo volumen pueda medirse con la mayor precisión posible. Dicho método necesita instrumentos de precisión y personal entrenado para cortar la madera.

Las alícuotas de madera y de hojas se someten luego a las mismas mediciones en el laboratorio (secado en cámara, pesaje del peso seco, etc.) que las descritas en la Sección 3.1.2.

3.2.3. Los cálculos

La biomasa seca del árbol se obtiene a través de la suma de la biomasa seca podada y de la biomasa seca sin podar:

$$B_{\text{seca}} = B_{\text{seca podada}} + B_{\text{seca sin podar}}$$

Cálculo de la biomasa podada

A partir de la biomasa fresca $B^{\text{alícuota}}_{\text{madera fresca}}$ de la alícuota de madera y de su biomasa seca $B^{\text{alícuota}}_{\text{madera seca}}$, se calcula como antes (cf. ecuación 3.2) la tasa de humedad de la madera (con la corteza):

$$\chi_{\text{madera}} = \frac{B^{\text{alícuota}}_{\text{madera seca}}}{B^{\text{alícuota}}_{\text{madera fresca}}}$$

Del mismo modo se calcula la tasa de humedad del follaje a partir de la biomasa fresca $B^{\text{alícuota}}_{\text{hoja fresca}}$ de la alícuota de hojas y de su biomasa seca $B^{\text{alícuota}}_{\text{hoja seca}}$:

$$\chi_{\text{hoja}} = \frac{B^{\text{alícuota}}_{\text{hoja seca}}}{B^{\text{alícuota}}_{\text{hoja fresca}}}$$

Se puede calcular así la biomasa seca podada:

$$B_{\text{seca podada}} = B_{\text{madera fresca podada}} \times \chi_{\text{madera}} + B_{\text{hoja fresca podada}} \times \chi_{\text{hoja}}$$

donde $B_{\text{hoja fresca podada}}$ es la biomasa fresca de las hojas de las ramas podadas y $B_{\text{madera fresca podada}}$ la biomasa fresca de la madera de las ramas podadas.

Cálculo de la biomasa sin podar

Para la biomasa seca de la parte sin podar (la que queda en pie), se hacen dos cálculos: uno para las ramas pequeñas y otro para las grandes y el tronco. La biomasa sin podar es la resultante de sumar ambos:

$$B_{\text{seca sin podar}} = B_{\text{rama seca sin podar}} + B_{\text{seca segmento}}$$

Cada segmento i del tronco y de las ramas grandes puede considerarse como un cilindro cuyo volumen es (fórmula de Smalian):

$$V_i = \frac{\pi}{8} L_i \left(D_{1i}^2 + D_{2i}^2\right) \tag{3.3}$$

donde V_i es el volumen del i-ésimo segmento, L_i su longitud, y D_{1i} y D_{2i} los diámetros de los dos extremidos del segmento i. La fórmula del cono truncado (véase la ecuación 3.1) puede usarse también en lugar de la fórmula (3.3) del cilindro, pero habrá pequeñas diferencias entre ambos cálculos paso que el ahusamiento en un metro de largo no es muy pronunciado para los árboles.

La biomasa seca de las ramas grandes y del tronco se obtiene como el producto de la densidad media de la madera y del volumen total de las ramas grandes y del tronco:

$$B_{\text{seca segmento}} = \bar{\rho} \times \sum_i V_i \tag{3.4}$$

donde la suma se refiere al conjunto de segmentos que componen las ramas grandes y el tronco (Figura 3.4B), y donde la densidad promedio de la madera se calcula mediante:

$$\bar{\rho} = \frac{B^{\text{alícuota}}_{\text{madera seca}}}{V^{\text{alícuota}}_{\text{madera fresca}}}$$

Habrá que tener cuidado para que las unidades de medida sean consistentes. Por ejemplo, si la densidad media de la madera $\bar{\rho}$ en (3.4) se expresa en $\mathrm{g\,cm^{-3}}$, entonces el volumen V_i

debe expresarse en cm^3, lo que lleva a expresar tanto la longitud L_i y los diámetros D_{1i} y D_{2i} en cm. La biomasa en este caso se expresa entonces en g.

La biomasa seca de las ramas pequeñas sin podar se calcula mediante un modelo entre la biomasa seca y el diámetro basal. Para ello, se elabora el modelo siguiendo el mismo procedimiento que para la elaboración de un modelo alométrico (véanse los Capítulos 4 a 7 del manual). Las ecuaciones de potencia son frecuentemente usadas:

$$B_{\text{rama seca}} = a + bD^c$$

donde a, b y c son los parámetros del modelo y D el diámetro basal de la rama, pero pueden hacerse pruebas con otras regresiones (cf. Cuadro 5.1). Con un modelo de este tipo, la biomasa seca de las ramas pequeñas sin podar sería:

$$B_{\text{rama seca sin podar}} = \sum_j (a + bD_j^c)$$

donde la suma se refiere al conjunto de ramas pequeñas sin podar y D_j es el diámetro en la base de la j-ésima rama.

3.3. Pesado parcial en el campo

El tercer caso que prevemos es el de los árboles de dimensiones demasiado grandes para un pesaje completo a mano. Damos un ejemplo mediante la construcción de una ecuación alométrica para estimar la biomasa epigea de los árboles de un bosque tropical muy húmedo por medición destructiva. El método propuesto debe adaptarse a las circunstancias locales y a los medios disponibles. El valor comercial y la demanda de madera son dos factores que hay que tener en cuenta para las mediciones en las concesiones forestales.

Los árboles seleccionados se cortan siguiendo prácticas adecuadas. Una vez que se ha cortado el árbol, las variables como la altura total y la altura de los aletones (cuando el árbol los tiene) pueden medirse mediante una cinta métrica. Luego, se analiza la arquitectura del árbol (Figura 3.6). El enfoque propuesto separa a los árboles que pueden pesarse manualmente en el campo (por ejemplo, los árboles de un diámetro ≤ 20 cm) de aquellos que necesitan medios técnicos más consecuentes (los árboles de un diámetro > 20 cm).

3.3.1. Árboles con un diámetro inferior a 20 cm

Para los árboles de un diámetro ≤ 20 cm, se actúa de forma similar a la descrita en el primer ejemplo (§ 3.1). En primer lugar, se separan las ramas y el tronco. La biomasa fresca del tronco ($B_{\text{tronco fresco}}$) y de las ramas ($B_{\text{ramas fresca}}$, madera y hojas juntas) se miden con básculas adecuadas. Para medir la biomasa de las hojas, se selecciona al azar un número limitado de ramas para cada árbol. La biomasa fresca de las hojas ($B^{\text{muestra}}_{\text{hoja fresca}}$) y la biomasa fresca de la madera ($B^{\text{muestra}}_{\text{madera fresca}}$) de esta muestra de ramas se miden separadamente con básculas. La proporción foliar de las ramas se calcula entonces como:

$$\omega_{\text{hoja}} = \frac{B^{\text{muestra}}_{\text{hoja fresca}}}{B^{\text{muestra}}_{\text{hoja fresca}} + B^{\text{muestra}}_{\text{madera fresca}}}$$

Las biomasas frescas foliares ($B_{\text{hoja fresca}}$) y madereras ($B_{\text{madera fresca}}$) de las ramas se calculan luego a partir de esta proporción promedio de la hoja:

$$\begin{aligned} B_{\text{hoja fresca}} &= \omega_{\text{hoja}} \times B_{\text{rama fresca}} \\ B_{\text{madera fresca}} &= (1 - \omega_{\text{hoja}}) \times B_{\text{rama fresca}} \end{aligned}$$

A continuación se toman alícuotas de hojas y de madera a distintos niveles en las ramas y a lo largo del tronco. La biomasa fresca ($B^{\text{alícuota}}_{\text{hoja fresca}}$ y $B^{\text{alícuota}}_{\text{madera fresca}}$) de las alícuotas se mide con una balanza electrónica en el campo. Las alícuotas se llevan al laboratorio y son secadas y pesadas, según el mismo protocolo descrito en el primer ejemplo (§ 3.1.2). La biomasa seca ($B^{\text{alícuota}}_{\text{hoja seca}}$ y $B^{\text{alícuota}}_{\text{madera seca}}$) de las alícuotas permite calcular la tasa de humedad de las hojas y de la madera:

$$\chi_{\text{hoja}} = \frac{B^{\text{alícuota}}_{\text{hoja seca}}}{B^{\text{alícuota}}_{\text{hoja fresca}}}, \qquad \chi_{\text{madera}} = \frac{B^{\text{alícuota}}_{\text{madera seca}}}{B^{\text{alícuota}}_{\text{madera fresca}}}$$

Por último, la biomasa seca de hojas y maderera se obtiene a partir de su biomasa fresca y de las tasas de humedad calculadas a partir de las alícuotas. Para la biomasa de madera, se agregará la biomasa fresca de la madera de las ramas y del tronco:

$$\begin{aligned} B_{\text{hoja seca}} &= \chi_{\text{hoja}} \times B_{\text{hoja fresca}} \\ B_{\text{maderera seca}} &= \chi_{\text{bois}} \times (B_{\text{madera fresca}} + B_{\text{tronco fresco}}) \end{aligned}$$

La masa seca total se obtiene finalmente como la suma de la biomasa seca foliar y de la biomasa seca de madera:

$$B_{\text{seca}} = B_{\text{hoja seca}} + B_{\text{maderera seca}}$$

3.3.2. Árboles con diámetro superior a 20 cm

No resulta práctico separar las ramas del tronco cuando los árboles son demasiado grandes debido a la cantidad de ramas y follaje. El método alternativo propuesto aquí consiste en tratar de forma diferente el tronco y las ramas grandes (de diámetro basal superior a 10 cm) por un lado, y las ramas pequeñas (de diámetro basal inferior a 10 cm) por otro. Mientras que las ramas grandes de diámetro basal > 10 cm sólo están hechas de madera, las pequeñas con diámetro basal ≤ 10 cm pueden incluir también follaje. Las ramas grandes de diámetro basal > 10 cm se tratan del mismo modo que el tronco. La primera etapa por tanto consiste en dividirlas en secciones de madera. Mientras la biomasa de las secciones de diámetro superior a 10 cm se deduce de su volumen medido ($V_{\text{troza},i}$) y de la densidad media de la madera ($\bar{\rho}$), la biomasa de las ramas de diámetro basal ≤ 10 cm se estima a partir de una regresión entre su diámetro en la base y la biomasa que tienen.

Medición del volumen de las secciones de diámetro superior a 10 cm (tronco o rama)

Una vez que se han dividido en secciones el tronco y las ramas de diámetro basal > 10 cm el volumen de las secciones se calcula a partir de su longitud y de sus diámetros (o de sus circunferencias) en los dos extremos (D_{1i} y D_{2i}). Una longitud fija (por ejemplo, dos metros) puede usarse como estándar para cada sección (Foto 3.10A). En ciertos lugares habrá que usar una longitud de secciones menor que la determinada porque una bifurcación impide dar una forma cilíndrica a la troza. En ese caso, el técnico anota la longitud y los diámetros de cada una de las secciones. Luego elabora un esquema que representa la arquitectura del árbol (Figura 3.6). Este esquema resulta particularmente útil para el análisis de los resultados y su interpretación.

Los árboles de diámetro > 20 cm pueden tener aletones. El volumen de los aletones puede estimarse partiendo del supuesto de que su forma corresponde a una pirámide cuya arista superior es un cuarto de elipse (recuadro de la Figura 3.6; Henry *et al.*, 2010). Para

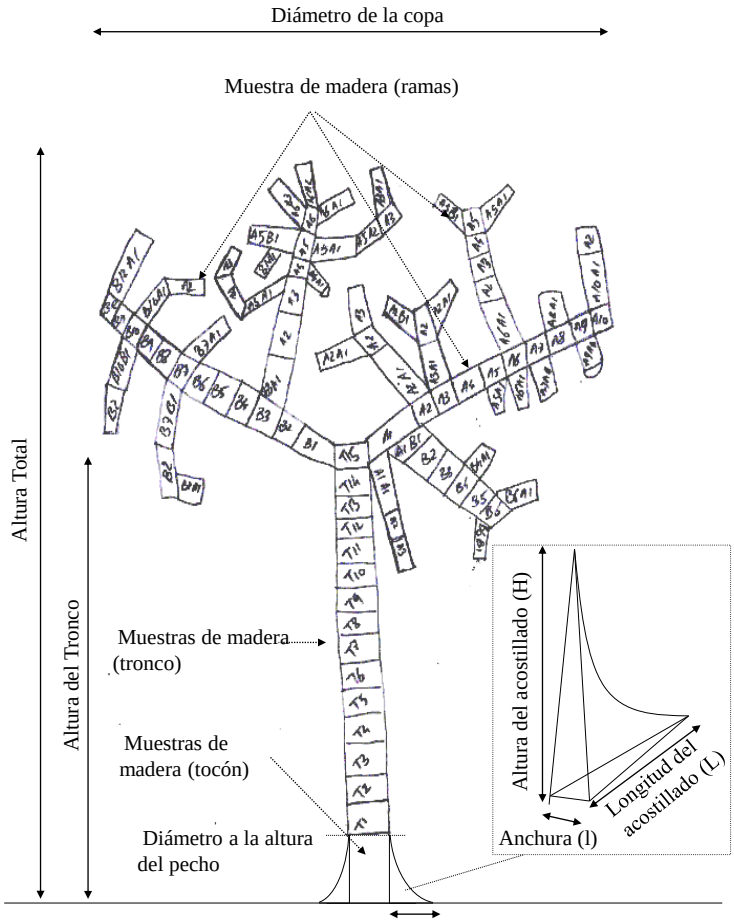


Figura 3.6 – *Esquema que representa las diferentes secciones de un árbol para el cálculo de su volumen.*

cada aletón j, se mide su altura H_j, su anchura l_j y su longitud L_j (recuadro de la Figura 3.6).

A continuación se toman alícuotas de madera en las distintas secciones de diámetro superior a 10 cm (tronco, ramas y aletones, de ser necesario; Foto 3.10B). Las alícuotas de madera fresca se colocan en bolsas herméticas y se las transporta hasta el laboratorio. En el laboratorio, se mide su volumen ($V^{\text{alícuota}}_{\text{madera fresca}}$) según el protocolo descrito en la Sección 3.2.2. Luego se secan y pesan las alícuotas de madera como se describe en la Sección 3.1.2, lo que permite obtener su biomasa seca ($B^{\text{alícuota}}_{\text{madera seca}}$).

Cálculo de la biomasa de las secciones de diámetro superior a 10 cm (tronco o rama)

Al igual que antes (cf. ecuación 3.3), el volumen $V_{\text{troza},i}$ de la i-ésima sección (tronco o rama de diámetro basal > 10 cm) se calcula mediante la fórmula de Smalian:

$$V_{\text{troza},i} = \frac{\pi \times L_i}{8}(D_{1i}^2 + D_{2i}^2)$$

donde L_i es la longitud de la i-ésima sección, D_{1i} es el diámetro de uno de sus extremos y D_{2i} es el diámetro de su otro extremo. Teniendo en cuenta su forma piramidal, una fórmula

Foto 3.10 – *Mediciones de un árbol grande en el campo: (A) medición del volumen de un árbol de diámetro* > 20 *cm, (B) toma de alícuotas de madera a nivel del tronco (Fotos: M. Henry).*

diferente se utiliza para calcular el volumen $V_{\text{aletón},j}$ del j-ésimo aletón:

$$V_{\text{aletón},j} = \left(1 - \frac{\pi}{4}\right) \frac{L_j H_j l_j}{3}$$

donde l_j es el ancho del j-ésimo aletón, L_j su longitud y H_j su altura.

Además, a partir de la biomasa seca y del volumen fresco de las alícuotas de madera, se puede calcular la densidad media de la madera:

$$\bar{\rho} = \frac{B^{\text{alícuota}}_{\text{madera seca}}}{V^{\text{alícuota}}_{\text{madera fresca}}}$$

La biomasa seca acumulada de las secciones (tronco y ramas de diámetro basal > 10 cm) es entonces:

$$B_{\text{seca secciones}} = \bar{\rho} \times \sum_i V_{\text{troza},i}$$

donde la suma se refiere al conjunto de las secciones, mientras que lal biomasa seca de los aletones es:

$$B_{\text{seca aletones}} = \bar{\rho} \times \sum_j V_{\text{aletón},j}$$

donde la suma se refiere al conjunto de los aletones. Como alternativa a la densidad media de la madera, se podrá utilizar una densidad de madera específica para cada parte del árbol (tronco, ramas, aletones). En este caso, la densidad media de la madera $\bar{\rho}$ se reemplazará en las fórmulas que figuran arriba por la densidad específica de cada compartimiento.

Medición de las ramas de diámetro inferior a 10 cm

Para todas las ramas de diámetro basal $\leq$ 10 cm, se mide el diámetro en la base. Su biomasa seca se puede estimarse a partir de una regresión entre el diámetro basal de la rama y la masa seca que contiene. Esta regresión se determina a partir de una muestra de ramas seleccionadas en el árbol con el objeto de representar las diferentes clases de diámetros en su base. Para cada rama de esta muestra, se separan las hojas de la madera. La biomasa fresca de las hojas ($B^{\text{muestra}}_{\text{hoja fresca},i}$ para la i-ésima rama) y la biomasa fresca de la madera ($B^{\text{muestra}}_{\text{madera fresca},i}$ para la i-ésima rama) de cada rama de la muestra se pesan separadamente en el campo.

Es posible que ciertas ramas tengan malformaciones y que no lleguen a una arquitectura ramificada. En ese caso, el volumen puede medirse y se registra la anomalía en las hojas de campo.

A continuación se toman alícuotas de madera y de hojas y de inmediato se pesa su biomasa fresca ($B^{\text{alícuota}}_{\text{madera fresca}}$ y $B^{\text{alícuota}}_{\text{hoja fresca}}$) en el campo. Las alícuotas se colocan en bolsas plásticas herméticas, se las transporta al laboratorio donde se las seca y pesa según el protocolo indicado en la Sección 3.1.2. Así se obtiene su biomasa seca ($B^{\text{alícuota}}_{\text{madera seca}}$ y $B^{\text{alícuota}}_{\text{hoja seca}}$).

Cálculo de la biomasa de las ramas de diámetro inferior a 10 cm

La biomasa fresca y seca de las alícuotas sirve para determinar el contenido de humedad de las hojas y de la madera:

$$\chi_{\text{hoja}} = \frac{B^{\text{alícuota}}_{\text{hoja seca}}}{B^{\text{alícuota}}_{\text{hoja fresca}}}, \qquad \chi_{\text{madera}} = \frac{B^{\text{alícuota}}_{\text{madera seca}}}{B^{\text{alícuota}}_{\text{madera fresca}}}$$

De ello se deduce, para cada rama i de la muestra de ramas, la biomasa seca de las hojas, la biomasa seca de la madera y luego la biomasa seca total de la rama i:

$$\begin{aligned} B^{\text{muestra}}_{\text{hoja seca},i} &= \chi_{\text{hoja}} \times B^{\text{muestra}}_{\text{hoja fresca},i} \\ B^{\text{muestra}}_{\text{madera seca},i} &= \chi_{\text{madera}} \times B^{\text{muestra}}_{\text{madera fresca},i} \\ B^{\text{muestra}}_{\text{rama seca},i} &= B^{\text{muestra}}_{\text{hoja seca},i} + B^{\text{muestra}}_{\text{madera seca},i} \end{aligned}$$

Como en la Sección 3.2.3, un modelo de biomasa para las ramas puede ser ajustado luego a los datos ($B^{\text{muestra}}_{\text{rama seca},i}$, D^{muestra}_i), donde D^{muestra}_i es el diámetro en la base de la i-ésima rama de la muestra. El modelo de biomasa para las ramas se determina siguiendo el mismo procedimiento que para la elaboración de una ecuación alométrica (véanse los Capítulos 4 a 7 del manual). Para aumentar el tamaño de la muestra, el modelo podrá determinarse a partir de todas las ramas medidas para todos los árboles de la misma especie o por grupos funcionales de especies (Hawthorne, 1995).

Al utilizar el modelo de biomasa para las ramas así determinado se puede calcular la biomasa seca de las ramas de diámetro basal $\leq$ 10 cm:

$$B_{\text{rama seca}} = \sum_i f(D_i)$$

donde la suma se refiere al conjunto de ramas de diámetro basal $\leq$ 10 cm, D_i es el diámetro basal de la i-ésima rama y f es el modelo de biomasa que predice la biomasa seca de una rama en función de su diámetro basal.

Cálculo de la biomasa del árbol

La biomasa seca del árbol se obtiene sumando la biomasa seca de las secciones (tronco y ramas de diámetro basal > 10 cm), la biomasa seca de los aletones y la biomasa seca de las ramas de diámetro ≤ 10 cm:

$$B_{\text{seca}} = B_{\text{seca secciones}} + B_{\text{seca aletones}} + B_{\text{rama seca}}$$

3.4. Mediciones radiculares

Las mediciones de la biomasa de las raíces son mucho más difíciles de realizar que las de la biomasa epigea. Los métodos que proponemos aquí son el resultado de campañas realizadas en diferentes ecosistemas y fueron objeto de un estudio comparativo en el Congo (Levillain *et al.*, 2011).

La primera etapa, independientemente del ecosistema, consiste en hacer un diagrama de Voronoi[1] alrededor del árbol seleccionado. La Figura 3.7 indica el proceso que se debe seguir: (*i*) trazar los segmentos que conectan el árbol seleccionado con cada uno de sus vecinos; (*ii*) trazar las mediatrices de cada segmento, (*iii*) unir las mediatrices entre sí para delimitar un espacio alrededor del árbol; (*iv*) a continuación puede dividirse dicho espacio en triángulos unidos por los bordes, siendo fácil de calcular la superficie de cada zona utilizando la fórmula del triángulo y conociendo las longitudes de sus tres lados (a, b y c):

$$A = \sqrt{p(p-2a)(p-2b)(p-2c)}$$

donde $p = a + b + c$ es el perímetro del triángulo y A su superficie. La Figura 3.8 ilustra este proceso para las plantaciones de cocoteros en Vanuatu (Navarro *et al.*, 2008).

El espacio así delimitado no constituye una materialización del espacio "vital" del árbol. Se trata solamente de una forma de separar el espacio en zonas unidas para facilitar luego el muestreo de la biomasa subterránea. La hipótesis principal es que las raíces de otro árbol que vienen a colonizar este espacio compensan a aquellas que salen de allí y que pertenecen al árbol seleccionado.

En el caso de masas pluriespecíficas o agroforestales, a veces resulta difícil, incluso imposible, separar las raíces de las diferentes especies. En este caso, sería muy arriesgado elaborar modelos individuales (la biomasa radicular asociada al árbol elegido para el muestreo) pero las estimaciones de la biomasa de las raíces por hectárea, sin distinción de especies, seguirá siendo perfectamente válida.

Los métodos de muestreo varían en función del grosor de las raíces. Levillain *et al.* (2011) efectuaron un estudio que compara los distintos métodos en el mismo árbol (Foto 3.11). Muestran que es más rentable, en términos de costo-precisión, muestrear raíces finas con cilindros de muestreo, al tiempo que las raíces medias necesitan una excavación parcial y las grandes una total del espacio de Voronoi.

El número de cilindros de muestreo y la dimensión de la fosa que hay que excavar varían de un ecosistema a otro. En el Congo, en las plantaciones de eucalipto, el número óptimo de cilindros para obtener una precisión de 10 % es de unos 300 en la superficie (0–10 cm) y de 100 para las raíces más profundas (10–50 y 50–100 cm). Para obtener esta precisión en 1 m de profundidad hacen falta 36 hombre·dias de trabajo. Por el contrario, si se desea una precisión de sólo el 30 % el tiempo necesario para el muestreo disminuye en 75 %. Este

[1] Un diagrama de Voronoi (también llamado descomposición de Voronoi, partición de Voronoi o polígonos de Voronoi) representa una descomposición particular de un espacio métrico determinado por las distancias a un conjunto discreto de objetos del espacio, en general un conjunto discreto de puntos.

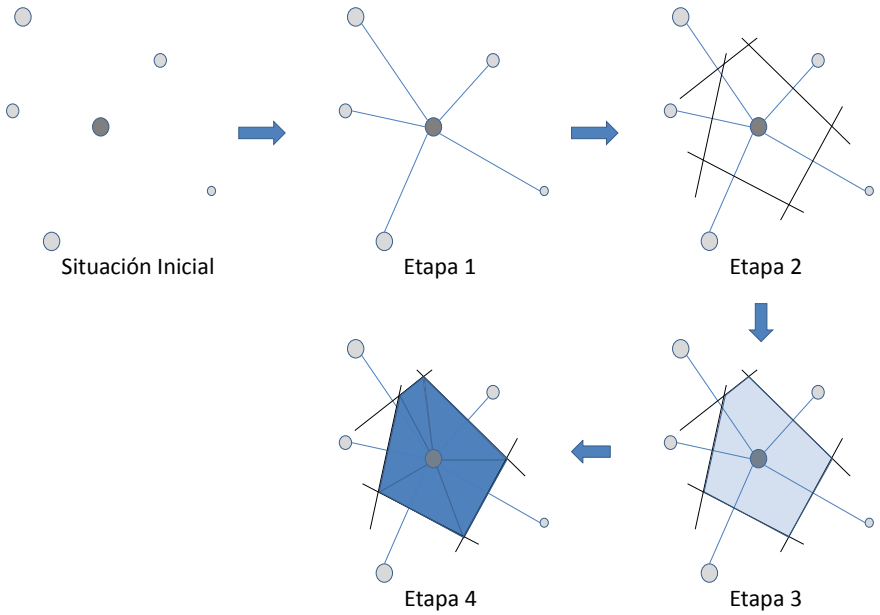


Figura 3.7 – *Método para delimitar un espacio de Voronoi y sus subdivisiones alrededor de un árbol y en una situación de vecindad cualquiera.*

ejemplo ilustra perfectamente la utilidad de hacer un premuestreo (cf. Capítulo 2) para evaluar la variabilidad en el ecosistema estudiado y luego adaptar el protocolo en función de los objetivos y de la precisión deseada.

Una vez que se han tomado muestras del suelo con las raíces, la selección puede hacerse en el laboratorio para los cilindros que contienen raíces finas. Por el contrario, para las raíces medianas y gruesas, hay que hacer la selección en el campo dado el volumen y el peso de la tierra excavada. En el laboratorio se lava el suelo teniendo cuidado en poner un filtro para poder recuperar después la raíces que flotan y/o se recuperan usando un tamiz. En el campo las muestras se separan del suelo manualmente sobre lonas. Para las raíces gruesas y las medianas se puede usar un compresor de aire que permite excavar completamente el sistema radicular conservando su arquitectura. Este método, paricularmente conveniente en suelos arenosos, permite satisfacer dos objetivos (biomasa y arquitectura) pero hace falta, no obstante, disponer de un compresor móvil en el lugar de trabajo (Foto 3.12).

Una vez clasificadas y recolectadas, las raíces se ponen a secar siguiendo los mismos principios que para la biomasa epigea. Las raíces finas necesitarán, en general, el mismo tiempo de secado que las hojas mientras que para las raíces medianas y gruesas serán necesarios más bien tiempos equivalentes a los de las ramas.

Para el tocón habrá que tomar una submuestra, de preferencia vertical, para tener mejor en cuenta las variaciones de densidad de la madera de esta parte del árbol. Se debe seguir los mismos procedimientos que para las rodajas del tronco.

Los cálculos que hay que hacer luego son los mismos que para la biomasa epigea.

Foto 3.11 – *A la izquierda, combinación de los métodos de muestreo (cilindros, excavaciones por cubos, excavación parcial de Voronoi, excavación total de Voronoi, según Levillain et al. (2011) (Foto: C. Jourdan). A la derecha, excavación manual de las raíces gruesas en una plantación de caucho en Tailandia (Foto: L. Saint-André).*

Foto 3.12 – *Utilización de un compresor de aire en el Congo para la extracción de los sistemas radiculares (raíces grandes y medianas) de eucaliptos. A la izquierda, el operador con los equipos de seguridad (protección contra el polvo y el ruido); en el centro, el compresor y una imagen ampliada del medidor que indica la presión del aire (unos 8 bares); a la derecha, el resultado (Fotos: C. Jourdan).*

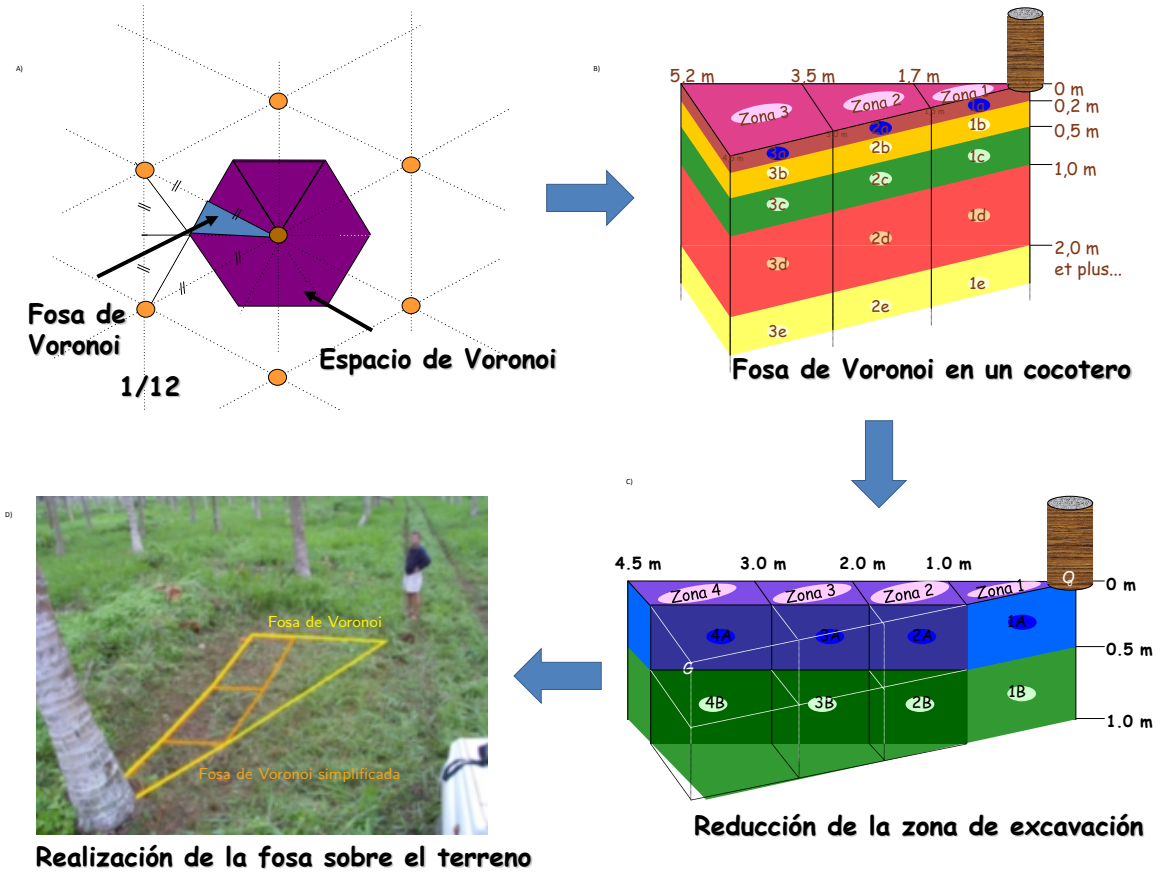


Figura 3.8 – *Ejemplo de división del espacio de Voronoi para el muestreo de las raíces en una plantación de cocoteros en Vanuatu (Foto: C. Jourdan). (A) División del espacio de Voronoi y decisión de trabajar sobre 1/12-ésimo de dicho espacio; (B) corte esquemático de las fosas realizadas; (C) simplificación del protocolo teniendo en cuenta la variabilidad observada en un primer muestreo; (D) materialización de los trazados en un caso real.*

3.5. Equipo recomendado

3.5.1. Material pesado y vehículos

- Automóviles, camiones, remolque: transporte de personas, de material y de muestras desde/hasta el laboratorio.
- Cuatriciclo (de ser posible): transporte de material voluminoso y de muestras en el campo.
- Pala cargadora para el pesar los haces de ramas.

3.5.2. Material básico

- Herramientas básicas en su caja
- Cajas de plástico (almacenamiento y transporte del material — aproximadamente 10).
- Bolsas de plástico (calcular una o dos grandes por árbol) para reunir las muestras de un árbol y evitar las pérdidas de humedad. Bolsas de papel (calcular una por

cada compartimiento y por cada árbol) para poner las muestras justo después de tomarlas. Lo ideal es que las bolsas ya estén marcadas con los números de árbol y de compartimiento (lo que permite ganar tiempo en el campo). También hay que prever llevar bolsas no marcadas y un rotulador negro con tinta indeleble para corregir los errores posibles o permitir tomar muestras adicionales.

- Grandes lonas para la copa (ya sea cortadas para atar los haces de ramas, o extendidas sobre el suelo para recuperar las hojas arrancadas de los árboles).
- Etiquetas (para engrapar en las rodajas), grapas y una engrapadora; o un lápiz de fucsina (si las muestras deben conservarse luego para medir la mineralomasa, hay que evitar la fucsina) (Foto 3.13).
- Cúteres, machetes, hachas, tijeras podadoras y sierras (Foto 3.13).
- Marcos para armar los haces de ramas (Foto 3.14) o bien contenedores de basura de distintos tamaños.
- Motosierra (lo ideal es una motosierra adecuada para la corta de árboles y otra, más pequeña y manejable, para cortar las ramas — Foto 3.13).
- Sogas resistentes para atar los haces de ramas (que se volverán a usar durante la campaña, por lo que hace falta hacer nudos reversibles).
- Bolsas grandes muy resistentes (tipo sacos o costales para cereales, arena o fertilizante — Foto 3.15) para transportar rodajas y muestras del campo a los vehículos (si éstos se encuentran lejos del área de corta).

Foto 3.13 – *Material de campo. A la izquierda, material para cortar las alícuotas y etiquetarlas; en el centro, ejemplos de plantillas para cortar las ramas; a la derecha, motosierra y equipo de seguridad (Foto: A. Genet).*

3.5.3. Ingreso de datos de campo usando computadoras

- Un ordenador de bolsillo (con cargador de batería y cables) o formularios de campo en papel impermeable o cartón, de ser posible encuadernadas y con tapas plastificadas.
- Flora o clave de determinación de las especies para los trabajos en bosques tropicales muy húmedos.
- Lápices de tipo 2B, gomas, sacapuntas.

FOTO 3.14 – *Atado de haces. A la izquierda, marco de hierro, lona y soga para atar las ramas (Foto: A. Genet); en el centro, la preparación de los haces en el campo (Foto: M. Rivoire); a la derecha, el haz listo para ser pesado (Foto: M. Rivoire).*

FOTO 3.15 – *Transportes de las rodajas y de las alícuotas en un costal para arena o cereales (Foto: J.-F. Picard).*

- Balanzas de campo o básculas (con 2 juego de baterías y cargador) para pesar las muestras (precisión de 1 g). Lo ideal es disponer de una gama completa adaptada al peso de las muestras (una troza de 1 o 2 m puede pesar centenares de kg mientras que las rodajas de madera van de algunas decenas de g a varias decenas de kg). El uso de una pala cargadora permite facilitar el pesado en el campo de las trozas grandes. Para ello hay que prever correas para atar las balanzas a la base y ganchos que se bloqueen automáticamente para enganchar la troza.

- Decámetro para medir las alturas a lo largo del tronco (perfiles de tronco).

- Forcípula y cinta para medir la circunferencia, o cinta diamétrica.

- Pintura en aerosol para marcar árboles (marcado de árboles en pie y marcado del fuste principal en las copas muy desarrolladas).

- Gancho de marcar para indicar los lugares donde se cortarán las rodajas (o marcado con la pintura en aerosol).

3.5.4. Equipo de laboratorio

- Cámaras de secado.

- Probeta de 500 ml como mínimo.

- Cuchilla descortezadora.

- Tijeras de podar.

- Balanza con una capacidad de 2 a 2000 g (precisión de 0,1 g a 1 g).

- Sierra sin fin.

3.6. Recomendación para la composición de los equipos de campo

Equipo de corta: un leñador, dos ayudantes de leñador, dos personas (para limpiar el área antes de la corta). Calcular dos días para cortar 40 árboles con circunferencias entre 31 y 290 cm (promedio 140 cm). Es posible cortar todos los árboles al comienzo de la campaña para liberar luego a este equipo que se ocupará de desramar los árboles. Para que puedan ser operacionales sin tiempos muertos, hace falta que una decena de árboles (de 20 metros de altura — o sea, unas 10 a 20 rodajas por árbol) estén listos para ser troceados en el área de trabajo.

Equipo de perfiles de fuste: dos personas (un encargado sujetar la cinta y otro de las mediciones). Este equipo empieza a trabajar apenas cortado el árbol y, por tanto, sigue al equipo de leñadores. En general ambos equipos son bastante sincrónicos. El equipo de perfiles de fuste nunca se queda esperando a que terminen los leñadores, salvo en casos muy poco frecuentes cuando hay problemas con la corta (por ejemplo, para fustes muy grandes o para troncos enredados en otros árboles y que hay que liberar).

Equipo de desrame: tres personas por unidad de trabajo. Cada unidad comprende un sierrista (con una sierra manejable) y dos agavilladores. Estos equipos pueden duplicarse o triplicarse en función de la dimensión de la copa que hay que desramar. Para hacerse una idea: para un árbole con 200 cm de circunferencia a la altura del pecho, se necesitan tres unidades; entre 80 y 200 cm de circonférence, de circunferencia, hacen falta dos; y por debajo de 80 cm de circunferencia, basta con una. Estas magnitudes tienen en cuenta el hecho de que las unidades no deben interferir entre ellas mientras trabajan. Si hay tres en el mismo árbol, hace falta que una esté en la parte baja del árbol y que suba a lo largo del tronco, mientras que las otras dos están una a cada lado del eje principal y que partan del medio de la copa aproximadamente para subir hacia la parte superior del árbol.

Equipo de troceado: incluye un leñador (para cortar los fustes) y otra persona (encargada de etiquetar las rodajas). En general, es el equipo de corta que se hace cargo de esta actividad al terminar con esa labor. Una vez que se han cortado todos los árboles, el leñador se incorpora a este equipo de troceado y los ayudantes de leñador se suman a aquellos de desrame.

Equipo de pesaje: tres personas (conductor de la pala cargadora y otras dos personas para el transporte de las trozas y los haces).

Equipo de muestreo de las ramas: una o dos personas.

4

Ingreso y estructura de los datos

Después de la fase de mediciones de campo y antes de la fase de análisis de los datos, viene la fase de estructuración de los mismos, que incluye su ingreso, la verificación de su exactitud y su formato.

4.1. Ingreso de los datos

El ingreso de los datos consiste en transferir a un archivo informático los datos que figuran en las fichas de campo. Para ello habrá que elegir con antelación un software adecuado. Si se trata de un conjunto de datos pequeño, bastará con una hoja electrónica tipo Microsoft Excel u OpenOffice Calc. Para iniciativas mayores, habrá que usar un sistema de gestión de base de datos, por ejemplo, Microsoft Access o MySQL (www.mysql.com).

4.1.1. Errores en el ingreso de los datos

El ingreso de los datos debe hacerse con todo el cuidado posible para limitar los errores. Una forma de lograrlo es hacer un registro doble: el primer operador ingresa la información y el segundo (de preferencia otra persona) vuelve a ingresar los datos en forma totalmente independiente. De ese modo, basta con comparar los archivos para descubrir los errores cometidos al ingresar los datos. Como es poco probable que ambos operadores cometan el mismo error, este método garantiza un ingreso de datos de buena calidad. La desventaja es que lleva mucho tiempo y resulta una tarea fastidiosa.

Al registrar los datos hay que prestar atención a ciertos detalles importantes. Primero, diferenciar los números de las cadenas de caracteres. Para el software estadístico que se encargará luego del tratamiento de los datos, un número no desempeña la misma función que una cadena de caracteres, por lo cual es importante hacer esa distinción desde un principio. Un número se interpretará como el valor de una variable numérica mientras que una cadena de caracteres será considerada como una categoría de una variable cualitativa. La diferencia entre ambos, en general, está obiva aunque no siempre es así. Por ejemplo, consideremos el caso de la latitud y la longitud. Si se desea calcular la correlación entre la latitud o la longitud y otra variable (para identificar un gradiente norte-sur o este-oeste), hace falta que el software considere las coordenadas geográficas como números. Por ello no hay que

registrar las coordenadas geográficas como, por ejemplo, "7°28′55,1″" o "13°41′25,9″". Esas coordenadas serían interpretadas como variables cualitativas y no se podría realizar ningún cálculo. Una solución posible consiste en convertir las coordenadas geográficas en valores decimales. Otra solución es registrar las coordenadas geográficas en tres columnas (una para los grados, otra para los minutos y la tercera para los segundos).

Cuando se ingresan variables cualitativas, hay que evitar cadenas de caracteres muy largas porque eso multiplica los riesgos de cometer errores. Es mejor ingresar un código abreviado y precisar en la metainformación (cf. a continuación) lo que significa ese código.

Otro detalle que tiene su importancia es el símbolo decimal utilizado. Prácticamente todos los paquetes de software estadístico permiten pasar de la coma decimal (símbolo utilizado en español) al punto (símbolo utilizado en inglés), así que resulta indiferente usar uno u otro. Por el contrario, una vez que se ha escogido la coma o el punto como símbolo decimal, hay que respetarlo a lo largo de todo el proceso de registro de los datos. Si se usa una vez uno y una vez otro, una parte de los datos normalmente numéricos será interpretada por el software estadístico como cadenas de caracteres.

4.1.2. La metainformación

Durante el ingreso de los datos hay que pensar también en la metainformación. Se trata de la información que acompaña los datos, sin ser por sí misma un dato medido. La metainformación dará, por ejemplo, la fecha en que se efectuaron las mediciones y quién las hizo. Si se utilizan códigos en el ingreso de datos, la metainformación indicará su significado. Por ejemplo, es frecuente que los nombres de las especies se ingresen en forma abreviada. Un código de especie de tipo ANO en una sabana seca de África occidental, por ejemplo, resulta ambiguo: puede tratarse de *Annona senegalensis* o de *Anogeissus leiocarpus*. La metainformación sirve para eliminar dicha ambigüedad. Asimismo debe precisar el carácter de las variables medidas. Por ejemplo, si se mide el diámetro de los árboles, no basta con anotar "diámetro" en el cuadro de los datos. La metainformación debe indicar a qué altura se midió (en la base, a 20 cm, a 1,30 m, etc.) y, algo sumamente importante, en qué unidad está expresado (en cm, en dm, etc.). Insistimos en el hecho de que la unidad de medida de cada una de las variables debe precisarse en la metainformación. Con demasiada frecuencia encontramos cuadros de datos que no indican las unidades en las cuáles se registraron, lo que da lugar a toda una serie de suposiciones arriesgadas.

Para la persona que concibió el dispositivo de medición y se encargó de supervisar las mediciones, la información contenida en la metainformación suele ser tan evidente que no ve la necesidad de perder tiempo en ingresarla. No obstante, hay que pensar en una persona ajena al proceso que se encuentra con ese conjunto de datos diez años más tarde. Si la metainformación se hizo bien, la persona podrá trabajar con esos datos como si fuera ella quien los hubiera elaborado.

4.1.3. Niveles anidados

Los datos se ingresan en las hojas de cálculo, con un renglón por cada individuo. Si los datos incluyen varios niveles anidados, debe haber tantos cuadros de datos como niveles. Supongamos que tratamos de construir un cuadro para las formaciones de tipo monte medio a escala de una región. Para los individuos multicaules de la muestra, se calcula el volumen de cada fuste por separado. Se seleccionan los individuos en las parcelas, a su vez seleccionadas en los macizos forestales distribuidos en la zona de estudio. En este caso se trata de cuatro niveles anidados: el rodal, que incluye varias parcelas; la parcela, que incluye diversos árboles; el árbol, que incluye varios fustes; y por último, el tronco. En este caso tendrá que haber

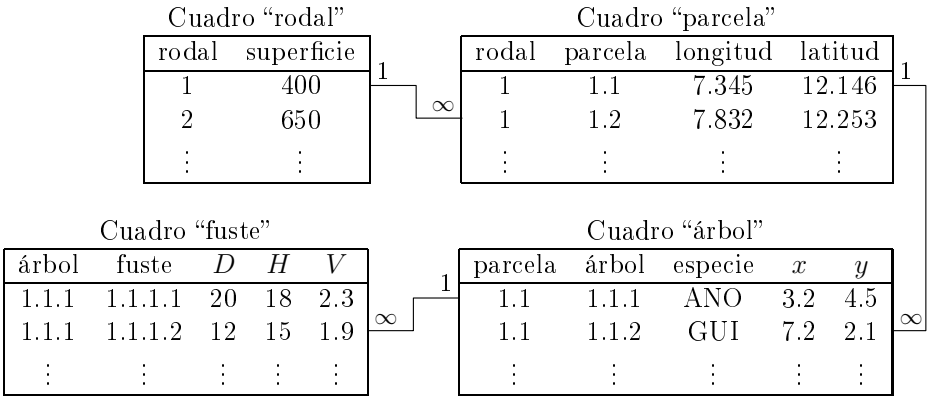


Figura 4.1 – *Ejemplo de cuatro cuadros de datos para cuatro niveles anidados.*

Cuadro 4.1 – *Registro de los datos con cuatro niveles anidados en un cuadro único.*

rodal	superficie	parcela	longitud	latitud	árbol	especie	x	y	fuste	D	H	V
1	400	1.1	7.345	12.146	1.1.1	ANO	3.2	4.5	1.1.1.1	20	18	2.3
1	40*1*	1.1	7.345	12.146	1.1.1	ANO	3.2	4.5	1.1.1.2	12	15	1.9
1	400	1.1	7.345	12.146	1.1.2	GUI	7.2	2.1	⋮	⋮	⋮	⋮
1	400	1.2	7.832	12.253	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
2	650	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

pues cuatro cuadros de datos (Figura 4.1). Cada uno reflejará las variables que describen a los individuos del nivel correspondiente, con un renglón del cuadro para cada individuo. Por ejemplo, el primer cuadro dará la superficie de cada uno de los macizos forestales. El segundo dará las coordenadas geográficas de cada una de las parcelas. El tercero dará la especie y las coordenadas de cada árbol dentro de la parcela. Por último, el cuarto dará el volumen y el tamaño de cada tronco. A cada renglón de un cuadro corresponden varios del cuadro del nivel inferior. Un identificador debe permitir establecer la correspondencia entre los renglones de los distintos cuadros. Así pues, el número del rodal se repetirá en los cuadros “rodal” y “parcela”, el número de la parcela se repetirá en los cuadros “parcela” y “árbol”, y el número del árbol se repetirá en los cuadros “árbol” y “fuste” (Figura 4.1).

Esta estructuración de los datos minimiza la repetición de la información y, con ello, los errores de ingreso. Una alternativa sería ingresar todos los datos en el mismo cuadro, como se indica en el ejemplo anterior en el Cuadro 4.1. No se recomienda esta alternativa porque repite inútilmente la información, multiplicando pues los riesgos de error de registro. Por ejemplo, en el Cuadro 4.1 introdujimos voluntariamente un error de registro en el segundo renglón del Cuadro, donde la superficie del rodal 1, normalmente igual a 400 ha, aquí es de 401 ha. Al repetir inútilmente la información, se multiplican este tipo de incongruencias que luego hay que corregir.

Una buena forma de resolver estos problemas de niveles anidados es construir una base de datos relacional. Este tipo de bases de datos se construyen justamente para gestionar los distintos cuadros con las relaciones que los unen entre sí. Permiten eliminar cualquier incongruencia como la ilustrada en el Cuadro 4.1 al verificar sistemáticamente la integridad

de las relaciones entre los cuadros. Sin embargo, la construcción de una base de datos relacional es una etapa técnica que exige a veces recurrir a una persona competente en ese ámbito.

Recapitulando, al ingresar datos, es preferible:

- evitar repetir la misma información,
- preferir las bases de datos relacionales,
- dar información adicional (metainformación),
- prestar atención a las unidades,
- establecer la diferencia entre la información cualitativa y aquella cuantitativa,
- verificar los datos,
- reducir o corregir los datos faltantes.

4.2. Verificación de los datos

La verificación exige que se comparen los formularios de campo contra la información del software estadístico (o eventualmente, un programa especialmente concebido para la verificación de los datos). Esta etapa sirve para eliminar cualquier incongruencia en los datos. Eventualmente, si todavía existe el dispositivo de medición, habrá que volver a efectuar algunas mediciones. La verificación permitirá eliminar:

- los datos aberrantes. Por ejemplo, un árbol de 50 metros de diámetro.
- los datos incoherentes. Por ejemplo, un árbol con una biomasa del tronco de 755 kg y una biomasa total de 440 kg, o bien un árbol de 5 cm de diámetro y de 40 m de altura.
- las modalidades falsas de las variables cualitativas. Por ejemplo, un software que hace la diferencia entre las mayúsculas y las minúsculas interpretará “sí” y “Sí” como dos categorías diferentes, cuando en realidad se trata de la misma.

La dificultad para detectar los datos aberrantes proviene de la elección del umbral entre lo que es una medición normal y lo que constituye una aberrante. Suele ocurrir que los datos aberrantes sean el resultado de un cambio de unidad durante el registro de los datos. Si la ficha de campo dice 1,2 kg y luego 900 g para las mediciones de biomasa foliar, habrá que tener cuidado al registrar 1,2 y 0,9 (en kg), o bien 1200 y 900 (en g), pero en ningún caso habrá que registrar 1,2 y 900. Los datos incongruentes son más difíciles de detectar porque hay que comparar varias variables entre sí. En el ejemplo anterior, un árbol con una biomasa del tronco de 755 kg no tiene nada de anormal y un árbol con una biomasa total de 440 kg tampoco pero, desde luego, ambas medidas no pueden ser correctas si se trata del mismo árbol. Del mismo modo, un árbol con un diámetro de 5 cm no tiene nada de anormal como tampoco lo tiene un árbol que mida 40 m de altura pero lo que es anormal es que haya un árbol de 5 cm de diámetro con una altura de 40 m.

La detección de los datos aberrantes e incongruentes podrá efectuarse con estadísticas descriptivas y gráficos que comparen dos variables al mismo tiempo: el examen de las medias, los cuantiles, los valores máximos y mínimos permiten con frecuencia detectar los datos aberrantes; los gráficos de dos variables permiten detectar incongruencias de los datos. En

el ejemplo anterior, se podrá hacer el gráfico de la biomasa total en función de la biomasa del tronco, y verificar que todos los puntos se sitúen por encima de la recta $y = x$. Los gráficos altura en función del diámetro, volumen en función del diámetro, etc., permiten detectar también los datos anormales. Las categorías de las variables cualitativas podrán inspeccionarse sacando la cuenta del número de observaciones por categoría. Dos variables cualitativas podrán cruzarse para elaborar la tabla de contingencia correspondiente. Durante esta inspección de los datos habrá que cerciorarse que el software estadístico ha interpretado los datos numéricos y los datos cualitativos como corresponde.

Las categorías falsas suelen resultar de una falta de rigor en el ingreso de los datos. Los errores de ortografía involuntarios, frecuentes cuando los nombres de las especies se escriben completos, por ejemplo, dan lugar a categorías falsas. Éstas pueden ser muy ambiguas y difíciles de corregir. Tomemos el ejemplo de un conjunto de datos sobre árboles del bosque tropical húmedo de África central, donde, entre otras, hay dos especies mubangu (*alombi* en francés) (*Julbernardia seretii*) y calabó (*ilomba* en francés) (*Picnanthus angolensis*). Supongamos que, por error, la categoría falsa "alomba" fue registrada por el técnico francés. Esta falsa categoría no se diferencia de las verdaderas alombi e ilomba más que en una letra. ¿Cómo saber cuál es la verdadera categoría? Los acentos suelen dar lugar también a categorías falsas, según se haya registrado el texto con o sin acentos. Tomemos el ejemplo del registro de un color: para la persona que registra los datos, puede ser obvio que "verde oscuró" y "verde oscuro" indican la misma modalidad pero para el software, se trata de dos diferentes. Los femeninos y masculinos de los adjetivos, según califiquen al objeto o al color, también pueden plantear problemas. Por ejemplo, "hoja verde claro" y "hoja verde clara" serán considerados como dos categorías diferentes por el software. Una categoría falsa que suele encontrarse con frecuencia se refiere al espacio. La modalidad "verde" (sin espacio) y la modalidad "verde␣" (con espacio, representado aquí como "␣") será considerada por el software como dos categorías diferentes. Esta falsa categoría resulta especialmente desconcertante puesto que el espacio no se ve en la pantalla, de modo que el usuario tiene realmente la impresión de que se trata de la misma categoría. Todos los caracteres invisibles (ir al renglón siguiente, tabulación, etc.) o los caracteres que aparecen del mismo modo en la pantalla aunque tengan códigos ASCII diferentes, pueden generar el mismo tipo de error desconcertante.

Las categorías falsas pueden evitarse usando máscaras de entrada que sólo permiten entrar las variables cualitativas a partir de una lista de categorías admisibles. El uso de scripts automáticos para verificar los datos, que supriman los espacios incorrectos, verifiquen los acentos o si se trata de minúsculas o mayúsculas, y que verifiquen que las variables cualitativas tengan un valor comprendido dentro de una lista de categorías admisibles, es algo necesario para grandes conjuntos de datos.

4.3. Estructura de los datos

La estructura de los datos consiste en organizarlos en un formato que permita realizar los cálculos necesarios para elaborar el modelo. Normalmente se trata de una tabla que tiene un renglón por cada individuo estadístico (un árbol para un modelo individual, una parcela para un modelo de rodal) y tantas columnas como variables descriptivas haya (tanto para las variables que haya que predecir: biomasa, volumen, etc., como para las variables explicativas: diámetro, altura, etc.). Esta fase de la estructuración de la base de datos puede exigir manipulaciones bastante avanzadas de los datos. En ciertos casos habrá que agregar los datos de un nivel de descripción a otro. Por ejemplo, si se quiere construir un modelo individual para un rebrote y las mediciones se hicieron sobre fustes individuales, habrá que

agregar los datos relativos a los fustes de un mismo tocón: sumar los volúmenes y las masas, calcular el diámetro equivalente del tocón (es decir, el diámetro cuadrático medio) a partir de los diámetros de sus fustes. Otro ejemplo es la elaboración de un modelo de rodal a partir de las mediciones de los árboles individuales. En ese caso habrá que sumar los datos relativos a los árboles a los datos que caracterizan el rodal (volumenpor hectárea, altura dominante, etc.). En otros casos, por el contrario, habrá que dividir los conjuntos de datos. Por ejemplo, se calculado el volumen de árboles escogidos al azar en una masa pluriespecífica y se quiere elaborar un modelo separado para las cinco especies dominantes. En ese caso hay que dividir el conjunto de datos en función de las especies arbóreas.

La estructura de los datos será mucho más fácil si su ingreso se hizo en el formato adecuado. Las bases de datos relacionales tienen la ventaja de ofrecer un lenguaje de búsqueda que permite elaborar fácilmente ese tipo de cuadros sintéticos. En el programa Microsoft Excel, la herramienta de "tablas dinámicas" se podrá aprovechar bien para estructurar los datos.

Conjunto de datos del "linea roja"

Para ilustrar algunas particularidades en este manual utilizaremos un conjunto de datos recopilados en Ghana por Henry *et al.* (2010). Dicho conjunto de datos da la biomasa seca de 42 árboles que pertenecen a 16 especies de un bosque muy húmedo tropical. Para cada árbol, se midió el diámetro a la altura del pecho, la altura, el diámetro de su copa, la densidad promedio de su madera, el volumen y la biomasa seca en cinco compartimientos: ramas, hojas,tronco, aletones y biomasa total.

El Cuadro 4.2 presenta los datos de Henry *et al.* (2010) tal y como deberán ser presentados en una hoja de cálculo. El Cuadro de los datos se presenta en una hoja de cálculo rectangular donde figuran los datos; no debe haber ningún renglón ni columna en blanco, ni ninguna presentación que se aleje de esta matriz de datos. Hay que evitar todo lo que sea puramente decorativo: tabulaciones o casillas vacías para aligerar la presentación, puesto que el software estadístico no podrá leer el conjunto de datos que no corresponda al formato de la matriz. Los títulos de las columnas se limitarán a palabras cortas, incluso abreviaturas. La información sobre el significado de las variables y su unidad de registro se pondrá en la metainformación.

Si hubiera que ingresar información sobre las especies, éstas se registrarían en un segundo cuadro dado que hay dos niveles anidados: el nivel de especie, con varios árboles por especie; y el nivel del árbol, anidado en el nivel de especie. De este modo, si se quisiera registrar el conjunto de especies que comparten los mismos recursos (*ecological guild*) y el nombre vernáculo de las especies, se obtendría un segundo Cuadro 4.3 propio a la especie, siendo el nombre científico de la especie el identificador que permite establecer la relación entre el Cuadro 4.2 y el Cuadro 4.3.

Lectura de los datos. Supongamos que los datos, presentados en forma de matriz, están guardados en un archivo Excel `Henry_et_al2010.xls`, cuya primera hoja se titula `biomasa` y contiene el Cuadro 4.2. En el soporte lógico R, la lectura de los datos se realiza mediante las instrucciones o comandos siguientes:

```
library(RODBC)
ch <- odbcConnectExcel("Henry_et_al2010.xls")
dat <- sqlFetch(ch,"biomasa")
odbcClose(ch)
```

Los datos se almacenan luego en el objeto `dat`.

Verificación de los datos. Algunas rutinas pueden hacerse para comprobar la calidad de los datos. En R, el comando `summary` da las estadísticas descriptivas básicas de las variables de un cuadro de datos:

```
summary(dat)
```

En particular para el diámetro, el resultado es:

```
      dbh
 Min.   :  2.60
 1st Qu.: 15.03
 Median : 59.25
 Mean   : 58.59
 3rd Qu.: 89.75
 Max.   :180.00
```

Así pues, el diámetro de los árboles medidos va de 2,6 cm a 180 cm, con un promedio de 58,59 cm y un diámetro mediano de 59,25 cm. Las estadísticas descriptivas básicas para la biomasa seca total son:

```
      Btot
 Min.   : 0.0000
 1st Qu.: 0.1375
 Median : 3.1500
 Mean   : 6.8155
 3rd Qu.: 9.6075
 Max.   :70.2400
```

La biomasa seca total del árbol más grande es de 70,24 toneladas. La biomasa seca del árbol más pequeño en el conjunto de datos es cero. Dado que las biomasas se expresan en toneladas con dos cifras significativas, ese valor no es un dato aberrante sino que simplemente significa que la biomasa seca de ese árbol es inferior a 0,01 toneladas = 10 kg. Sin embargo, ese valor cero planteará problemas más tarde cuando se quiera realizar una transformación logarítmica.

Por último, podemos asegurarnos de que la biomasa seca total sea realmente la suma de las biomasas de los otros cuatro compartimientos:

```
max(abs(dat$Btot-rowSums(dat[,c("Bbran","Bfol","Btronc","Bctf")])))
```

La mayor diferencia en valor absoluto es igual a 0,01 toneladas, lo que corresponde bien a la precisión de los datos (dos cifras significativas). Por ende, no hay ninguna incongruencia a este nivel en el conjunto de datos.

Cuadro 4.2 – *Datos de biomasa de los árboles de Henry et al. (2010) en Ghana.* `dbh` *es el diámetro en cm,* `haut` *es la altura en m,* `houp` *es el diámetro de la copa en m,* `dens` *es la densidad promedia de la madera en* $g\,cm^{-3}$, `volume` *es el volumen en* m^3, `Bbran` *es la biomasa seca de las ramas en toneladas,* `Bfol` *es la biomasa foliar seca en toneladas,* `Btronc` *es la biomasa seca del tronco en toneladas,* `Bctf` *es la biomasa seca de los aletones en toneladas, y* `Btot` *es la biomasa seca total en toneladas.*

especie	dbh	haut	houp	dens	volume	Bbran	Bfol	Btronc	Bctf	Btot
Heritiera utilis	7,3	5,1	3,7	0,58	0,03	0,02	0	0	0	0,02
Heritiera utilis	12,4	12	5	0,62	0,11	0,02	0	0,05	0	0,07
Heritiera utilis	31	22	9	0,61	1,34	0,1	0,01	0,71	0,02	0,83
Heritiera utilis	32,5	27,5	7,1	0,61	1,12	0,07	0,01	0,61	0,01	0,7
Heritiera utilis	48,1	35,6	7,9	0,61	3,83	0,24	0,01	2,07	0,01	2,33
Heritiera utilis	56,5	35,1	8	0,6	5,43	0,85	0,03	2,28	0,14	3,31
Heritiera utilis	62	40,4	11,1	0,6	6,84	0,68	0,04	3,28	0,15	4,15
Heritiera utilis	71,9	42,3	20	0,6	9,84	1,34	0,05	4,43	0,11	5,93
Heritiera utilis	83	39,4	15,9	0,6	11,89	2,2	0,09	4,83	0,04	7,16
Heritiera utilis	100	50,5	19,1	0,58	31,71	8,71	0,11	8,39	1,4	18,61
Heritiera utilis	105	50,5	19,2	0,58	35,36	8,81	0,13	11,18	0,65	20,76
Heritiera utilis	6,5	8,1	1,5	0,78	0,01	0,01	0	0	0	0,01
Tieghemella heckelii	12	17	4,7	0,78	0,15	0,12	0,01	0	0	0,13
Tieghemella heckelii	73,5	45	11,1	0,66	11,08	1,27	0,04	5,91	0,14	7,36
Tieghemella heckelii	80,5	50,7	13	0,66	12,25	1,54	0,05	6,45	0,09	8,13
Tieghemella heckelii	93	45	17	0,66	17,79	3,66	0,06	7,8	0,21	11,73
Tieghemella heckelii	180	61	41	0,62	112,81	27,28	0,74	35,07	7,16	70,24
Piptadeniastrum africanum	70	39,7	10,5	0,58	10,98	2,97	0,06	3,29	0,07	6,39
Piptadeniastrum africanum	89	50	18,8	0,57	15,72	3,69	0,05	5,16	0,16	9,06
Piptadeniastrum africanum	90	50,2	16	0,57	22,34	5,73	0,38	6,23	0,74	13,08
Aubrevillea kerstingii	65	32,5	9	0,62	4,79	1,52	0,02	1,45	0	2,99
Afzelia bella	83,6	40	13,5	0,67	14,57	3,17	0,03	6	0,58	9,79
Cecropia peltata	7,8	2,3	2,5	0,17	0,07	0	0	0,01	0	0,01
Cecropia peltata	20,5	21,2	6,2	0,23	0,44	0,03	0	0,07	0	0,11
Cecropia peltata	29,3	22,5	8,9	0,27	1,11	0,13	0,01	0,16	0	0,31
Cecropia peltata	35,5	12	7,3	0,26	1,39	0,12	0,02	0,25	0	0,38
Ceiba pentandra	132	45	16	0,54	28,55	1,53	0,04	13,37	0,44	15,39
Ceiba pentandra	170	51	27,1	0,26	64,84	3,2	0,1	11,87	1,88	17,05
Nauclea diderrichii	2,6	4,9	8,4	0,76	0	0	0	0	0	0
Nauclea diderrichii	94,6	50,5	12	0,5	17,19	1,06	0,02	7,49	0,06	8,64
Nauclea diderrichii	110	58,8	14,1	0,4	28,71	3,47	0,06	7,9	0,07	11,49
Nauclea diderrichii	112	40	13,2	0,47	22,74	3,41	0,1	7,19	0,13	10,82
Daniellia thurifera	9	9,3	8	0,42	0,11	0,05	0,01	0	0	0,05
Guarea cedrata	12,8	13	3,1	0,62	0,12	0,08	0,01	0	0	0,08
Guarea cedrata	71,5	45,5	14	0,5	10,12	0,65	0,02	4,3	0,13	5,1
Strombosia glaucescens	7,6	11,3	3,9	0,66	0,07	0,05	0,01	0	0	0,05
Strombosia glaucescens	26,5	26	12,2	0,73	1,09	0,2	0,01	0,58	0	0,8
Garcinia epunctata	7,1	5,7	3,8	0,65	0,08	0,05	0,01	0	0	0,06
Drypetes chevalieri	13,2	15,7	5	0,65	0,22	0,15	0,02	0	0	0,16
Cola nitida	23,6	23,4	6,3	0,56	0,68	0,09	0,01	0,28	0	0,39
Nesogordonia papaverifera	24,3	30,2	6,5	0,69	0,73	0,12	0,01	0,36	0,02	0,51
Dialium aubrevilliei	98	43,7	98	0,65	18,49	2,55	0,05	9,07	0,4	12,07

Cuadro 4.3 – *Datos sobre las especies objeto del muestreo por Henry et al. (2010) en Ghana.*

guild	especie	vernacular
heliófila no pionera	Heritiera utilis	Nyankom
heliófila no pionera	Tieghemella heckelii	Makore
heliófila no pionera	Piptadeniastrum africanum	Dahoma
heliófila no pionera	Aubrevillea kerstingii	Dahomanua
heliófila no pionera	Afzelia bella	Papao-nua
pionera	Cecropia peltata	Odwuma
pionera	Ceiba pentandra	Onyina
pionera	Nauclea diderrichii	Kusia
pionera	Daniellia thurifera	Sopi
tolerante a la sombra	Guarea cedrata	Kwabohoro
tolerante a la sombra	Strombosia glaucescens	Afena
tolerante a la sombra	Garcinia epunctata	Nsokonua
tolerante a la sombra	Drypetes chevalieri	Katreka
tolerante a la sombra	Cola nitida	Bese
tolerante a la sombra	Nesogordonia papaverifera	Danta
tolerante a la sombra	Dialium aubrevilliei	Dua bankye

5

Exploración gráfica de los datos

La exploración gráfica de los datos es la primera etapa de su análisis. Consiste en estudiar visualmente las relaciones entre las variables para hacerse una idea del tipo de modelo que hay que ajustar. Concretamente, se proyecta en un gráfico una nube de puntos cuyas coordenadas corresponden a dos variables: la variable explicativa en el eje de las x y la variable dependiente en el eje de las y. Un gráfico sólo puede elaborarse para un máximo de dos variables al mismo tiempo (desde el punto de vista práctico los gráficos tridimensionales no pueden ser analizados visualmente). Para explorar gráficamente las relaciones entre p variables (con $p > 1$), se harán pues $p(p-1)/2$ gráficos de dos variables y/o se tratará de construir variables explicativas sintéticas a partir de varias variables explicativas (volveremos a abordar este punto en el § 5.1.1).

Supongamos pues que tenemos una variable de respuesta denominada Y (el volumen, la biomasa, etc.) y p variables explicativas denominadas $X_1, X_2, ..., X_p$ (el diámetro, la altura, etc.). El objetivo de la exploración gráfica no es seleccionar entre las n variables explicativas aquellas que se utilizarán realmente para el modelo: la selección de las variables supone que sabemos probar el carácter significativo de una variable, lo que ocurre en la fase siguiente de ajuste del modelo. Las p variables explicativas se consideran entonces como fijas y se busca la forma del modelo que vincula mejor la variable Y a las variables X_1 a X_p. Un modelo se compone de dos términos: la media y el error (o residuo). La exploración gráfica pretende precisar al mismo tiempo la forma de la relación promedio y aquella del error pero sin preocuparse del valor de los parámetros del modelo (ésta será la etapa siguiente del ajuste del modelo). La relación media puede ser lineal o no lineal, linealizable o no; el error residual puede ser aditivo o multiplicativo, de varianza constante (homocedasticidad) o no (heterocedasticidad). Como ejemplo en la Figura 5.1 muestra estos cuatro casos posibles, dependiendo de que la relación sea lineal o no y la varianza de los residuos, constante o no.

La fase de exploración gráfica de los datos también es necesaria para evitar caer en la trampa del ajuste "a ciegas": en efecto, se puede tener la impresión de que el ajuste de un modelo a los datos es de buena calidad cuando en realidad se trata de un "artefacto". Esto se ilustra en la Figura 5.2 en el caso de la relación lineal. En los cuatro casos que aparecen en esta Figura, el R^2 de la regresión lineal de Y con respecto a X es elevado mientras que, en realidad, la relación lineal $Y = a + bX + \varepsilon$ no se adapta a los datos. En la Figura 5.2A, la nube de puntos se estructura en tres subconjuntos en los cuáles la relación entre Y y X

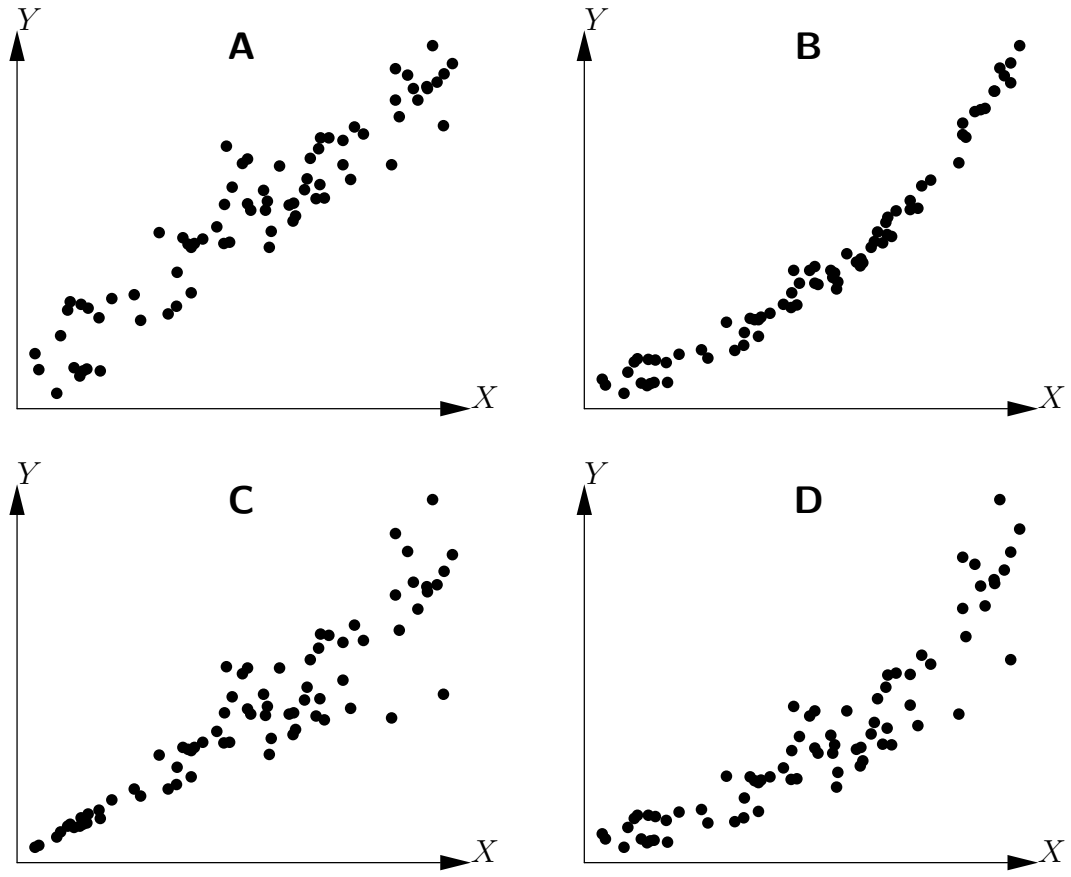


Figura 5.1 – *Ejemplo de las relaciones entre las dos variables X e Y: A) relación lineal y varianza de los residuos constante, B) relación no lineal y varianza de los residuos constante, C) relación lineal y varianza de los residuos no constante, D) relación no lineal y varianza de los residuos no constante.*

es lineal con un coeficiente de correlación negativo. Sin embargo, estos tres subconjuntos se organizan a lo largo de una recta de pendiente positiva, que es la recta originada por la regresión lineal. En 5.2B, la nube de puntos, salvo un único dato aislado (probablemente un dato aberrante), no presenta ninguna relación entre Y y X. Pero el dato aislado basta para hacer creer que existe una relación positiva entre Y y X. En 5.2C, la relación entre Y y X es parabólica. Por último en 5.2D, la nube de puntos, salvo un único dato excéntrico, se estructura a lo largo de una recta de pendiente positiva. En este caso una relación lineal entre Y y X sería adaptada para describir los datos una vez excluido el dato aislado. Este dato aislado hace que se reduzca artificialmente el valor de R^2 (por oposición al gráfico 5.2B donde el dato aislado aumenta artificialmente R^2).

Como su nombre lo indica, la fase de exploración gráfica es más exploratoria que un método sistemático. Aun cuando puedan darse una serie de consejos para encontrar el buen modelo, hacen falta experiencia e intuición para lograrlo.

5.1. Exploración de la relación promedio

En esta Sección nos interesamos en la forma gráfica de determinar el carácter de la relación promedio entre dos variables X e Y, es decir, en encontrar la forma de la función

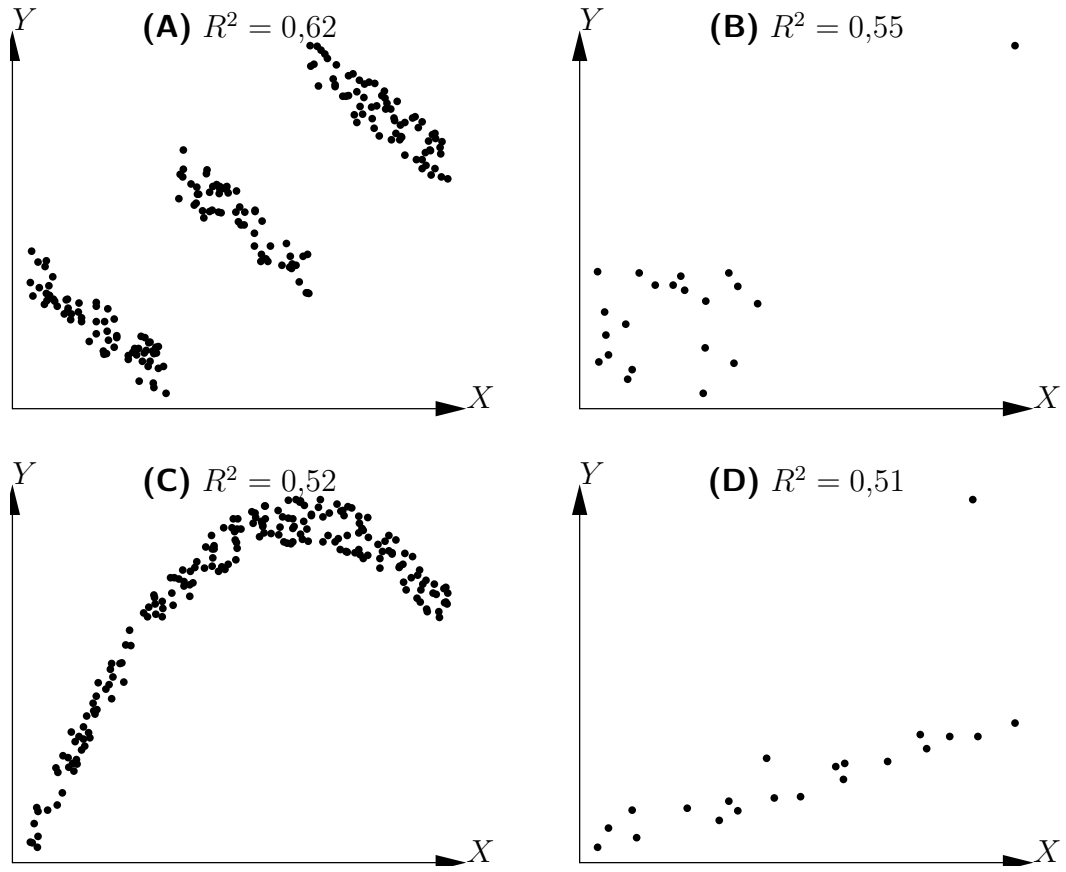


Figura 5.2 – *Coeficientes de determinación (R^2) de las regresiones lineales realizadas en las nubes de puntos que no presentan relaciones lineales.*

f (¡si existe!) tal que $\mathrm{E}(Y) = f(X)$. Cuando no hay más que una variable explicativa X, la exploración gráfica consiste en trazar la nube de puntos de Y en función de X.

②

Explorando la relación biomasa–diámetro

Para ver la forma de la relación entre la biomasa seca total y el diámetro a la altura de pecho de los árboles, se dibuja la nube de puntos de la biomasa en función del diámetro. Una vez leído el conjunto de datos (cf. Línea roja 1), el comando para dibuja la nube de puntos es:

```
plot(dat$dbh,dat$Btot,xlab="Diámetro (cm)",ylab="Biomasa (t)")
```

La nube de puntos resultante se muestra en la Figura 5.3. Esta nube de puntos es del mismo tipo que el gráfico de la Figura 5.1D: la relación entre la biomasa y el diámetro no es lineal y la varianza de la biomasa aumenta cuando aumenta el diámetro.

Como el método gráfico de la nube de puntos no puede usarse más que para una sola variable explicativa, se tratará de reducirlo a este caso cuando haya varias variables explicativas. Ante todo, expliquemos este último punto.

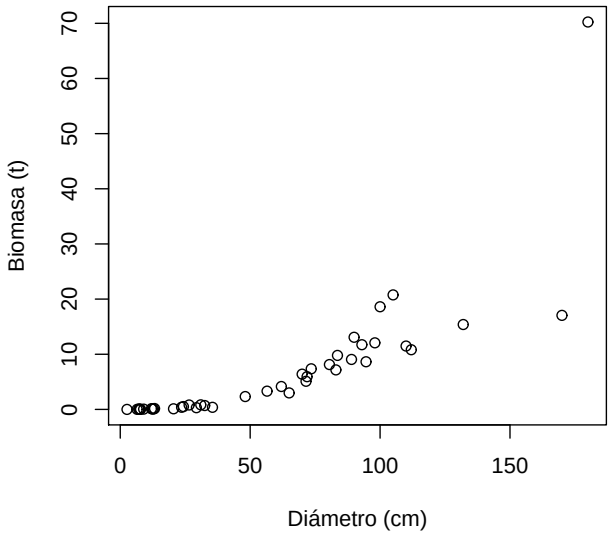


Figura 5.3 – *Nube de puntos de la biomasa seca total (toneladas) en función del diámetro a la altura del pecho (cm) para los 42 árboles medidos en Ghana por Henry et al. (2010).*

5.1.1. Cuando hay más de una variable explicativa

Lo primero es ver si es posible formar, a partir de varias variables explicativas, una sola variable explicativa sintética. Por ejemplo, si queremos predecir el volumen del tronco a partir de su diámetro a la altura del pecho D y de su altura H, podemos estar seguros que la nueva variable D^2H será un predictor efectivo. En ese caso, a partir de dos variables explicativas D y H se formó una nueva (¡y única!) variable explicativa D^2H. Por ejemplo Louppe *et al.* (1994) crearon el modelo de volumen individual siguiente para *Afzelia africana* en la Reserva Forestal Badénou en Côte d'Ivoire:

$$V = -0{,}0019 + 0{,}04846C^2H$$

donde V es el volumen total en m^3, C la circunferencia a 1,30 m en m y H la altura en m. Aunque se trate de una tabla con dos entradas (la circunferencia y la altura), no hay más que una variable explicativa: C^2H. Otro ejemplo es la tabla de cubicación del rodal establecida por Fonweban & Houllier (1997) en Camerún para *Eucalyptus saligna*:

$$V = \beta_1 G^{\beta_2} \left(\frac{H_0}{N}\right)^{\beta_3}$$

donde V es el volumen del rodal en $m^3\,ha^{-1}$, G es el área basal en $m^2\,ha^{-1}$, H_0 es la altura dominante del rodal, N es la densidad del mismo (número de árboles por hectárea) y los β son parámetros constantes. Aun cuando se trate de un modelo de tres entradas (el área basal, la altura dominante y la densidad), en realidad sólo hay dos variables explicativas: G y la relación H_0/N.

③

Explorando la relación biomasa–D^2H

Comparado con un modelo de biomasa de dos entradas usando el diámetro a la altura del pecho D y la altura H, la cantidad D^2H constituye una aproximación del volumen del tronco

(dejando de lado el coeficiente de forma) y puede usarse por tanto como variable explicativa sintética. La nube de puntos de la biomasa en función de D^2H se obtiene mediante el comando:

```
with(dat,plot(dbh^2*haut,Btot,xlab="D2H (cm2.m)",ylab="Biomasa (t)"))
```

El resultado se representa en la Figura 5.4. Esta nube de puntos es del mismo tipo que el gráfico de la Figura 5.1C: la relación entre la biomasa y D^2H es lineal pero la varianza de la biomasa aumenta a medida que aumenta D^2H.

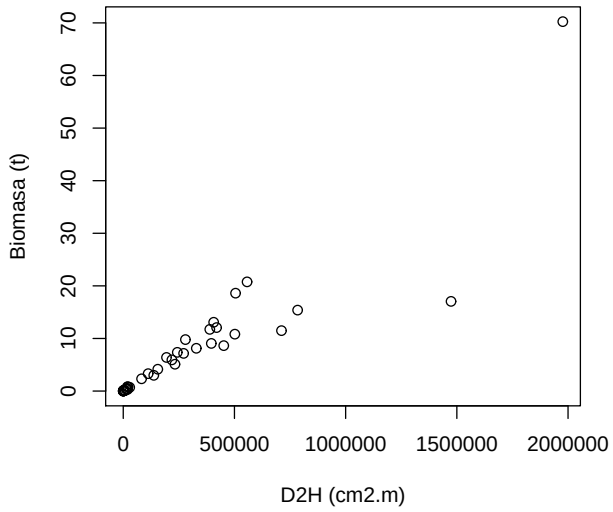


Figura 5.4 – *Nube de puntos de la biomasa seca total (toneladas) en función de D^2H, donde D es el diámetro a la altura del pecho (cm) y H la altura (m) para los 42 árboles medidos en Ghana por Henry et al. (2010).*

Supongamos que, después de esta fase de agregación de las variables explicativas, todavía quedan p variables explicativas $X_1, \ldots, X_p$ (con $p > 1$). Primero se podrían explorar las p relaciones entre Y y cada una de las p variables explicativas. Se trata en efecto de relaciones entre dos variables y los métodos gráficos que presentaremos luego, en consecuencia, se aplican bien a esos casos. Sin embargo este enfoque suele ser muy poco informativo ya que la relación entre Y y p variables no se reduce a las p relaciones entre Y y cada una de las p variables separadamente. Un ejemplo sencillo puede ilustrar este concepto: supongamos que la variable Y sea (dentro del margen de error) la suma de dos variables explicativas:

$$Y = X_1 + X_2 + \varepsilon \tag{5.1}$$

donde ε es un error de esperanza cero y que las variables X_1 y X_2 están vinculadas de forma tal que X_1 varía entre 0 y $-X_{\text{máx}}$ y que, con una X_1 dada, X_2 varía entre máx$(0, -X1)$ y mín$(X_{\text{máx}}, 1 - X_1)$. La Figura 5.5 muestra los dos gráficos de Y en función de cada una de las variable explicativas X_1 y X_2 para datos simulados según el modelo (con $X_{\text{máx}} = 5$). La nube de puntos parece no tener una estructura particular y por ende no se puede detectar el modelo $\mathrm{E}(Y) = X_1 + X_2$.

Una forma de resolver este problema es a través del condicionamiento. Esto se trata de examinar la relación entre la variable de respuesta Y una de las variables explicativas (supongamos X_2) condicionalmente con respecto a los valores de la otra variable explicativa

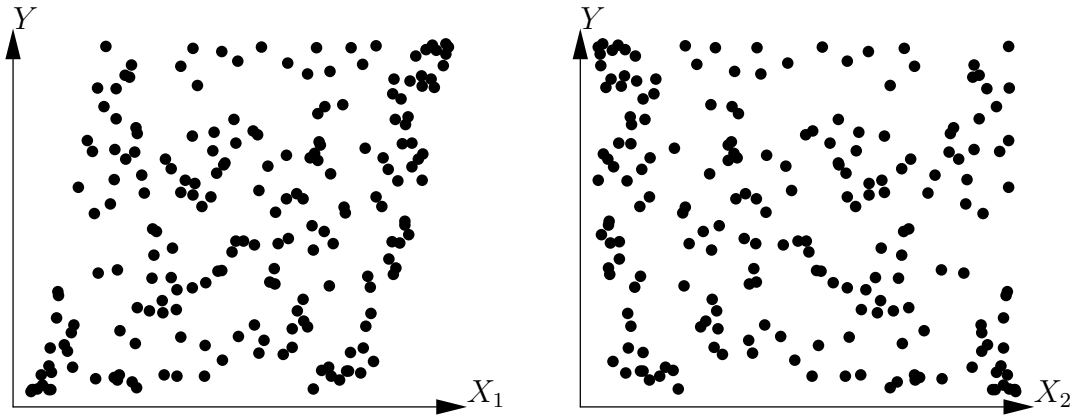


Figura 5.5 – *Gráficos de una variable Y en función de cada una de las dos variables explicativas X_1 y X_2 tales que $E(Y) = X_1 + X_2$.*

(en este caso X_1). En la práctica, se divide el conjunto de datos según las clases de valores de X_1, luego se explora la relación entre Y y X_2 en cada uno de los subconjuntos de datos. Siguiendo el ejemplo anterior, se dividieron los valores de X_1 en 12 intervalos grandes de 0,5 unidades: el primero va de -5 a $-4{,}5$, el segundo de $-4{,}5$ a -4, etc., hasta el último intervalo que va de 0,5 a 1. El conjunto de datos representado en la Figura 5.5 se dividió en 12 subconjuntos de datos en función de 12 clases de valores de X_1, luego se trazaron los 12 gráficos de Y en función de X_2 para estos 12 subconjuntos de datos. El resultado se representa en la Figura 5.6. La superposición de los gráficos de dicha Figura daría nuevamente el gráfico de la derecha de la Figura 5.5. Estos gráficos muestran que, para un valor dado de X_1, la relación entre Y y X_2 es realmente lineal. Además se puede ver que la pendiente de la línea de Y en función de X_2 para una X_1 dada, es constante para todos los valores de X_1. Esta exploración gráfica demuestra por tanto que el modelo es de tipo:

$$\mathrm{E}(Y) = f(X_1) + aX_2$$

donde a es un coeficiente constante (en este caso igual a 1, pero la exploración gráfica no se ocupa del valor de los parámetros), y $f(X_1)$ representa la intersección de la recta que une Y a X_2 para un valor dado de X_1. Esta intersección potencialmente varía en función de X_1, según una función f que queda por determinar.

Para explorar la forma de la función f, podemos ajustar por regresión lineal una recta a cada uno de los 12 subconjuntos de datos de Y y X_2 que corresponden a las 12 clases de valores de X_1. Se determina la intersección y_0, de cada una de esas 12 rectas y se grafica y_0 en función del punto medio de cada clase de valores X_1. La Figura 5.7 muestra este gráfico para los mismos datos simulados anteriormente. Esta exploración gráfica revela que la relación entre y_0 y X_1 es lineal, es decir: $f(X_1) = bX_1 + c$. Al final la exploración gráfica basada en el condicionamiento con respecto a X_1 reveló que un modelo adecuado era:

$$\mathrm{E}(Y) = aX_2 + bX_1 + c$$

Como las variables X_1 y X_2 desempeñan un papel simétrico en el modelo (5.1), el condicionamiento también resulta simétrico con respecto a ambas variables. Aquí hemos estudiado la relación entre Y y X_2 condicionalmente con respecto a X_1, pero hubiéramos llegado del mismo modo al mismo modelo explorando la relación entre Y y X_1 condicionalmente con respecto a X_2.

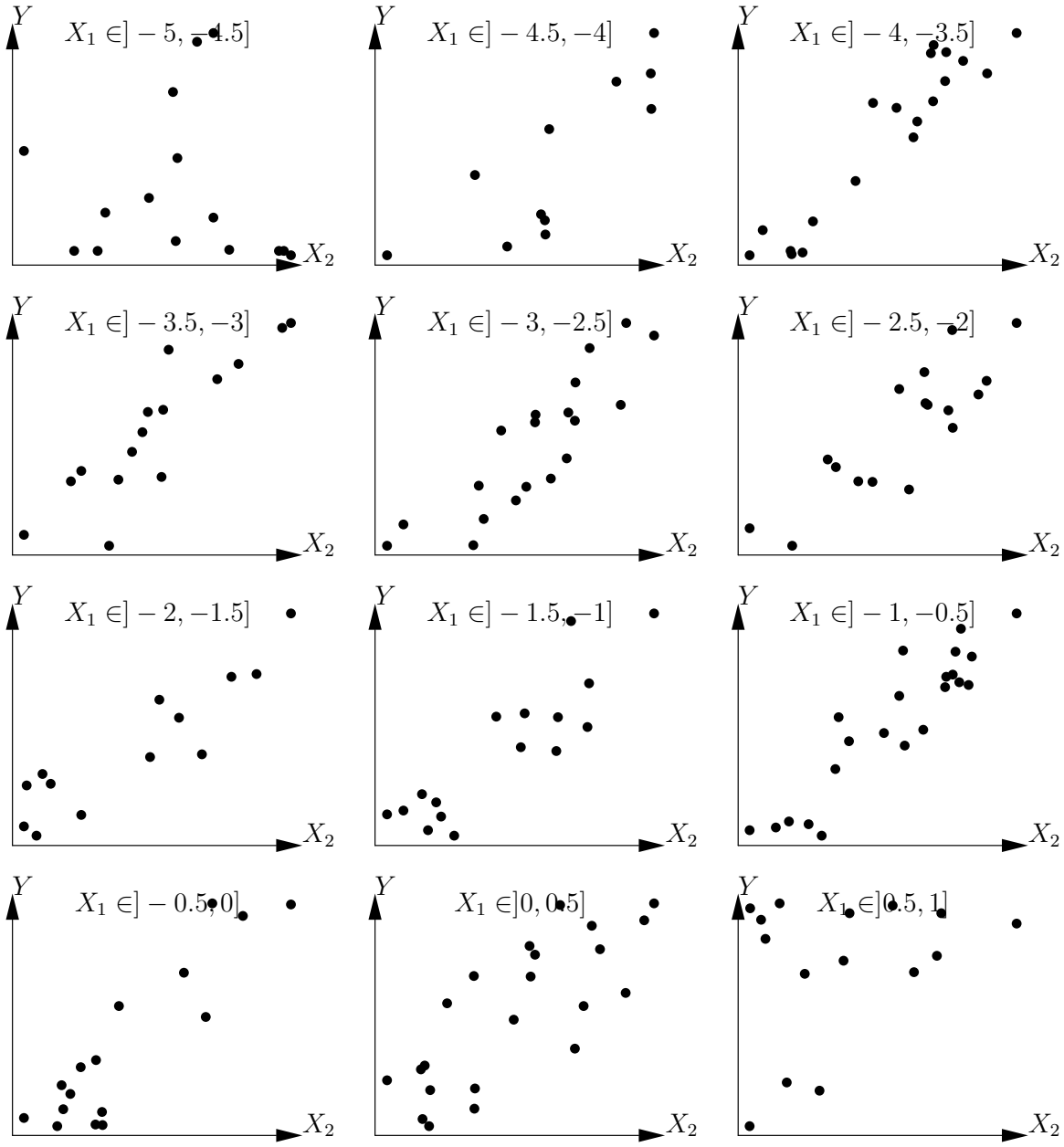


Figura 5.6 – *Gráficos de una variable Y en función de una variable explicativa X_2 para cada uno de los subconjuntos de datos definidos por las clases de valores de otra variable explicativa X_1, con $E(Y) = X_1 + X_2$.*

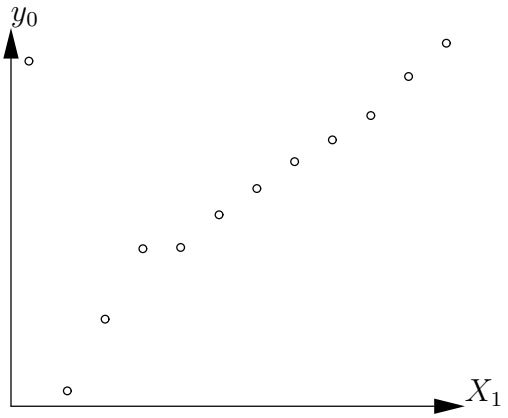


Figura 5.7 – *Trazado de la intersección de la regresión lineal de Y con respecto a X_2 para un subconjunto de datos correspondiente a una clase de valores de X_1 en función del medio de estas clases, para datos simulados según el modelo $Y = X_1 + X_2 + \varepsilon$.*

En este ejemplo, la relación entre Y y X_2 para un X_1 dado, es una recta cuya pendiente es independiente de X_2: se dice que no hay interacción entre X_1 y X_2. Un modelo con interacción sería, por ejemplo, $\mathrm{E}(Y) = X_1 + X_2 + X_1X_2$. En este caso, la relación entre Y y X_2 a X_1 dado es una recta cuya pendiente, igual a $1 + X_1$, depende en efecto de X_1. El condicionamiento permite, sin mayor dificultad, explorar la forma de los modelos con interacciones entre las variables explicativas.

El condicionamiento se extiende, en principio, a cualquier número de variables explicativas. Para tres variables explicativas X_1, X_2, X_3 por ejemplo, se podrá explorar la relación entre Y y X_3 para X_1 y X_2 fijas; tengamos en cuenta f la función que define esta relación, así como $\theta(X_1, X_2)$ los parámetros de f (que dependen potencialmente de X_1 y X_2):

$$\mathrm{E}(Y) = f[X_3; \theta(X_1, X_2)]$$

A continuación, se explora la relación entre θ y las dos variables X_1 y X_2. Nuevamente se condiciona explorando la relación entre θ y X_2 para X_1 fija; observemos g, la función que define esta relación, y $\phi(X_1)$ los parámetros de g (que dependen potencialmente de X_1):

$$\theta(X_1, X_2) = g[X_2; \phi(X_1)]$$

Por último, exploramos la relación entre ϕ y X_1; siendo h la función que define esta relación. Al final de cuentas, el modelo que describe los datos será:

$$\mathrm{E}(Y) = f\{X_3; g[X_2; h(X_1)]\}$$

Este razonamiento se aplica, en principio, a cualquier número de variables explicativas pero en la práctica se ve bien que resulta difícil aplicarlo a $p > 3$. El condicionamiento exige, además, abundantes datos ya que cada subconjunto de datos, definido por clases de valores de variables condicionales, debe incluir una cantidad suficiente de datos para poder explorar gráficamente las relaciones entre las variables. En el caso de las tres variables explicativas, los subconjuntos de datos se definen mediante el cruce de las clases de valores de X_1 y X_2 (por ejemplo). Si el conjunto de datos completo incluye n observaciones, si X_1 y X_2 se dividen en 10 clases de valores y si los datos se distribuyen equitativamente según sus clases, entonces cada subconjunto de datos sólo incluye $n/100$ observaciones. En la práctica,

a menos que el conjunto de datos sea particularmente grande, es difícil utilizar el principio de condicionamiento para más de dos variables explicativas.

Para ajustar los modelos de biomasa o de volumen, el número de entradas del modelo suele ser limitado (dos o tres entradas como máximo), de modo que, generalmente, no tenemos que enfrentarnos al problema de la exploración gráfica con un elevado número de variables explicativas. De ser ese el caso, se podrían utilizar análisis multivariados, como el análisis de los componentes principales (Philippeau, 1986; Härdle & Simar, 2003). Estos análisis consisten en proyectar las observaciones en un subespacio de dimensión reducida (con mucha frecuencia dos o tres), construido a partir de combinaciones lineales de variables explicativas y de forma tal que se maximice la variabilidad de las observaciones en ese subespacio. En otras palabras, estos análisis multivariados permiten visualizar las relaciones entre variables, perdiendo el mínimo posible de información, lo que constituye el objetivo buscado por la exploración gráfica.

④

Condicionamiento relativo a la densidad de la madera

Exploremos ahora la relación entre la biomasa, D^2H y la densidad de la madera ρ. Se definen `n` clases de densidad de madera, de forma tal que cada una contenga aproximadamente el mismo número de observaciones:

```
d <- quantile(dat$dens,(0:n)/n)
i <- findInterval(dat$dens,d,rightmost.closed=TRUE)
```

El objeto `d` define los límites de las clases de densidad mientras que el objeto `i` contiene el número de la clase de densidad a la que corresponde cada observación. El gráfico de la biomasa en función de D^2H en coordenadas logarítmicas, con los distintos símbolos y colores según la clase de densidad, se obtiene con el comando:

```
with(dat,plot(dbh^2*haut,Btot,xlab="D2H (cm2m)",ylab="Biomasa (t)",log="xy",pch=i, col=i))
```

y está representado en la Figura 5.8 para `n` $= 4$ clases de densidad de la madera. Anticipándonos al contenido del Capítulo 6, ajustamos una regresión lineal entre $\ln(B)$ e $\ln(D^2H)$ para cada subconjunto de observaciones correspondiente a cada clase de densidad de la madera:

```
m <- as.data.frame(lapply(split(dat,i), function(x)
coef(lm(log(Btot)~I(log(dbh^2*haut)),data=x[x$Btot>0,]))))
```

Para graficar la intersección de la regresión y su pendiente en función de la densidad mediana de la clase:

```
dmid <- (d[-1]+d[-n])/2
plot(dmid,m[1,],xlab="Densidad de la madera (g/cm3)",ylab="Intersección")
plot(dmid,m[2,],xlab="Densidad de la madera (g/cm3)",ylab="Pendiente")
```

a primera vista, no observamos ninguna relación en particular (Figura 5.9).

5.1.2. ¿Cómo detectar si una relación es adecuada?

En adelante, supongamos que tenemos una sola variable explicativa X y que buscamos explorar la relación entre X y la variable Y que hay que explicar. La primera etapa es

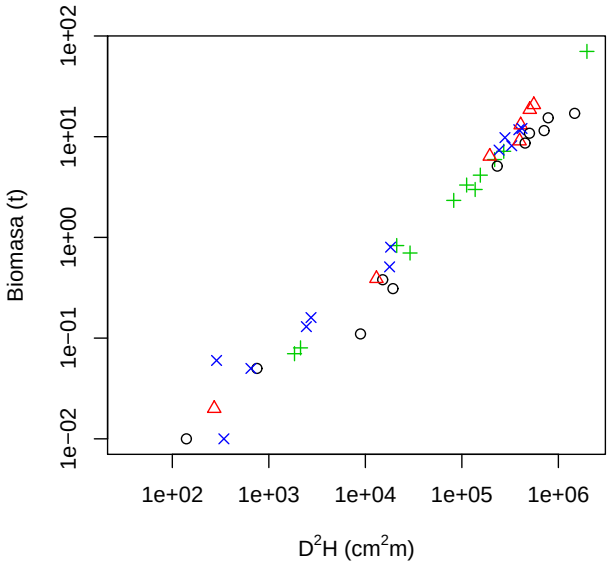


Figura 5.8 – *Nube de puntos (datos transformados logarítmicamente) de la biomasa seca total (toneladas) en función de D^2H, donde D es el diámetro a altura del pecho (cm) y H la altura (m) para los 42 árboles medidos en Ghana por Henry et al. (2010) con distintos símbolos según las clases de densidad de la madera: círculo negro, $0{,}170 \leq \rho < 0{,}545\ g\,cm^{-3}$; triángulo rojo, $0{,}545 \leq \rho < 0{,}600\ g\,cm^{-3}$; signo más verde, $0{,}600 \leq \rho < 0{,}650\ g\,cm^{-3}$; cruz azul, $0{,}650 \leq \rho < 0{,}780\ g\,cm^{-3}$.*

graficar la nube de puntos que corresponde a los datos con X como abscisa e Y como ordenada. A continuación se trata de adivinar visualmente la función que pasa por el medio de dicha nube, siguiendo su forma. Se pone de manifiesto que el ojo humano es poco hábil para diferenciar entre formas similares. Por ejemplo, la Figura 5.10 presenta tres nubes de puntos que corresponden, en desorden, a los tres modelos siguientes (aquí asumimos que el término de error es cero):

modelo de potencia:	$Y = aX^b$
modelo exponencial:	$Y = a\exp(bX)$
modelo polinomial:	$Y = a + bX + cX^2 + dX^3$

Las tres nubes de puntos tienen una apariencia similar y habría que ser muy hábil para poder decir a qué modelo corresponde cada una de ellas.

Por el contrario, el ojo humano es hábil para distinguir si una relación es lineal o no. Para detectar visualmente si la forma de una nube de puntos se ajusta o no a una función conviene mucho, cuando es posible, utilizar una transformación de variables que vuelva la relación lineal. En el caso del modelo de potencia, por ejemplo, $Y = aX^b$ implica que $\ln Y = \ln a + b \ln X$. La transformación de las variables:

$$\begin{cases} X' = \ln X \\ Y' = \ln Y \end{cases} \tag{5.2}$$

vuelve la relación lineal. En el caso del modelo exponencial, $Y = a\exp(bX)$ implica $\ln Y = \ln a + bX$, donde la transformación de las variables:

$$\begin{cases} X' = X \\ Y' = \ln Y \end{cases} \tag{5.3}$$

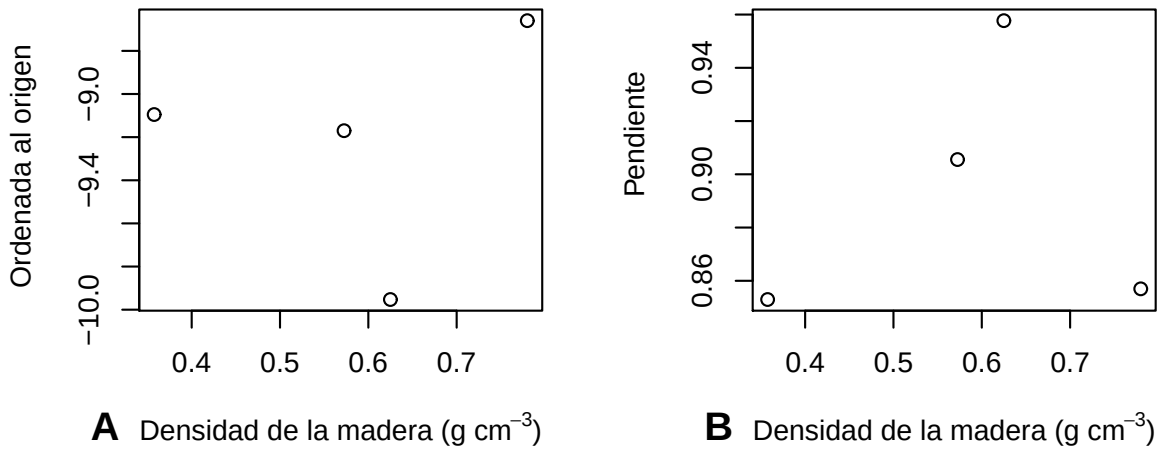


Figura 5.9 – *Intersección a y pendiente b de la regresión lineal* $\ln(B) = a + b\ln(D^2H)$ *condicional a la clase de densidad de la madera, en función de la densidad mediana de la madera mediana de las clases. Las regresiones se ajustan a los datos de los 42 árboles medidos por Henry et al. (2010) en Ghana.*

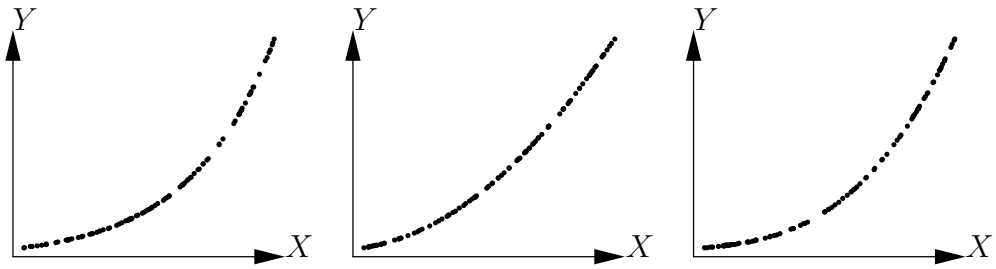


Figura 5.10 – *Tres nubes de puntos que corresponden, en desorden, a tres modelos: modelo de potencia, modelo exponencial y modelo polinomial.*

vuelve pues la relación lineal. Por el contrario, ninguna de las dos transformaciones permite linealizar el modelo polinomial. Si aplicamos estas transformaciones de variables a los datos representados en la Figura 5.10, vamos a estar en condiciones de descubrir cuál de las nubes corresponde a cada uno de estos modelos. La Figura 5.11 representa las tres nubes de puntos después de aplicar la transformación de variables como en (5.3). La primera nube de puntos toma la forma de una recta mientras que las otras dos mantienen la forma curva. La nube de puntos más a la izquierda de la Figura 5.10, corresponde así al modelo exponencial.

La Figura 5.12 representa las tres nubes de puntos después de efectuar la transformación de las variables (5.2). La segunda nube toma la forma de una recta mientras que las otras dos mantienen la forma curva. La nube de puntos en el centro de la Figura 5.10 corresponde también al modelo de potencia. Por deducción, la nube de puntos más a la derecha en dicha Figura 5.10 corresponde al modelo polinomial.

No siempre es posible encontrar una transformación de variables que vuelva lineal la relación. Este es, precisamente, el caso del modelo polinomial $Y = a + bX + cX^2 + dX^3$: no se puede encontrar una transformación de X en X' ni de Y en Y' que permita que la relación entre Y' y X' sea una recta, independientemente de los coeficientes a, b, c y d. Debe quedar claro también que la linealidad de la que hablamos aquí es la de la relación entre la variable dependiente Y, y la variable explicativa X. No se trata de la linealidad en el

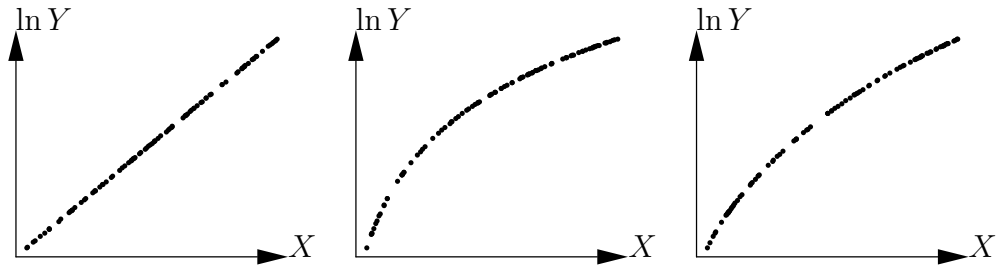


Figura 5.11 – *Aplicación de la transformación de variables* $X \to X$, $Y \to \ln Y$ *a las nubes de puntos representadas en la Figura 5.10.*

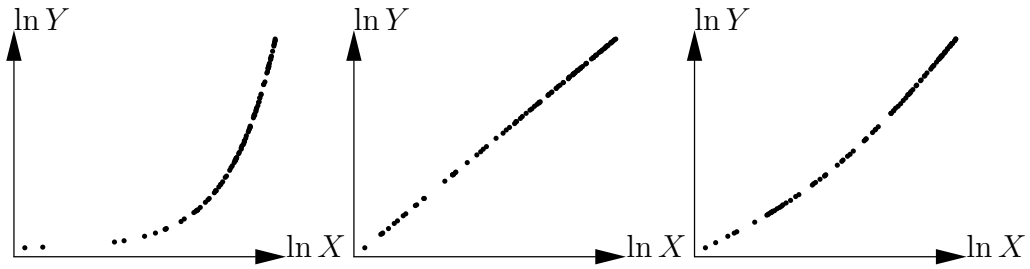


Figura 5.12 – *Aplicación de la transformación de variables* $X \to \ln X$, $Y \to \ln Y$ *a las nubes de puntos representadas en la Figura 5.10.*

sentido del modelo lineal, que describe la linealidad respecto de los coeficientes del modelo (por tanto, el modelo $Y = a + bX^2$ es lineal en el sentido del modelo lineal mientras que este modelo define una relación no lineal entre X y Y).

Cuando ninguna transformación de variables permite linealizar la relación entre X y Y, lo mejor es ajustar el modelo y determinar visualmente si la curva ajustada pasa por el medio de la nube de puntos adaptándose a su forma. En este caso, convendrá además evaluar el gráfico de los residuos en función de los valores predichos.

⑤

Exploración de la relación biomasa–diámetro: transformación de las variables

Utilicemos la transformación logarítmica para transformar simultáneamente el diámetro y la biomasa. El gráfico de la nube de puntos de datos transformados logarítmicamente se obtiene del modo siguiente:

```
plot(dat$dbh,dat$Btot,xlab="Diametro (cm)",ylab="Biomasa (t)",log="xy")
```

La nube de puntos resultante se muestra en la Figura 5.13. La transformación logarítmica linealizó la relación entre la biomasa y el diámetro: la relación entre $\ln(D)$ e $\ln(B)$ tiene la forma de una recta y la varianza de $\ln(B)$ no varía con el diámetro (como en la Figura 5.1A).

⑥

Exploración de la relación biomasa–D^2H: transformación de las variables

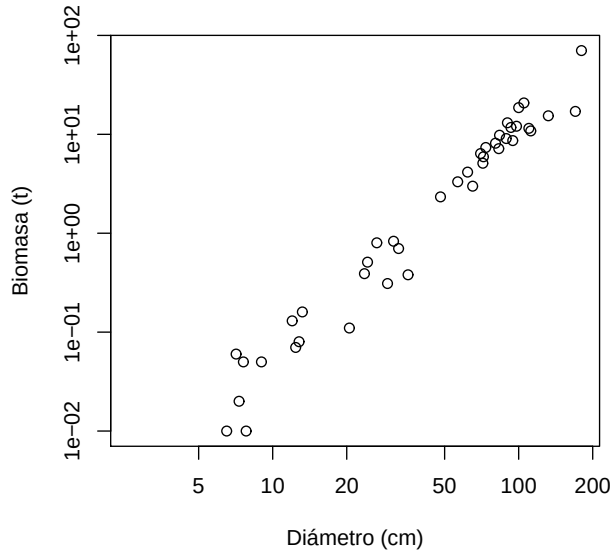


Figura 5.13 – *Nube de puntos (datos transformados logarítmicamente) de la biomasa seca total (toneladas) en función del diámetro a la altura del pecho (cm) para los 42 árboles medidos en Ghana por Henry et al. (2010).*

Utilicemos la transformación logarítmica para transformar simultáneamente D^2H y la biomasa. La nube de puntos de datos transformados logarítmicamente se obtiene del modo siguiente:

```
with(dat,plot(dbh^2*haut,Btot,log="xy",xlab="D2H (cm2.m)",ylab="Biomasa (t)"))
```

La nube de puntos resultante se muestra en la Figura 5.14. La transformación logarítmica linealizó la relación entre la biomasa y D^2H: la relación entre $\ln(D^2H)$ e $\ln(B)$ tiene la forma de una recta y la varianza de $\ln(B)$ no varía con D^2H (como en la figure 5.1A).

5.1.3. Catálogo de primitivos

Las síntesis de modelos realizado por Zianis *et al.* (2005) para Europa, por Henry *et al.* (2011) para África subsahariana o más específicamente por Hofstad (2005) para África austral permitirá hacerse una idea de la forma de los modelos de biomasa y volumen más frecuentes en la bibliografía. Los dos modelos que se ven con más frecuencia son: el modelo de potencia y el modelo polinomial (de grado dos o, como máximo, tres). Estos dos tipos de modelos serán entonces el punto de partida de la exploración gráfica de los datos para la elaboración de un modelo de volumen o de biomasa. El modelo de potencia $Y = aX^b$ se conoce también como una relación alométrica y existen bastantes interpretaciones biológicas del mismo (Gould, 1979; Franc *et al.*, 2000, § 1.1.5). En particular, la "teoría de escalamiento metabólico" (Enquist *et al.*, 1998, 1999; West *et al.*, 1997, 1999) predice de forma teórica y apoyándose en una descripción fractal de la estructura interna de los árboles, que la biomasa de un árbol está vinculada a su diámetro por una relación de potencia con un exponente igual a $8/3 \approx 2{,}67$:

$$B \propto \rho D^{8/3}$$

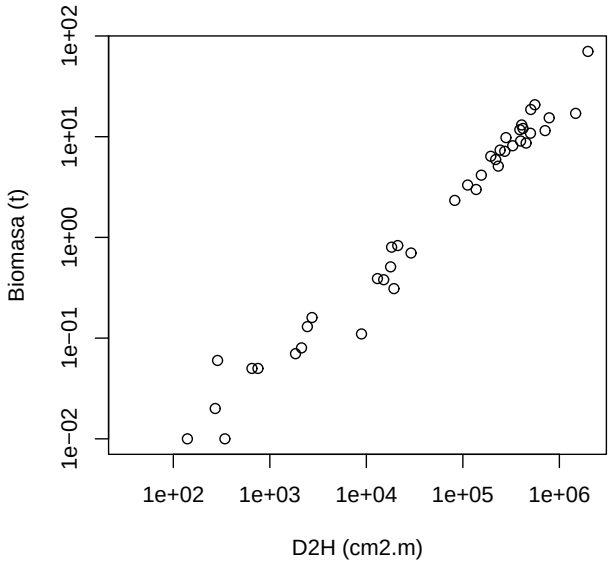


Figura 5.14 – *Nube de puntos (datos transformados logarítmicamente) de la biomasa seca total (toneladas) en función de D^2H, donde D es el diámetro a la altura del pecho (cm) y H la altura (m) para los 42 árboles medidos en Ghana por Henry et al. (2010).*

donde ρ es la densidad específica de la madera. Aun cuando se haya cuestionado mucho la "teoría de escalamiento metabólico" (Muller-Landau *et al.*, 2006), ésta tiene al menos el mérito de dar una base biológica a la relación de potencia que se observa con frecuencia.

Además del modelo de potencia $B = aD^b$ y el modelo polinomial de segundo grado $B = a_0 + a_1 D + a_2 D^2$, y sin pretender ser exhaustivos, los modelos de biomasa siguientes suelen encontrarse con frecuencia (Yamakura *et al.*, 1986; Brown *et al.*, 1989; Brown, 1997; Martinez-Yrizar *et al.*, 1992; Araújo *et al.*, 1999; Nelson *et al.*, 1999; Ketterings *et al.*, 2001; Chave *et al.*, 2001, 2005; Nogueira *et al.*, 2008; Basuki *et al.*, 2009; Návar, 2009; Djomo *et al.*, 2010; Henry *et al.*, 2010):

1. modelo de dos entradas en forma de potencia con respecto a la variable D^2H: $B = a(\rho D^2 H)^b$
2. modelo de dos entradas: $\ln(B) = a_0 + a_1 \ln(D) + a_2 \ln(H) + a_3 \ln(\rho)$
3. modelo de una entrada: $\ln(B) = a_0 + a_1 \ln(D) + a_2[\ln(D)]^2 + a_3[\ln(D)]^3 + a_4 \ln(\rho)$,

donde ρ es la densidad específica de la madera. Dejando de lado el factor de forma, la variable D^2H representa al volumen del tronco, lo que explica su frecuente uso como variable explicativa. La segunda ecuación puede verse como una generalización de la primera. En efecto, aplicando la transformación logarítmica, la primera ecuación equivale a: $\ln(B) = \ln(a) + 2b\ln(D) + b\ln(H) + b\ln(\rho)$. La primera ecuación es pues equivalente a la segunda en el caso particular donde $a_2 = a_3 = a_1/2$. Por último, la tercera ecuación puede verse como una generalización del modelo de potencia $B = aD^b$.

En líneas más generales, el Cuadro 5.1 recapitula cierto número de funciones que pueden modelar la relación entre dos variables. Es apropiado cuando las variables han sido transformadas para linealizar la relación entre X e Y. Cabe señalar que el modelo de potencia modificado no es más que una reescritura del modelo exponencial y que el modelo de la raíz no es más que la reescritura del modelo exponencial modificado. Asimismo se observará

Cuadro 5.1 – *Algunos modelos que vinculan dos variables.*

Nombre	Écuación	Transformación
Modelos polinomiales		
lineal	$Y = a + bX$	identidad
parabólico o cuadrático	$Y = a + bX + cX^2$	
polinomial de orden p	$Y = a_0 + a_1X + a_2X^2 + \ldots + a_pX^p$	
Modelos exponenciales		
exponencial o de Malthus	$Y = a\exp(bX)$	$Y' = \ln Y$, $X' = X$
exponencial modificado	$Y = a\exp(b/X)$	$Y' = \ln Y$, $X' = 1/X$
logaritmo	$Y = a + b\ln X$	$Y' = Y$, $X' = \ln X$
log recíproco	$Y = 1/(a + b\ln X)$	$Y' = 1/Y$, $X' = \ln X$
presión de vapor	$Y = \exp(a + b/X + c\ln X)$	
Modelos ley de potencia		
potencia	$Y = aX^b$	$Y' = \ln Y$, $X' = \ln X$
potencia modificada	$Y = ab^X$	$Y' = \ln Y$, $X' = X$
potencia desplazada	$Y = a(X - b)^c$	
geométrico	$Y = aX^{bX}$	$Y' = \ln Y$, $X' = X\ln X$
geométrico modificado	$Y = aX^{b/X}$	$Y' = \ln Y$, $X' = (\ln X)/X$
raíz	$Y = ab^{1/X}$	$Y' = \ln Y$, $X' = 1/X$
de Hoerl	$Y = ab^X X^c$	
de Hoerl modificado	$Y = ab^{1/X} X^c$	
Modelos de producción–densidad		
inverso	$Y = 1/(a + bX)$	$Y' = 1/Y$, $X' = X$
inverso cuadrático	$Y = 1/(a + bX + cX^2)$	
de Bleasdale	$Y = (a + bX)^{-1/c}$	
de Harris	$Y = 1/(a + bX^c)$	
Modelos de crecimiento		
de crecimiento saturado	$Y = aX/(b + X)$	$Y' = X/Y$, $X' = X$
mononuclear o de Mitscherlisch	$Y = a[b - \exp(-cX)]$	
Modelos sigmoidales		
de Gompertz	$Y = a\exp[-b\exp(-cX)]$	
de Sloboda	$Y = a\exp[-b\exp(-cX^d)]$	
logístico o de Verhulst	$Y = a/[1 + b\exp(-cX)]$	
de Nelder	$Y = a/[1 + b\exp(-cX)]^{1/d}$	
de von Bertalanffy	$Y = a[1 - b\exp(-cX)]^3$	
de Chapman-Richards	$Y = a[1 - b\exp(-cX)]^d$	
de Hossfeld	$Y = a/[1 + b(1 + cX)^{-1/d}]$	
de Levakovic	$Y = a/[1 + b(1 + cX)^{-1/d}]^{1/e}$	
du factor multiplicativo múltiple	$Y = (ab + cX^d)/(b + X^d)$	
de Johnson-Schumacher	$Y = a\exp[-1/(b + cX)]$	
de Lundqvist-Matérn o de Korf	$Y = a\exp[-(b + cX)^d]$	
de Weibull	$Y = a - b\exp(-cX^d)$	
Modelos diversos		
hiperbólico	$Y = a + b/X$	$Y' = Y$, $X' = 1/X$
sinusoidal	$Y = a + b\cos(cX + d)$	
de capacidad de calor	$Y = a + bX + c/X^2$	
de Gauss	$Y = a\exp[-(X - b)^2/(2c^2)]$	
de fracción racional	$Y = (a + bX)/(1 + cX + dX^2)$	

que una gran parte de estos modelos son sólo casos particulares de modelos más complejos (y que conllevan más parámetros). Por ejemplo, el modelo lineal no es más que un caso particular del modelo polinomial, el modelo de Gompertz no es más que un caso particular del modelo de Sloboda, etc.

El modelo polinomial de tipo p debe usarse con prudencia puesto que los polinomios son capaces de ajustarse a cualquier forma siempre que el grado p sea suficientemente elevado (las funciones usuales se pueden descomponer todas en una base de polinomios: es el principio del desarrollo limitado). Concretamente, se puede tener un polinomio que se adapte muy bien a la forma de la nube de puntos en el ámbito de los valores de datos disponibles pero que tome una forma muy inverosímil fuera de dicho ámbito. En otras palabras, el modelo polinomial puede presentar peligros de extrapolación, mucho mayores cuanto más importante sea el grado de p. En la práctica, se debe evitar a toda costa ajustar polinomios de grado superior a 3.

5.2. Exploración de la varianza

Consideremos ahora el término de error ε del modelo relaciona la variable Y que hay que explicar a una variable explicativa X. La exploración de la forma de la varianza equivale prácticamente a responder a la pregunta: ¿la varianza de los residuos es constante (homocedasticidad) o no (heterocedasticidad)? La respuesta a esta pregunta depende implícitamente de la forma precisa de la relación que se utilizará para ajustar el modelo. Como ejemplo, para la relación de potencia $Y = aX^b$, se puede

- o ajustar el modelo no lineal $Y = aX^b + \varepsilon$, lo que equivale a estimar directamente los parámetros a y b;
- o bien ajustar el modelo lineal $Y' = a' + bX' + \varepsilon$ a los datos transformados $Y' = \ln Y$ y $X' = \ln X$, lo que equivale a estimar los parámetros $a' = \ln a$ y b.

Ambas opciones, desde luego, no son intercambiables ya que el término de error ε (que supondremos que sigue una distribución normal de desviación estándar constante) no desempeña el mismo papel en ambos casos. En el primero, tenemos un error aditivo con respecto al modelo de potencia. En el segundo, tenemos un error aditivo con respecto al modelo liniealizado así que, si volvemos al modelo de potencia:

$$Y = \exp(Y') = aX^b \exp(\varepsilon) = aX^b \varepsilon'$$

lo que corresponde a un error multiplicativo con respecto al modelo de potencia, donde ε' sigue una ley log-normal. La diferencia entre estas dos opciones se representa en la Figura 5.15. El error aditivo se traduce en una varianza constante en el gráfico (A) de Y en función de X y por una varianza decreciente con X en el gráfico (C) de estos mismos datos transformados logarítmicamante. El error multiplicativo se refleja en una varianza creciente con X en el gráfico (B) de Y en función de X y por una varianza constante en el gráfico (D) de estos mismos datos transformados logarítmicamante.

De este modo, el proceso de linealización del modelo que relaciona Y a X mediante una transformación de variables afecta tanto la forma de la relación media como el término de error. Por otro lado, esta propiedad puede aprovecharse para estabilizar los residuos que varían con X afín de volverlos constantes pero este punto se abordará en el Capítulo siguiente. Por el momento, intentamos explorar la forma del error $Y - \mathrm{E}(Y)$ en función de X, sin procurar transformar las variables X e Y.

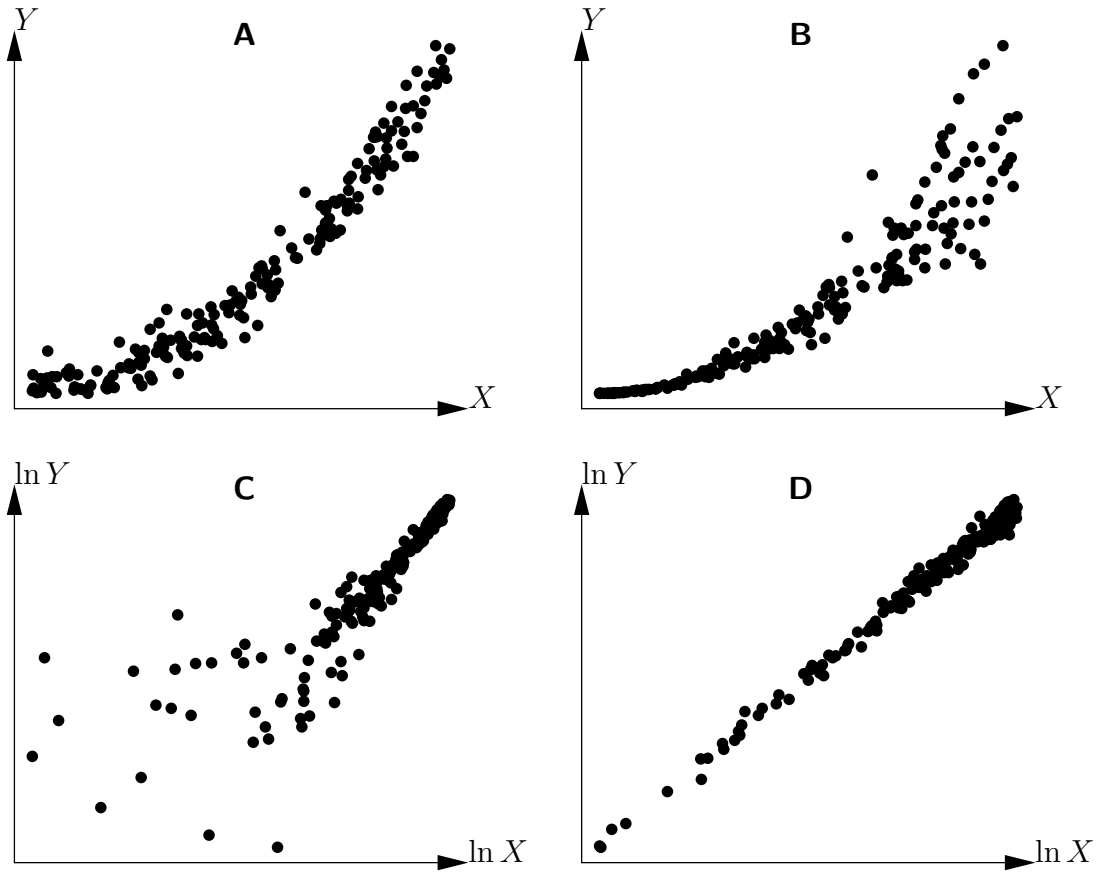


Figura 5.15 – *Modelo de potencia con error aditivo (A y C) o multiplicativo (B y D). El gráfico (C) (respectivamente D) resulta del gráfico (A) (respectivamente B) por transformación de las variables* $X \to \ln X$ *e* $Y \to \ln Y$.

Como forma de la relación media $\mathrm{E}(Y) = f(X)$ fue determinada anteriormente en forma gráfica, basta examinar visualmente en el gráfico de Y en función de X si los puntos se reparten igualmente de cada lado de la curva f independientemente del valor de X . Los gráficos (A) y (B) de la Figura 5.1, por ejemplo, muestran el caso de residuos de varianza constante para todos los valores de X, mientras que los gráficos (C) y (D) de esa misma Figura ilustran el caso de residuos cuya varianza aumenta con X. Relaciones más complejas, como las de la Figura 5.16, también pueden concebirse. Tratándose de la Figura 5.16, la varianza de los residuos fluctúa en forma periódica con X. En la práctica hay pocas posibilidades de encontrar tales situaciones en el contexto de los modelos de biomasa o de volumen. En casi todos los casos, habrá que escoger entre dos situaciones: la varianza de los residuos es constante o aumenta con X. En el primer caso, ya no hay nada que hacer. En el segundo, se tratará de precisar la forma exacta de la relación entre X y la varianza de los residuos pero se adoptará de plano un modelo de potencia para vincular la varianza de los residuos ε a X:

$$\mathrm{Var}(\varepsilon) = \alpha X^{\beta}$$

Los valores de los coeficientes α y β se estimarán al mismo tiempo que los otros coeficientes del modelo durante la fase de ajuste del modelo, que se tratará en el próximo Capítulo.

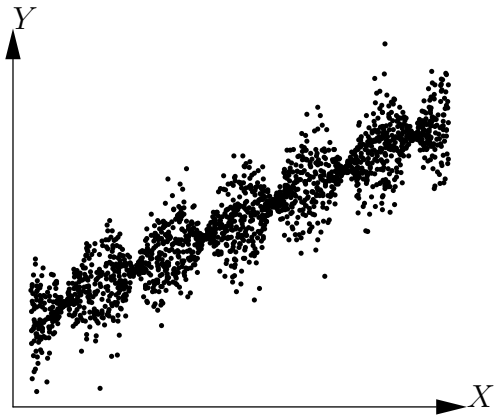


Figura 5.16 – *Gráfico de una nube de puntos generados por el modelo* $Y = a + bX + \varepsilon$*, donde* ε *sigue una distribución normal de media cero y de una desviación estándar proporcional al coseno de* X*.*

5.3. La exploración no es una selección

Para concluir, queremos precisar que la exploración gráfica no pretende seleccionar una sola forma de modelo sino más bien separar los modelos que son aceptables para describir el conjunto de datos de aquellos que no lo son. En vez de procurar seleccionar "el" modelo que sea mejor para describir los datos, hay que intentar más bien seleccionar tres o cuatro modelos posibles que permitan describir los datos. La elección final entre esos tres o cuatro modelos identificados durante la exploración gráfica se hará después de la fase de ajuste de los datos que veremos en el Capítulo siguiente.

6
Ajuste del modelo

El ajuste de un modelo consiste en estimar sus parámetros a partir de los datos. Esto implica que los datos están disponibles y en la presentación adecuada y que se conoce la expresión matemática del modelo que hay que ajustar. Por ejemplo, ajustar el modelo de potencia $B = aD^b$ consiste en estimar los coeficientes a y b a partir de un conjunto de datos que dan los valores B_i y D_i de la biomasa y del diámetro de n árboles ($i = 1, \ldots, n$). La variable de respuesta (también llamada en la bibliografía variable de salida, variable de interés, variable dependiente) del modelo es la variable que predice el modelo. Hay sólo una. En el marco del presente manual, la variable de respuesta será siempre un volumen o una biomasa. Las variables explicativas son variables usadas para predecir la variable de respuesta. Puede haber varias y su número se indica con p. No hay que confundir las variables explicativas y las entradas de datos del modelo. El modelo $B = a(D^2H)^b$ contiene una única variable explicativa (a saber D^2H) pero dos entradas (el diámetro D la altura H). Al contrario, el modelo $B = a_0 + a_1D + a_2D^2$ contiene dos variables explicativas (D y D^2) pero una sola entrada (el diámetro D). A cada variable explicativa va asociado un coeficiente que hay que estimar. A ello se agrega, cuando corresponde, una intersección o un coeficiente multiplicador, de forma tal que el número total de coeficientes por estimar en un modelo con p variables explicativas será p o $p + 1$.

Una observación consiste en el dato de la variable de respuesta (volumen o biomasa) y de las variables explicativas para un árbol. Para retomar el ejemplo del modelo $B = aD^b$, una observación consiste en el doblete (B_i, D_i). La cantidad de observaciones es pues n. Una observación se deriva de una medición sobre el terreno. La predicción del modelo es el valor de la variable de respuesta predicha para el modelo dadas las variables explicativas. Una predicción se deriva de un cálculo. Por ejemplo, la predicción del modelo $B = aD^b$ para un árbol de diámetro D_i es $\hat{B}_i = aD_i^b$. Hay tantas predicciones como observaciones. Un concepto clave del ajuste de los modelos es el residuo. El residuo o error residual es la diferencia entre el valor observado de la variable de respuesta y su predicción. Siempre para el mismo ejemplo, el residuo de la i-ésima observación es: $\varepsilon_i = B_i - \hat{B}_i = B_i - aD_i^b$. Hay tantos residuos como observaciones. El ajuste de un modelo será mucho mejor cuanto menores sean los residuos. Además, las propiedades estadísticas del modelo se derivarán de las propiedades que los residuos hayan tenido que verificar *a priori*, en particular la forma de su distribución. El tipo de ajuste del modelo dependerá directamente de las propiedades

de sus residuos.

En todos los modelos que veremos, se supondrá que las observaciones son *independientes* o, lo que es lo mismo, se supondrá que los residuos son independientes: para todo $i \neq j$, ε_i se supone que es independiente de ε_j. Esta propiedad de independencia es relativamente fácil de garantizar *por medio* del protocolo de muestreo. Típicamente habrá que asegurarse de que las características de un árbol medido en un lugar determinado no influyan en las características de otro árbol de la muestra. En general, seleccionar para la muestra árboles que estén bastante alejados entre sí basta para garantizar esta propiedad de independencia. Si los residuos no son independientes, se puede modificar el modelo para tenerlo en cuenta. Por ejemplo, se podrá introducir una estructura de dependencia espacial en los residuos para considerar una autocorrelación espacial de las mediciones. No abordaremos estos modelos porque son muy complejos de poner en práctica.

En todos los modelos que veremos, se partirá además del supuesto de que los residuos tienen una distribución *normal* de esperanza cero. La media cero de los residuos es en realidad una propiedad que se deriva automáticamente del ajuste del modelo y que garantiza que las predicciones no estén sesgadas. Son los residuos, y no las observaciones, los que se supone que tienen una distribución normal. Para los datos de volumen o biomasa, esta hipótesis no es en absoluto restrictiva. En el caso poco probable en que la distribución de los residuos se alejara mucho de una distribución normal, podríamos eventualmente considerar el ajuste de otros tipos de modelos, como el modelo lineal generalizado, pero eso no se abordará dentro del marco de este manual. La hipótesis de independencia y de distribución normal de los residuos son las dos primeras que sustentan el ajuste de los modelos. Conviene comprobar que estas dos hipótesis estén realmente verificadas. Más tarde veremos una tercera hipótesis. En la medida en que dichas hipótesis se refieren a los residuos del modelo y no a las observaciones, no se pueden probar hasta que no se hayan calculados los residuos, es decir, hasta que no se haya ajustado el modelo. Se trata pues de hipótesis que se verifican *a posteriori*, después de ajustar el modelo. Asimismo, los modelos que veremos son *robustos* con respecto a estas hipótesis, es decir que las calidades de predicción de los modelos ajustados siguen siendo correctas aunque las hipótesis de independencia y de distribución normal de los residuos no hayan sido completamente verificadas. Por este motivo no trataremos de probar de manera muy formal esas dos hipótesis. En la práctica nos contentaremos con una verificación visual basada en los gráficos.

6.1. Ajuste de un modelo lineal

El modelo lineal es el más simple de los modelos por ajustar. El adjetivo *lineal* significa aquí que el modelo depende *linealmente* de sus coeficientes. Por ejemplo, $Y = a + bX^2$ y $Y = a + b\ln(X)$ son modelos lineales puesto que la variable de respuesta Y depende linealmente de los coeficientes a y b, aun cuando Y no depende linealmente de la variable explicativa X. Por el contrario, $Y = aX^b$ no es un modelo lineal porque Y no depende linealmente del coeficiente b. Otra propiedad del modelo lineal es que el residuo es aditivo. Para destacar esto, se escribe explícitamente el residuo ε en la expresión del modelo. Por ejemplo, para una regresión lineal de Y con respecto a X, escribiremos: $Y = a + bX + \varepsilon$.

6.1.1. Regresión lineal simple

La regresión lineal simple es el más simple de los modelos lineales. Supone (*i*) que no hay más que una sola variable explicativa X, (*ii*) que la relación entre la variable de respuesta

Y y X tiene la forma de una recta:

$$Y = a + bX + \varepsilon$$

donde a es la intersección de la recta y b su pendiente, y (*iii*) que los residuos tienen una varianza constante: $\mathrm{Var}(\varepsilon) = \sigma^2$. Por ejemplo, el modelo

$$\ln(B) = a + b\ln(D) + \varepsilon \tag{6.1}$$

es un ejemplo de regresión lineal simple, que tiene como variable de respuesta $Y = \ln(B)$ y como variable explicativa $X = \ln(D)$. Corresponde a un modelo de potencia para la biomasa: $B = \exp(a)D^b$. Este modelo se usa frecuentemente para ajustar un modelo de biomasa monoespecífico. Otro ejemplo es el modelo de biomasa de dos entradas:

$$\ln(B) = a + b\ln(D^2H) + \varepsilon \tag{6.2}$$

La hipótesis de varianza constante de los residuos se suma a las dos hipótesis de independencia y de distribución normal (se habla también de homocedasticidad). Se resumen las tres hipótesis al escribir:

$$\varepsilon \underset{\text{i.i.d.}}{\sim} \mathcal{N}(0,\, \sigma)$$

donde $\mathcal{N}(\mu,\, \sigma)$ designa la ley normal de esperanza μ y la desviación estándar σ, el tilde "$\sim$" significa "está distribuido según", e "i.i.d." es la abreviatura "independiente e idénticamente distribuido".

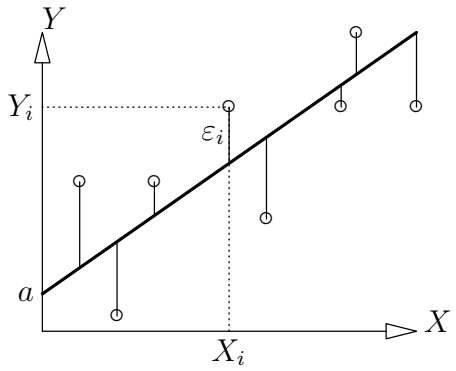


Figura 6.1 – *Esquema de las observaciones (puntos), de la recta de regresión (trazo grueso) y de los residuos (trazos finos).*

Estimación de los coeficientes

La Figura 6.1 esquematiza las observaciones y la recta de los valores predichos. El mejor ajuste será el que minimice el error residual. Se pueden considerar diversas formas de cuantificar dicho error. Desde un punto de vista matemático, eso equivale a elegir una norma para medir ε, y varias normas podrían servir para ello. La que suele usarse es la norma L_2, que equivale a cuantificar la diferencia residual entre las observaciones y las predicciones mediante la suma de los cuadrados de los residuos, lo que también se denomina suma de cuadrados o suma de cuadrado del error (SCE):

$$\mathrm{SCE}(a, b) = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} (Y_i - a - bX_i)^2$$

El mejor ajuste es pues aquel que minimiza la SCE. En otras palabras, las estimaciones $\hat{a}$ y $\hat{b}$ de los coeficientes a y b son los valores de a y b que minimizan la suma de los cuadrados de las diferencias:

$$(\hat{a},\, \hat{b}) = \arg\min_{(a,\, b)} \mathrm{SCE}(a, b)$$

Este mínimo se obtiene calculando las derivadas parciales de SCE con respecto a a y b, y al buscar los valores de a y b que anulan esas derivadas parciales. Los cálculos simples dan los resultados siguientes: $\hat{b} = \widehat{\mathrm{Cov}}(X,\, Y)/S_X^2$ y $\hat{a} = \bar{Y} - \hat{b}\bar{X}$, donde $\bar{X} = (\sum_{i=1}^n X_i)/n$ es la media empírica de la variable explicativa, $\bar{Y} = (\sum_{i=1}^n Y_i)/n$ es la media empírica de de la variable de respuesta,

$$S_X^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

es la varianza empírica de la variable explicativa, y

$$\widehat{\mathrm{Cov}}(X,\, Y) = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})$$

es la covarianza empírica entre la variable explicativa y la variable de respuesta. La estimación de la varianza residual, por su parte, es:

$$\hat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n}(Y_i - \hat{a} - \hat{b}X_i)^2 = \frac{\mathrm{SCE}(\hat{a},\, \hat{b})}{n-2}$$

Puesto que este método de estimación de los coeficientes se basa en la minimización de la suma de los cuadrados de las diferencias, se lo llama método *de los mínimos cuadrados* (a veces se especifica como "mínimos cuadrados ordinarios", para diferenciarlo de los mínimos cuadrados ponderados que veremos en § 6.1.3). La ventaja de este método de estimación es que da una expresión explícita de los coeficientes estimados.

Interpretación de los resultados de una regresión

Si se ajusta una regresión lineal simple, hay que analizar varias salidas. El coeficiente de determinación, en general llamado R^2, mide la calidad del ajuste. El R^2 está directamente vinculado a la varianza residual porque:

$$R^2 = 1 - \frac{\hat{\sigma}^2(n-2)/n}{S_Y^2}$$

donde $S_Y^2 = [\sum_{i=1}^n (Y_i - \bar{Y})^2]/n$ es la varianza empírica de Y. La diferencia $S_Y^2 - \hat{\sigma}^2(n-2)/n$ entre la varianza de Y y la varianza residual representa la varianza explicada por el modelo. El coeficiente de determinación R^2 se interpreta pues como la razón entre la varianza explicada por el modelo y la varianza total. Está comprendido entre 0 y 1 y, cuanto más próximo de uno es, mejor es la calidad del ajuste. En el caso de una regresión lineal simple, y únicamente en ese caso, R^2 es también igual al cuadrado del coeficiente de correlación lineal (también llamado coeficiente de Pearson) entre X y Y. En el Capítulo 5 hemos visto (en particular en la Figura 5.2) los límites de la interpretación de R^2.

Además de los valores estimados de los coeficientes a y b, el ajuste del modelo brinda también la desviación estándar de estas estimaciones (es decir, las desviaciones estándar de los estimadores $\hat{a}$ y $\hat{b}$), así como los resultados de las pruebas de significancia de estos coeficientes. Hay una prueba para la intersección a, que prueba la hipótesis nula $a = 0$, y también una prueba para la pendiente b, que prueba la hipótesis nula $b = 0$.

Por último, hay que analizar el resultado de la prueba de significancia global del modelo. Este test se basa en la descomposición de la varianza total de Y como la suma de la varianza explicada por el modelo y de la varianza residual. Como en un análisis de varianza, se usa la prueba de Fisher que usa como estadístico de prueba una relación ponderada de la varianza explicada sobre la varianza residual. En el caso de la regresión lineal simple, y únicamente en ese caso, la prueba de significatividad global del modelo da el mismo resultado que la prueba de la hipótesis nula $b = 0$. Esto se comprende intuitivamente: una recta que une X a Y sólo es significativa si la pendiente de dicha recta no es nula.

Verificación de las hipótesis

El ajuste del modelo se logra verificando que se han comprobado las hipótesis planteadas *a priori* sobre los residuos. No volveremos a abordar la hipótesis de la independencia de los residuos, que consideramos verificada gracias al plan de muestreo adoptado. Eventualmente, si existiera un orden natural en las observaciones, se podría usar el test de Durbin-Watson para probar que los residuos son realmente independientes (Durbin & Watson, 1971). La hipótesis de distribución normal de los residuos se verifica visualmente a partir del gráfico cuantil-cuantil. Éste representa los cuantils empíricos de los residuos en función de los cuantils teóricos de la distribución normal estándar. Si la hipótesis de distribución normal de los residuos es aceptable, los puntos se alinean aproximadamente a lo largo de una recta, como en la 6.2 (gráfico de la derecha).

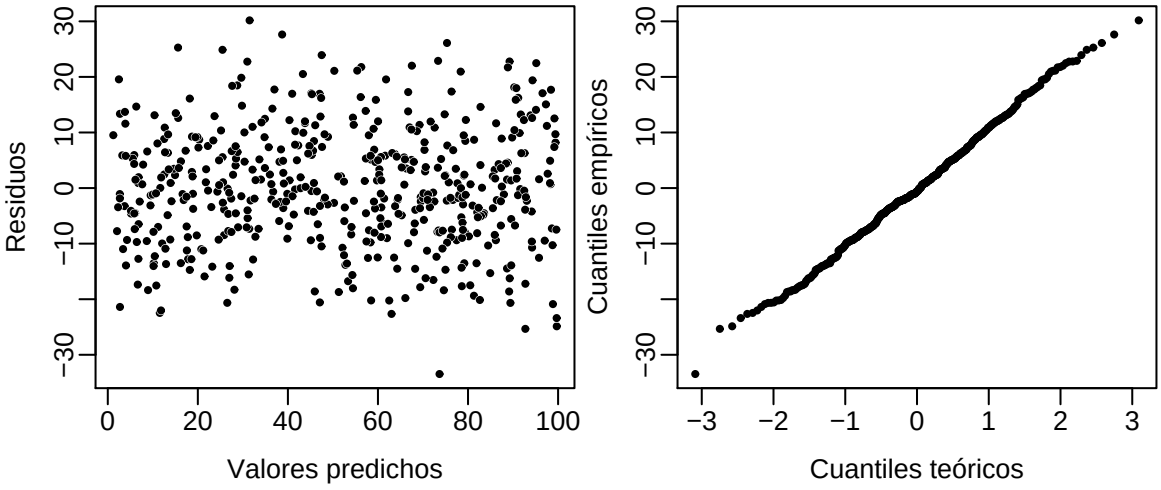


Figura 6.2 – *Apariencia del gráfico de los residuos en función de los valores predichos (a la izquierda) y del gráfico cuantile-cuantile (a la derecha) cuando las hipótesis de distribución normal y de varianza constante de los residuos se han verificado bien.*

En el caso del ajuste de modelos de volumen o de biomasa, la hipótesis más importante que hay que verificar es la de la constancia de la varianza de los residuos. Se la verifica visualmente trazando la nube de puntos de los residuos $\varepsilon_i = Y_i - \hat{Y}_i$ en función de los valores predichos $\hat{Y}_i = \hat{a} + \hat{b}X_i$. Si la varianza de los residuos es constante, dicha nube no debe mostrar ninguna tendencia, ninguna estructuración particular. Por ejemplo, es el caso del gráfico de la izquierda de la Figura 6.2. Por el contrario, si aparece una estructuración particular en dicha nube, cabe replantearse la hipótesis. Ese es el caso, por ejemplo, en la Figura 6.3, donde la nube de puntos de los residuos, en función de los valores predichos, tiene forma de embudo. Esta forma es típica de un aumento de la varianza residual con la

variable explicativa (es lo que se llama heterocedasticidad). Si tal es el caso, hay que ajustar otro modelo distinto de la regresión lineal simple.

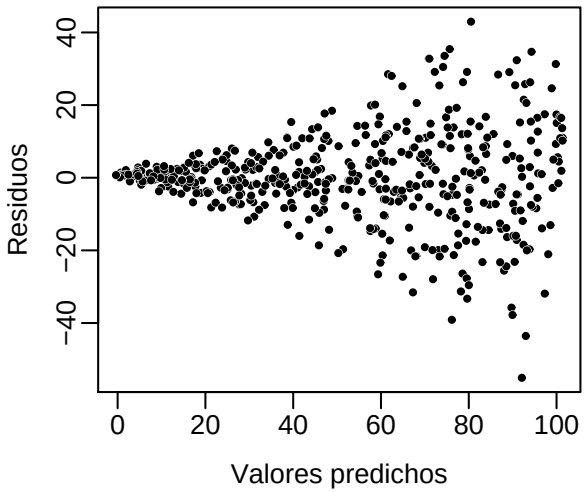


Figura 6.3 – *Apariencia del gráfico de los residuos en función de los valores predichos cuando los residuos no tienen una varianza constante (heterocedasticidad).*

En el caso de datos biológicos tales como el volumen de la biomasa de los árboles, la heterocedasticidad es la regla y la homocedasticidad la excepción. Esto significa simplemente que la variabilidad de la biomasa (o del volumen) de los árboles es mucho mayor cuanto más grandes son ellos. Dicha variabilidad creciente de la biomasa de los individuos con su tamaño es un principio general en biología. Por tanto, en el caso de ajustar modelos de biomasa o volumen, la regresión lineal simple que usa la biomasa como variable de respuesta ($Y = B$) resultará generalmente poco útil. La transformación logarítmica (es decir, $Y = \ln(B)$) permite resolver este problema, de forma tal que las regresiones lineales que usemos para ajustar modelos serán casi siempre regresiones sobre datos transformados logarítmicamente. Volveremos luego a tratar con detalle este punto fundamental.

⑦

Regresión lineal simple entre $\ln(B)$ **y** $\ln(D)$

El análisis exploratorio (Línea roja 5) ha demostrado que la relación entre el logaritmo de la biomasa y la longitud del diámetro era lineal, con una varianza de $\ln(B)$ que era aproximativamente constante. Se puede entonces ajustar una regresión lineal simple para predecir $\ln(B)$ en función de $\ln(D)$:

$$\ln(B) = a + b\ln(D) + \varepsilon$$

con

$$\mathrm{Var}(\varepsilon) = \sigma^2$$

La regresión se ajusta mediante mínimos cuadrados ordinarios. Como no se puede aplicar la transformación logarítmica a un valor cero, los datos de biomasa nulos (cf. Línea roja 1) se retiran antes del conjunto de datos:

```
m <- lm(log(Btot)~I(log(dbh)),data=dat[dat$Btot>0,])
summary(m)
```

La desviación estándar residual es $\hat{\sigma} = 0{,}462$, R^2 es 0,9642 y el modelo es altamente significativo (prueba de Fisher: $F_{1,39} = 1051$, p-value $< 2{,}2 \times 10^{-16}$). Los valores de los coeficientes se dan en el Cuadro siguiente:

```
             Estimate  Std. Error  t value  Pr(>|t|)
(Intercept)  -8.42722     0.27915   -30.19    <2e-16  ***
I(log(dbh))   2.36104     0.07283    32.42    <2e-16  ***
```

La primera columna de este Cuadro da los valores de los coeficientes. El modelo se escribe entonces como: $\ln(B) = -8{,}42722 + 2{,}36104 \ln(D)$. La segunda columna da las desviaciones estándar de los estimadores de los coeficientes. La tercera columna da el valor del estadístico de prueba para la hipótesis nula que "el coeficiente es cero". Por último, la cuarta columna da el p-value de esta prueba. En nuestro caso, tanto la pendiente como la intersección son significativamente diferentes de cero.

Queda por comprobar gráficamente que se verifiquen las hipótesis de la regresión lineal:

```
plot(m,which=1:2)
```

El resultado se representa en la Figura 6.4. Aunque el gráfico cuantil-cuantil de los residuos parezca ligeramente estructurado, se considerará que las hipótesis de la regresión lineal simple se han respetado como corresponde.

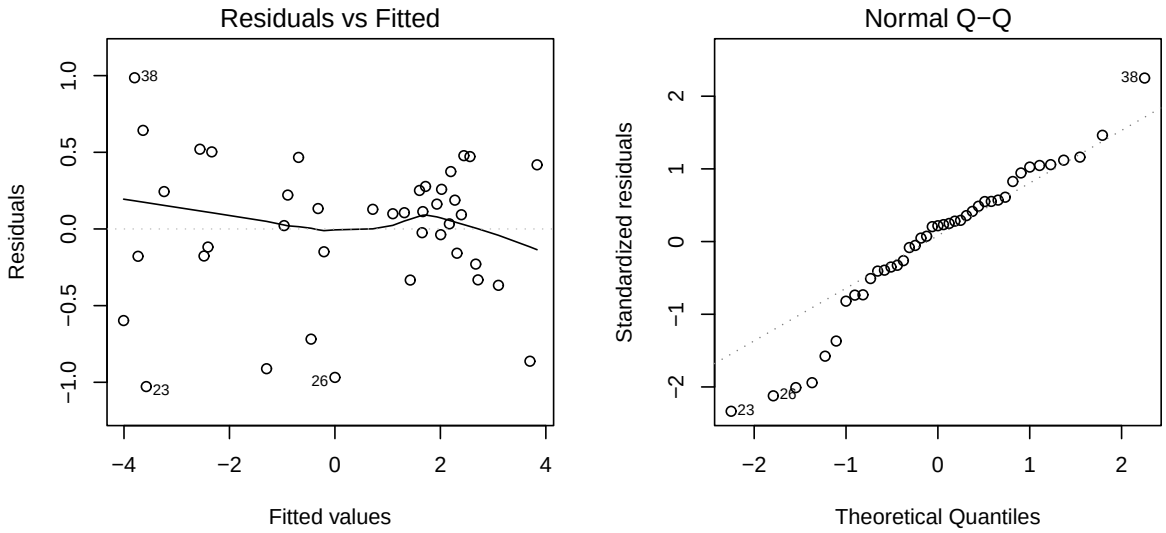


Figura 6.4 – *Gráfico de los residuos en función de los valores predichos (a la izquierda) y gráfico cuantil-cuantil (a la derecha) de los residuos de la regresión lineal simple de* $\ln(B)$ *con respecto a* $\ln(D)$ *ajustada a los 42 árboles medidos por Henry et al. (2010) en Ghana.*

⑧

Regresión lineal simple entre $\ln(B)$ e $\ln(D^2H)$

El análisis exploratorio (Línea roja 6) ha demostrado que la relación entre el logaritmo de la biomasa y el logaritmo de D^2H era lineal con una varianza aproximadamente constante de $\ln(B)$. Se puede ajustar entonces una regresión lineal simple para predecir $\ln(B)$ en función de $\ln(D^2H)$:

$$\ln(B) = a + b \ln(D^2H) + \varepsilon$$

con

$$\mathrm{Var}(\varepsilon) = \sigma^2$$

La regresión se ajusta mediante los mínimos cuadrados ordinarios. Como no puede aplicarse la transformación logarítmica a un valor cero, los datos de biomasa nulos (cf. Línea roja 1) se retiran primero del conjunto de datos:

```
m <- lm(log(Btot)~I(log(dbh^2*haut)),data=dat[dat$Btot>0,])
summary(m)
```

La desviación estándar es $\hat{\sigma} = 0{,}4084$, R^2 es 0,972 y el modelo es altamente significativo (prueba de Fisher: $F_{1,39} = 1356$, p-value $< 2{,}2 \times 10^{-16}$). Los valores de los coeficientes son los siguientes:

```
                  Estimate  Std. Error  t value  Pr(>|t|)
(Intercept)       -8.99427     0.26078   -34.49    <2e-16  ***
I(log(dbh^2*haut))  0.87238     0.02369    36.82    <2e-16  ***
```

La primera columna de este Cuadro da los valores de los coeficientes. El modelo se escribe entonces como: $\ln(B) = -8{,}99427 + 0{,}87238 \ln(D^2H)$. La segunda columna da las desviaciones estándar de los estimadores de los coeficientes. La tercera da el valor del estadístico para la prueba de la hipótesis nula que "el coeficiente es cero". Por último, la cuarta columna da el p-value para esta prueba. En nuestro caso, tanto la pendiente como la intersección son significativamente diferentes de cero.

Queda por verificar gráficamente que se verifiquen las hipótesis de la regresión lineal:

```
plot(m,which=1:2)
```

El resultado está representado en la Figura 6.5. Incluso si el gráfico de los residuos en función de los valores predichos parece ligeramente estructurado, se considerará que las hipótesis de la regresión lineal simple se han respetado como corresponde.

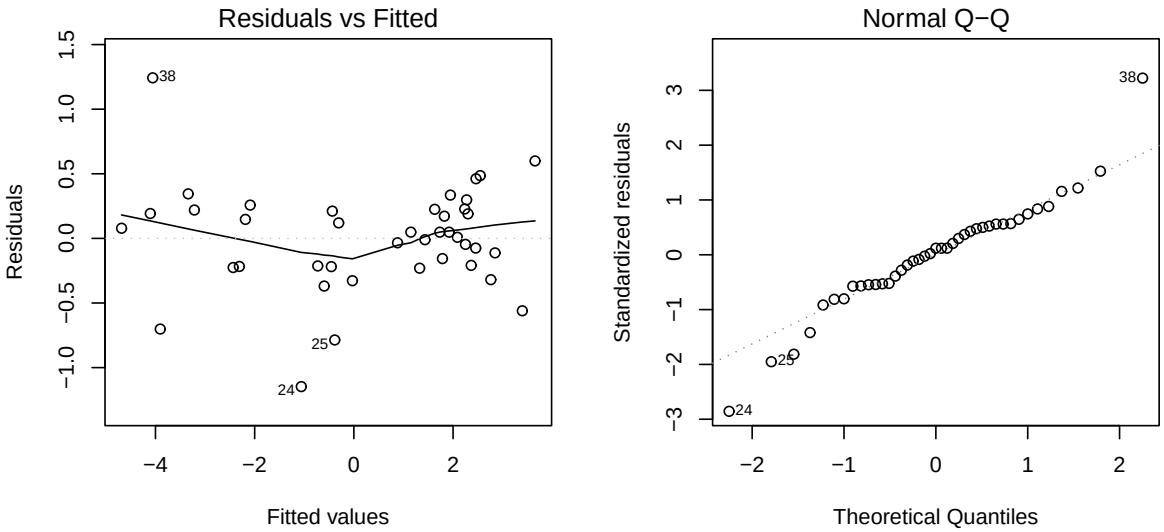


Figura 6.5 – *Gráfico de los residuos en función de los valores predichos (a la izquierda) y gráfico cuantil-cuantil (a la derecha) de los residuos de la regresión lineal simple de* $\ln(B)$ *con respecto a* $\ln(D^2H)$ *ajustada a los 42 árboles medidos por Henry et al. (2010) en Ghana.*

6.1.2. Regresión múltiple

La regresión múltiple es la extensión de la regresión lineal simple, cuando hay varias variables explicativas, y se escribe:

$$Y = a_0 + a_1 X_1 + a_2 X_2 + \ldots + a_p X_p + \varepsilon \tag{6.3}$$

donde Y es la variable de respuesta, $X_1, \ldots, X_p$ las p variables explicativas, $a_0, \ldots, a_p$ son los coeficientes por estimar, y ε es el error residual. Contando la intersección a_0, hay $p+1$ coeficientes por estimar. Como para la regresión lineal simple, se supone que la varianza de los residuos es constante, igual a σ^2:

$$\varepsilon \underset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma)$$

Los siguientes modelos de biomasa son ejemplos de regresión múltiple:

$$\ln(B) = a_0 + a_1 \ln(D^2 H) + a_2 \ln(\rho) + \varepsilon \tag{6.4}$$
$$\ln(B) = a_0 + a_1 \ln(D) + a_2 \ln(H) + \varepsilon \tag{6.5}$$
$$\ln(B) = a_0 + a_1 \ln(D) + a_2 \ln(H) + a_3 \ln(\rho) + \varepsilon \tag{6.6}$$
$$\ln(B) = a_0 + a_1 \ln(D) + a_2 [\ln(D)]^2 + a_3 [\ln(D)]^3 + \varepsilon \tag{6.7}$$
$$\ln(B) = a_0 + a_1 \ln(D) + a_2 [\ln(D)]^2 + a_3 [\ln(D)]^3 + a_4 \ln(\rho) + \varepsilon \tag{6.8}$$

donde ρ es la densidad de la madera. En todos estos ejemplos, la variable de respuesta es el logaritmo de la biomasa: $Y = \ln(B)$. El modelo (6.4) generaliza (6.2) al agregar la dependencia con respecto a la densidad específica de la madera: por lo general se preferirá (6.4) a (6.2) cuando el conjunto de datos sea pluriespecífico. El modelo (6.5) generaliza (6.2) al considerar que el exponente asociado a la altura H no es necesariamente igual a la mitad del exponente asociado al diámetro D. Así introduce un poco más de flexibilidad en la forma de la relación entre la biomasa y $D^2 H$. El modelo (6.6) generaliza (6.2) al considerar al mismo tiempo que hay varias especies y que la biomasa no es totalmente una potencia de $D^2 H$. El modelo (6.7) generaliza (6.1) al considerar que la relación entre $\ln(B)$ e $\ln(D)$ no es exactamente lineal. Ofrece así un poco más de flexibilidad en la forma de esta relación. El modelo (6.8) es una extensión de (6.7) para tener en cuenta la presencia de varias especies en el conjunto de datos.

Estimación de los coeficientes

Igual que para la regresión lineal simple, la estimación de los coeficientes de la regresión múltiple se basa en el método de mínimos cuadrados. Los estimadores $\hat{a}_0, \hat{a}_1, \ldots, \hat{a}_p$ son los valores de los coeficientes $a_0, a_1, \ldots, a_p$ que minimizan la suma de los cuadrados de las diferencias:

$$\text{SCE}(a_0, a_1, \ldots, a_p) = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} (Y_i - a_0 - a_1 X_{i1} - \ldots - a_p X_{ip})^2$$

donde X_{ij} es el valor de la j-ésima variable explicativa para la i-ésima observación ($i = 1, \ldots, n$ y $j = 1 \ \ldots, p$). Nuevamente las estimaciones de los coeficientes se obtienen calculando las derivadas parciales de SCE con respecto a los coeficientes y buscando los valores de los coeficientes que anulan esas derivadas parciales. Los cálculos no son más complicados que para la regresión lineal simple, siempre y cuando se los ponga en forma matricial. Digamos

que $\mathbf{X}$ es la matriz con n lineas y p columnas, llamada matriz de diseño, que reúne los valores observados de las variables explicativas:

$$\mathbf{X} = \begin{bmatrix} 1 & X_{11} & \cdots & X_{1p} \\ \vdots & \vdots & & \vdots \\ 1 & X_{n1} & \cdots & X_{np} \end{bmatrix}$$

Digamos que $\mathbf{Y} = {}^{\mathrm{t}}[Y_1, \ldots, Y_n]$ es el vector de los n valores observados de la variable de respuesta, y $\mathbf{a} = {}^{\mathrm{t}}[a_0, \ldots, a_p]$ el vector de los $p+1$ coeficientes por estimar. Entonces

$$\mathbf{Xa} = \begin{bmatrix} a_0 + a_1 X_{11} + \ldots + a_p X_{1p} \\ \vdots \\ a_0 + a_1 X_{n1} + \ldots + a_p X_{np} \end{bmatrix}$$

es el vector $\hat{\mathbf{Y}}$ de los n valores predichos para el modelo de la variable respuesta. Al usar esas notaciones de matrices, la suma de los cuadrados de las diferencias se escribe:

$$\mathrm{SCE}(\mathbf{a}) = {}^{\mathrm{t}}(\mathbf{Y} - \hat{\mathbf{Y}})(\mathbf{Y} - \hat{\mathbf{Y}}) = {}^{\mathrm{t}}(\mathbf{Y} - \mathbf{Xa})(\mathbf{Y} - \mathbf{Xa})$$

Al usar las reglas de cálculo diferencial matricial (Magnus & Neudecker, 2007), se obtiene finalmente:

$$\hat{\mathbf{a}} = \arg\min_{\mathbf{a}} \mathrm{SCE}(\mathbf{a}) = ({}^{\mathrm{t}}\mathbf{XX})^{-1\,\mathrm{t}}\mathbf{XY}$$

La estimación de la varianza residual, por su parte, es:

$$\hat{\sigma}^2 = \frac{\mathrm{SCE}(\hat{\mathbf{a}})}{n - p - 1}$$

Al igual que para la regresión lineal simple, este método de estimación tiene la ventaja de aportar una expresión explícita de los coeficientes estimados. La regresión lineal simple, al ser un caso particular de la regresión múltiple (caso en que $p = 1$), podemos cerciorarnos de que las expresiones matriciales de las estimaciones de los coeficientes y de $\hat{\sigma}$ vuelven a dar, cuando $p = 1$, las expresiones dadas anteriormente en el caso de la regresión lineal simple.

Interpretación de los resultados de una regresión múltiple

Al igual que para la regresión lineal simple, el ajuste de una regresión múltiple da un coeficiente de determinación R^2 que representa la parte de varianza explicada por el modelo; los valores $\hat{\mathbf{a}}$ de los coeficientes $a_0, a_1, \ldots, a_p$ del modelo; las desviaciones estándar de dichas estimaciones; los resultados de las pruebas de significatividad de los coeficientes (hay $p+1$, una por cada coeficiente, hipótesis nulas $a_i = 0$ pour $i = 0, \ldots, p$); y el resultado de la prueba de significatividad global del modelo.

Igual que antes, el valor de R^2 está comprendido entre 0 y 1. Su valor será mucho mayor cuanto mejor sea la calidad de ajuste del modelo. No obstante, hay que tener cuidado porque el valor de R^2 aumenta automáticamente con el número de variables explicativas usadas. Por ejemplo, si se predice Y para un polinomio de grado p en X,

$$Y = a_0 + a_1 X + a_2 X^2 + \ldots + a_p X^p$$

R^2 será automáticamente una función creciente del grado p. Esto puede dar la ilusión de que una regresión polinomial será mucho mejor cuanto más elevado sea el grado p del polinomio. Evidentemente, no es así. Un valor demasiado elevado del grado p conllevará una sobreparametrización del modelo. En otras palabras, R^2 no es un criterio válido para hacer una selección de modelo. Volveremos a este punto en la Sección 6.3.

Verificación de las hipótesis

Al igual que la regresión lineal simple, la regresión múltiple se basa en tres hipótesis: independencia de los residuos, distribución normal de los residuos y varianza constante de los residuos. Estas hipótesis se verifican exactamente del mismo modo que para la regresión lineal simple. Para comprobar la distribución normal de los residuos, haremos un gráfico cuantil-cuantil y nos aseguraremos visualmente de que la nube de puntos forma una recta. Para verificar la varianza constante de los residuos, haremos un gráfico de los residuos en función de los valores predichos y nos aseguraremos visualmente de que la nube de puntos no presente ninguna tendencia en particular.

La misma restricción que para la regresión lineal simple se aplica a los datos biológicos de volumen o de biomasa, que presentan casi siempre (por no decir siempre) heterocedasticidad. De hecho, la regresión múltiple sólo será generalmente aplicable para el ajuste modelos cuando los datos hayan sido transformados logarítmicamente.

⑨

Regresión polinomial entre $\ln(B)$ e $\ln(D)$

El análisis exploratorio (Línea roja 5) ha demostrado que la relación entre el logaritmo de la biomasa y el logaritmo del diámetro correspondía a una relación lineal. Podemos preguntarnos si dicha relación es realmente lineal o bien si no tiene una forma más compleja. Para ello, se puede hacer una regresión polinomial de grado p, es decir, una regresión múltiple de $\ln(B)$ con respecto a $\ln(D)$, $[\ln(D)]^2$, ..., $[\ln(D)]^p$:

$$\ln(B) = a_0 + a_1 \ln(D) + a_2[\ln(D)]^2 + \ldots + a_p[\ln(D)]^p + \varepsilon$$

con

$$\mathrm{Var}(\varepsilon) = \sigma^2$$

La regresión se ajusta mediante mínimos cuadrados ordinarios. Como la transformación logarítmica estabiliza la varianza residual, las hipótesis de la regresión múltiple se verifican *a priori*. Para un polinomio de grado 2, la regresión polinomial se ajusta mediante la siguiente linea de comando:

```
m2 <- lm(log(Btot)~I(log(dbh))+I(log(dbh)^2),data=dat[dat$Btot>0,])
print(summary(m2))
```

Se obtiene:

```
                Estimate  Std. Error  t value  Pr(>|t|)
 (Intercept)   -8.322190    1.031359   -8.069  9.25e-10  ***
 I(log(dbh))    2.294456    0.633072    3.624  0.000846  ***
 I(log(dbh)^2)  0.009631    0.090954    0.106  0.916225
```

con $R^2 = 0{,}9642$. En cuanto a la regresión polinomial de grado 3:

```
m3 <- lm(log(Btot)~I(log(dbh))+I(log(dbh)^2)+I(log(dbh)^3),data=dat[dat$Btot>0,])
print(summary(m3))
```

da:

```
                Estimate  Std. Error  t value  Pr(>|t|)
 (Intercept)   -5.46413     3.80855   -1.435     0.160
 I(log(dbh))   -0.42448     3.54394   -0.120     0.905
 I(log(dbh)^2)  0.82073     1.04404    0.786     0.437
 I(log(dbh)^3) -0.07693     0.09865   -0.780     0.440
```

con $R^2 = 0{,}9648$. Por último, la regresión polinomial de grado 4:

```
m4 <- lm(log(Btot)~I(log(dbh))+I(log(dbh)^2)+I(log(dbh)^3)+I(log(dbh)^4),data=dat[
dat$Btot>0,])
print(summary(m4))
```

da:

```
                Estimate  Std. Error  t value  Pr(>|t|)
 (Intercept)    -26.7953     15.7399   -1.702    0.0973  .
 I(log(dbh))     26.3990     19.5353    1.351    0.1850
 I(log(dbh)^2)  -11.2782      8.7301   -1.292    0.2046
 I(log(dbh)^3)    2.2543      1.6732    1.347    0.1863
 I(log(dbh)^4)   -0.1628      0.1166   -1.396    0.1714
```

con $R^2 = 0{,}9666$. El agregar términos de grado superior a 1 no aporta nada al modelo. Los coeficientes asociados a dichos términos no son significativamente diferentes de cero. Sin embargo, el R^2 del modelo no deja de aumentar con el grado p del polinomio. Así pues R^2 no es un buen criterio para seleccionar el grado del polinomio. Podemos superponer a la nube de puntos biomasa–diámetro las curvas predichas por estos diferentes polinomios: el objeto `m` que designa la regresión lineal de $\ln(B)$ con respecto a $\ln(D)$ ajustada en la Línea roja 7,

```
with(dat,plot(dbh,Btot,xlab="Diámetro (cm)",ylab="Biomasa (t)",log="xy"))
D <- 10^seq(par("usr")[1],par("usr")[2],length=200)
lines(D,exp(predict(m,newdata=data.frame(dbh=D))))
lines(D,exp(predict(m2,newdata=data.frame(dbh=D))))
lines(D,exp(predict(m3,newdata=data.frame(dbh=D))))
lines(D,exp(predict(m4,newdata=data.frame(dbh=D))))
```

Las curvas resultantes aparecen en la Figura 6.6: cuanto más elevado es el grado del polinomio, más se deforma la curva para ajustarse a los datos, con una extrapolación fuera del ámbito de los datos que es cada vez más irrealista (lo que es típico de una sobreparametrización del modelo).

10

Regresión múltiple entre $\ln(B)$, $\ln(D)$ e $\ln(H)$

La exploración gráfica (Líneas rojas 3 e 6) demostró que la variable sintética D^2H estaba vinculada a la biomasa a través de una relación de potencia (o sea, una relación lineal en coordenadas logarítmicas): $B = a(D^2H)^b$. No obstante, podemos preguntarnos si las variables D^2 y H tienen realmente el mismo exponente b, o bien si pueden tener exponentes diferentes: $B = a \times (D^2)^{b_1} H^{b_2}$. Al trabajar con los datos transformados logarítmicamente (lo que, dicho sea de paso, estabiliza la varianza residual), es como si ajustásemos una regresión múltiple de $\ln(B)$ con respecto a $\ln(D)$ y $\ln(H)$:

$$\ln(B) = a + b_1 \ln(D) + b_2 \ln(H) + \varepsilon$$

con

$$\mathrm{Var}(\varepsilon) = \sigma^2$$

La regresión se ajusta mediante mínimos cuadrados ordinarios. El ajuste de esta regresión múltiple:

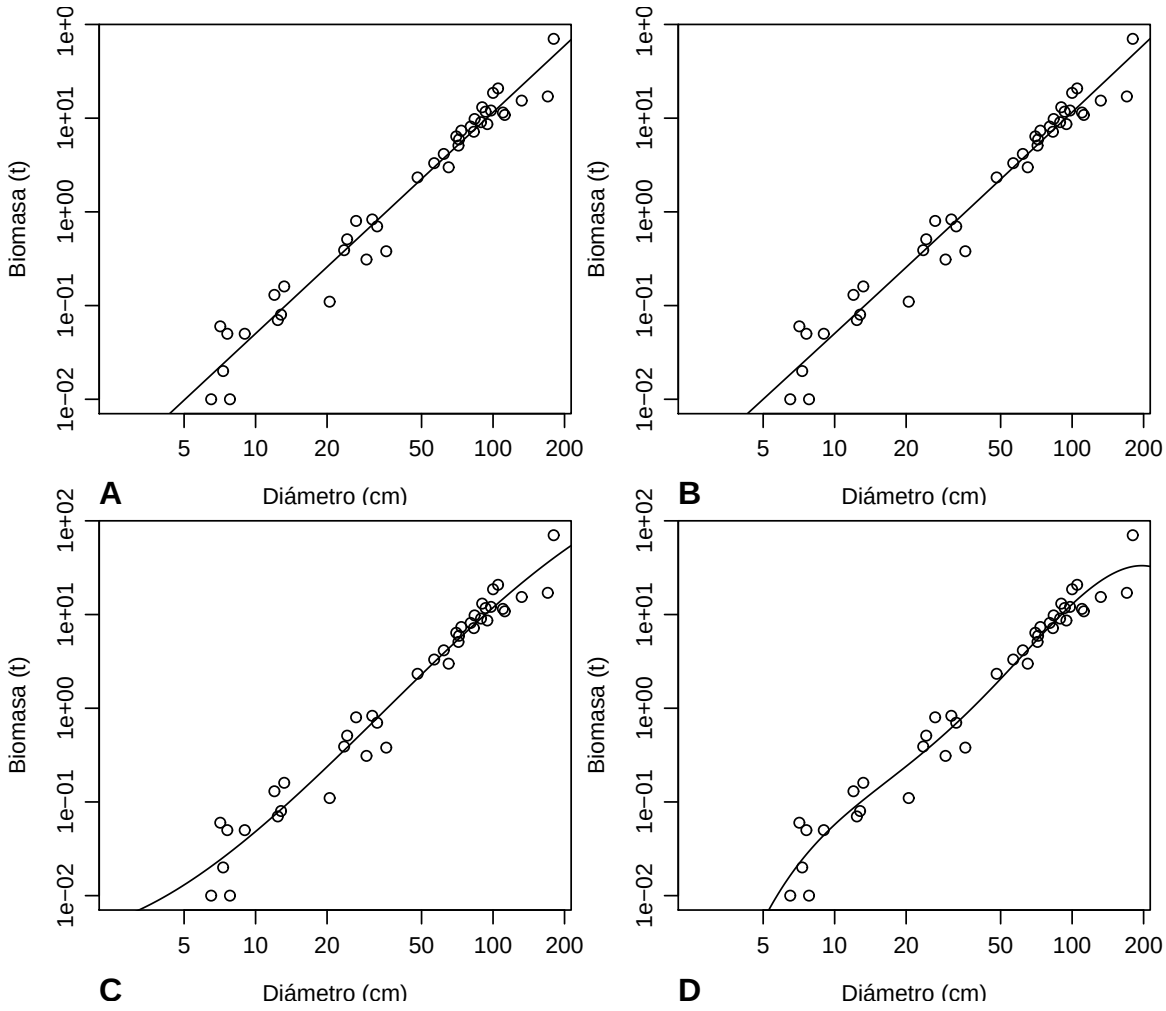


Figura 6.6 – *Biomasa en función del diámetro (en coordenadas logarítmicas) para 42 árboles medido en Ghana por Henry et al. (2010), (puntos), y predicciones (curvas) por medio de una regresión polinomial de* $\ln(B)$ *con respecto a* $\ln(D)$*: (A) polinomio de grado 1 (línea recta); (B) polinomio de grado 2 (parábola); (C) polinomio de grado 3; (D) polinomio de grado 4.*

```
m <- lm(log(Btot)~I(log(dbh))+I(log(haut)),data=dat[dat$Btot>0,])
summary(m)
```

da:

```
                Estimate  Std. Error  t value  Pr(>|t|)
 (Intercept)     -8.9050      0.2855  -31.190    <2e-16  ***
 I(log(dbh))      1.8654      0.1604   11.632  4.35e-14  ***
 I(log(haut))     0.7083      0.2097    3.378   0.00170  **
```

con una desviación estándar residual de 0,4104 y $R^2 = 0{,}9725$. El modelo es altamente significativo (prueba de Fisher: $F_{2,38} = 671{,}5$, p-value $< 2{,}2 \times 10^{-16}$). El modelo, en el que todos los coeficientes son significativamente diferentes de cero, se escribe: $\ln(B) = -8{,}9050 + 1{,}8654 \ln(D) + 0{,}7083 \ln(H)$. Al aplicar la función exponencial para volver a los datos de partida, el modelo se convierte en: $B = 1{,}357 \times 10^{-4} D^{1{,}8654} H^{0{,}7083}$. El exponente asociado a la altura vale un poco menos de la mitad de aquel asociado al diámetro y es un poco menor que el exponente 0,87238 que había sido encontrado para la variable sintética D^2H (cf. Línea roja 8). El examen de los residuos:

```
plot(m,which=1:2)
```

no revela nada en particular (Figura 6.7).

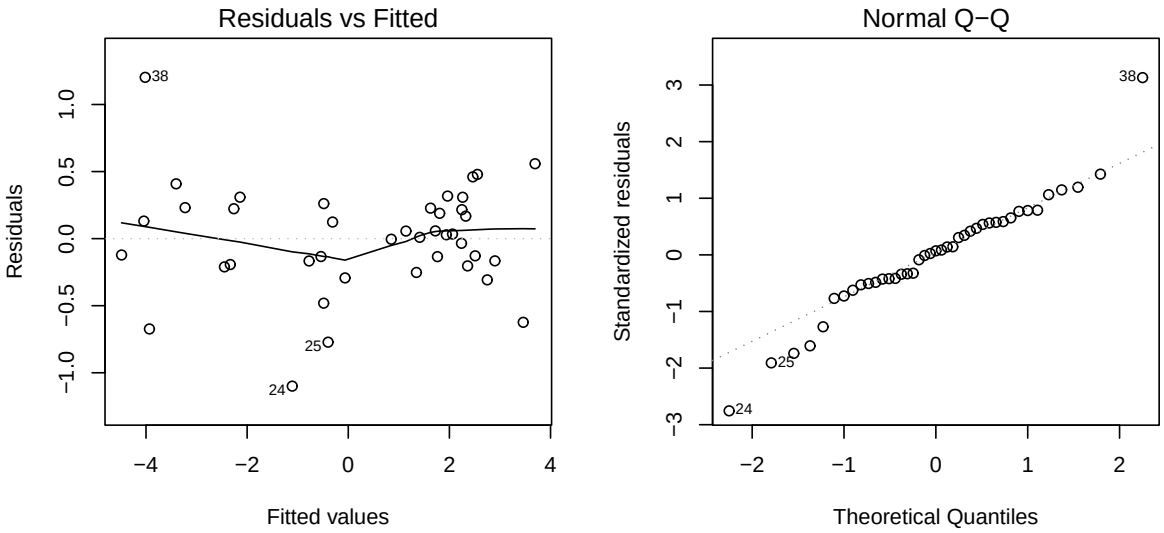


Figura 6.7 – *Gráfico de los residuos en función de los valores predichos (izquierda) y gráfico cuantile–cuantile (derecha) de los residuos de la regresión múltiple de* $\ln(B)$ *con respecto a* $\ln(D)$ *e* $\ln(H)$ *ajustada a los 42 árboles medidos por Henry et al. (2010) en Ghana.*

6.1.3. Regresión ponderada

Supongamos ahora que queremos ajustar directamente un modelo polinomial de la biomasa B con respecto al diámetro D. Por ejemplo, un polinomio de grado 2:

$$B = a_0 + a_1 D + a_2 D^2 + \varepsilon \tag{6.9}$$

Como evocamos anteriormente, la biomasa tiene casi siempre (por no decir siempre) una variabilidad que aumenta con el diámetro D del árbol. En otras palabras, la varianza de

ε aumenta con D, en contradicción con la hipótesis de homocedasticidad necesaria para la regresión múltiple. Por ende, no podríamos ajustar el modelo (6.9) con una regresión múltiple. La transformación logarítmica permite estabilizar la varianza residual (volveremos a ello en la Sección 6.1.5). Al tomar $\ln(B)$ como variable de respuesta, el modelo por ajustar se convierte en:

$$\ln(B) = \ln(a_0 + a_1 D + a_2 D^2) + \varepsilon \tag{6.10}$$

Es razonable suponer que la varianza de los residuos de dicho modelo es constante. Pero, desgraciadamente, ya no se trata de un modelo lineal puesto que la dependencia de la varianza de respuesta con respecto a los coeficientes a_0, a_1 y a_2 no es lineal. Por eso es posible ajustar el modelo (6.10) mediante un modelo lineal. Más adelante (§ 6.2) veremos cómo ajustar este modelo no lineal.

La regresión ponderada permite ajustar un modelo tal como (6.9) en el que la varianza de los residuos no es constante, apoyándose en el formalismo del modelo lineal. Se la puede considerar como una extensión de la regresión múltiple en caso en que la varianza de los residuos no sea constante. La regresión ponderada se desarrolló en ingeniería forestal a partir del decenio de 1960 y hasta el decenio de 1980, en particular gracias a los trabajos de Cunia (1964, 1987a). Fue ampliamente usada para ajustar modelos lineales (Whraton & Cunia, 1987; Brown *et al.*, 1989; Parresol, 1999), antes de ser remplazada por métodos de ajuste más eficaces que veremos en la Sección 6.1.4.

La regresión ponderada se escribe de forma idéntica a la regresión múltiple (6.3):

$$Y = a_0 + a_1 X_1 + a_2 X_2 + \ldots + a_p X_p + \varepsilon$$

con la sola excepción de que ya no suponemos que la varianza de los residuos es constante. Cada observación tiene ahora su propia varianza residual σ_i^2:

$$\varepsilon_i \sim \mathcal{N}(0,\, \sigma_i)$$

A cada observación se asocia un *peso* de ponderación positivo w_i (de allí que se use el adjetivo "ponderada" para calificar esta regresión), que es inversamente proporcional a la varianza residual:

$$w_i \propto 1/\sigma_i^2$$

El coeficiente de proporcionalidad entre w_i y $1/\sigma_i^2$ no es precisado porque el método es insensible en realidad a cualquier nueva normalización del peso (como se verá en la párrafo siguiente). El hecho de asociar a cada observación un peso que es inversamente proporcional a su varianza es bastante natural. Una observación que tiene una fuerte varianza residual se interpreta como una observación que tiene una fuerte variabilidad intrínseca, y es natural pues que tenga menos peso en el ajuste del modelo. Como no se pueden estimar n pesos a partir de n observaciones, hay que modelar la ponderación. Para datos biológicos tales la biomasa o el volumen, la heterocedasticidad de los residuos corresponde casi siempre a una relación de potencia entre la varianza residual y el tamaño de los árboles. Supondremos entonces que, entre las p variables explicativas de la regresión ponderada, hay una (típicamente el diámetro de los árboles) tal que σ_i es una función de potencia de dicha variable. Sin pérdida de generalidad, se puede suponer que esta variable es X_1, de forma que:

$$\sigma_i = k\, X_{i1}^c$$

con $k > 0$ y $c \geq 0$. En consecuencia:

$$w_i \propto X_{i1}^{-2c}$$

El exponente c no puede estimarse del mismo modo que a_0, a_1, ..., a_p, sino que debe determinarse *a priori*. Es el principal inconveniente de este método de ajuste. Veremos más adelante cómo elegir el valor del exponente c. Por el contrario, el coeficiente multiplicador k no hay que estimarlo porque los pesos w_i están definidos sólo dentro de un factor multiplicador. En la práctica, se podría plantear entonces $w_i = X_{i1}^{-2c}$.

Estimación de los coeficientes

El método de los mínimos cuadrados se ajusta para tener en cuenta la ponderación de las observaciones. Se habla entonces del método de los mínimos cuadrados ponderados. Para un exponente c fijo, las estimaciones de los coeficientes a_0, ..., a_p son los valores que minimizan la suma ponderada de los cuadrados de las diferencias:

$$\mathrm{SCE}(a_0,\, a_1,\, \ldots,\, a_p) = \sum_{i=1}^{n} w_i\, \varepsilon_i^2 = \sum_{i=1}^{n} w_i (Y_i - a_0 - a_1 X_{i1} - \ldots - a_p X_{ip})^2$$

o, en escritura matricial:

$$\mathrm{SCE}(\mathbf{a}) = {}^{\mathrm{t}}(\mathbf{Y} - \hat{\mathbf{Y}})\mathbf{W}(\mathbf{Y} - \hat{\mathbf{Y}}) = {}^{\mathrm{t}}(\mathbf{Y} - \mathbf{X}\mathbf{a})\mathbf{W}(\mathbf{Y} - \mathbf{X}\mathbf{a})$$

donde $\mathbf{W}$ es la matriz diagonal $n \times n$ que tiene w_i en su diagonal:

$$\mathbf{W} = \begin{bmatrix} w_1 & & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & w_n \end{bmatrix}$$

El mínimo de SCE se obtiene para (Magnus & Neudecker, 2007):

$$\hat{\mathbf{a}} = \arg\min_{\mathbf{a}} \mathrm{SCE}(\mathbf{a}) = ({}^{\mathrm{t}}\mathbf{X}\mathbf{W}\mathbf{X})^{-1\,\mathrm{t}}\mathbf{X}\mathbf{W}\mathbf{Y}$$

Este mínimo no cambia cuando los pesos w_i son multiplicados todos por el mismo escalar, lo que demuestra sin duda que el método no es sensible a la normalización de los pesos. Podemos cerciorarnos de que la estimación por el método de mínimos cuadrados ponderados aplicados a las observaciones X_{ij} e Y_i dé el mismo resultado que la estimación por el método de los mínimos cuadrados ordinarios aplicados a las observaciones $\sqrt{w_i}\, X_{ij}$ y $\sqrt{w_i}\, Y_i$. Como anteriormente, una ventaja de este método de ajuste es que las estimaciones de los coeficientes tienen una expresión explícita.

Interpretación de los resultados y verificación de las hipótesis

La interpretación de los resultados de la regresión ponderada se hace exactamente del mismo modo que para aquellos de la regresión múltiple. La verificación de las hipótesis relativas a los residuos es similar, con la diferencia que los residuos se remplazan por los residuos ponderados $\varepsilon'_i = \sqrt{w_i}\, \varepsilon_i = \varepsilon_i / X_i^c$. Hay que cerciorarse de que el gráfico de los residuos ponderados ε'_i en función de los valores predichos no presenta ninguna tendencia particular (como en la Figura 6.8B). Si la nube de puntos de los residuos comparada con los valores predichos tiene forma de embudo que se abre hacia la derecha (como en la Figura 6.8A), significa que el valor del exponente c es demasiado pequeño (el valor más pequeño posible es cero). Si la nube de puntos tiene forma de embudo que se cierra hacia la derecha (como en la Figura 6.8C), significa que el valor del exponente c es demasiado grande.

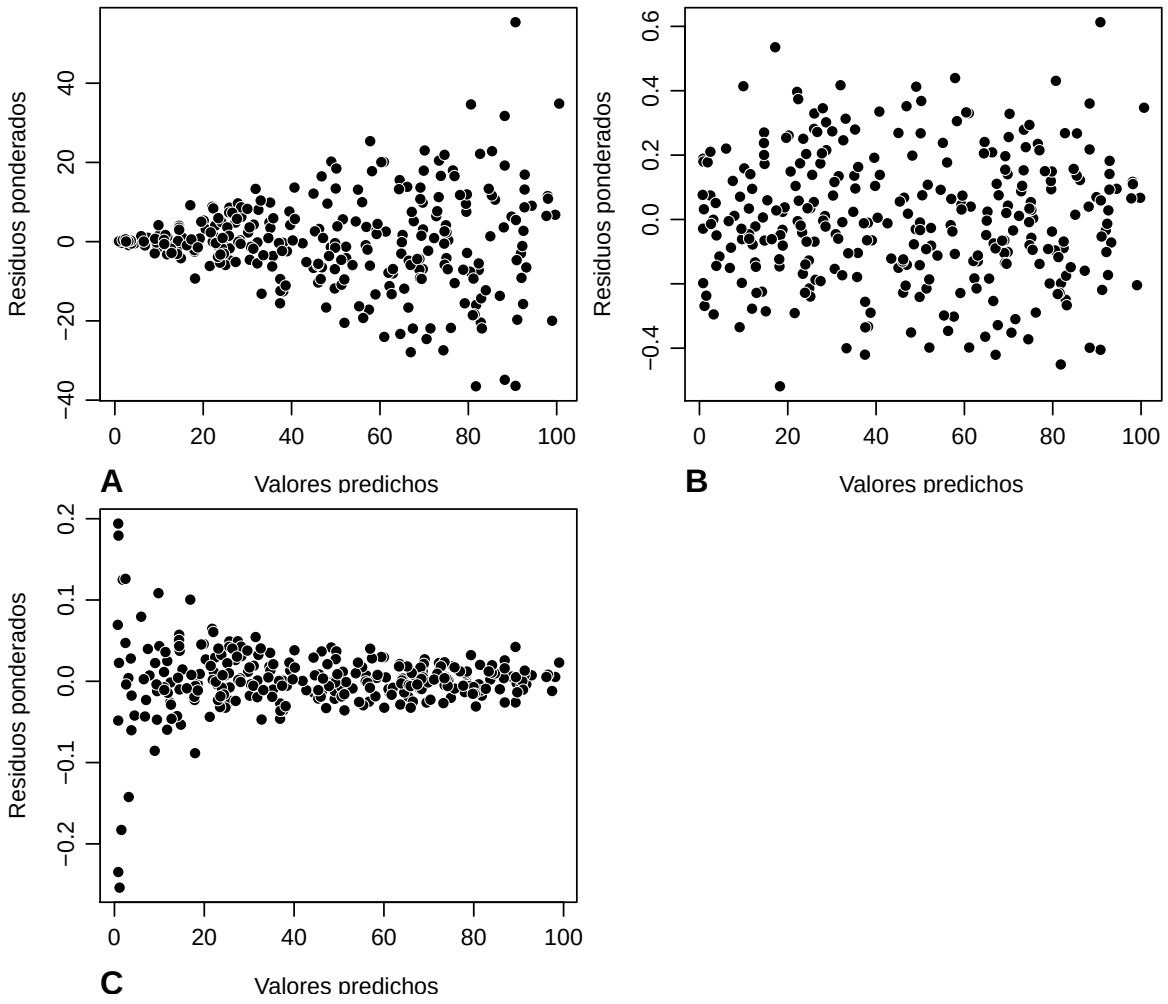


Figura 6.8 – *Gráfico de los residuos ponderados en función de los valores predichos para una regresión ponderada: (A) el valor del exponente c para la ponderación es demasiado pequeño; (B) el valor del exponente c es adecuado; (C) el valor del exponente c es demasiado grande. Cabe señalar que a medida que el exponente c aumenta, disminuye el rango de valores de los residuos ponderados* ε/X^c.

Elección de la ponderación

Un punto crucial de la regresión ponderada es la elección *a priori* del valor del exponente c que define la ponderación. Varios métodos pueden usarse para determinar c. El primer método consiste en proceder por tanteo, en función de la apariencia del gráfico de los residuos ponderados en función de los valores predichos. Dado que la apariencia del gráfico nos indica la pertinencia del valor de c (Figura 6.8), basta con probar varios valores de c hasta que la nube de puntos de los residuos ponderados en función de los valores predichos no presente ya una tendencia particular. Como la regresión lineal es apropiada con respecto a la hipótesis de varianza constante de los residuos, no hace falta determinar c con gran precisión. Comúnmente es suficiente probar valores enteros de c. Concretamente, se podrá ajustar la regresión ponderada para c con valor de 0, 1, 2, 3 o 4 (rara vez resulta útil ir más allá de 4) y retener el valor entero que garantice la mejor apariencia de la nube de puntos de los residuos ponderados en función de los valores predichos. Este método simple suele ser ampliamente suficiente.

Si queremos obtener un valor más preciso del exponente c, se puede proceder a calcular aproximativamente la varianza condicional de la variable de respuesta Y ya que conocemos X_1:

1. subdividir X_1 en K clases centradas en X_{1k} $(k = 1, \ldots, K)$;
2. calcular la varianza empírica, σ_k^2, de Y para las observaciones que pertenezcan a la clase k (con $k = 1, \ldots, K$);
3. hacer una regresión lineal de $\ln(\sigma_k)$ con respecto a $\ln(X_{1k})$.

La pendiente de esta regresión es una estimación de c.

La tercera forma de estimar c consiste en buscar el valor de c que minimice el índice de Furnival (1961). Este índice se define en la página 161.

11

Regresión lineal ponderada entre B y D^2H

El análisis exploratorio de la relación entre la biomasa y D^2H demostró (Línea roja 3) que esta relación era lineal con una varianza de la biomasa que aumentaba con D^2H. Así pues se puede ajustar una regresión ponderada de la biomasa B con respecto a D^2H:

$$B = a + bD^2H + \varepsilon$$

con

$$\mathrm{Var}(\varepsilon) \propto D^{2c}$$

La regresión lineal se ajusta mediante mínimos cuadrados ponderados, lo que exige conocer *a priori* el valor del exponente c.

Estimemos primero el coeficiente c para la ponderación de las observaciones. Para ello, distribuiremos las observaciones en clases de diámetro y calcularemos la desviación estándar de la biomasa en cada clase de diámetro:

```
D <- quantile(dat$dbh,(0:5)/5)
i <- findInterval(dat$dbh,D,rightmost.closed=TRUE)
sdB <- data.frame(D=(D[-1]+D[-6])/2,sdB=tapply(dat$Btot,i,sd))
```

El objeto `D` contiene los límites de las clases de diámetro, calculados de forma tal que tengamos 5 clases que contengan aproximadamente el mismo número de observaciones. El

objeto `i` contiene el número de la clase de diámetro a la que pertenece cada observación. La Figura 6.9, obtenida mediante el comando:

```
with(sdB,plot(D,sdB,log="xy",xlab="Diámetro (cm)",ylab=
"Desviación estándar de la biomasa (t)"))
```

muestra la desviación estándar de la biomasa en función del diámetro mediano de cada clase de diámetro, en una escala logarítmica. Los puntos se alinean aproximadamente a lo largo de una recta, lo que confirma que el modelo de potencia es adecuado para modelizar la varianza residual. La regresión lineal del logaritmo de la desviación estándar de la biomasa con respecto al logaritmo del diámetro mediano de cada clase, ajustado usando al comando:

```
summary(lm(log(sdB)~I(log(D)),data=sdB))
```

da:

```
              Estimate  Std. Error  t value  Pr(>|t|)
 (Intercept)   -7.3487      0.7567   -9.712   0.00232  **
 I(log(D))      2.0042      0.1981   10.117   0.00206  **
```

La pendiente de la regresión es igual a $c = 2$. De esta forma la desviación estándar σ de la biomasa es aproximativamente proporcional a D^2, y se tomará una ponderación de las observaciones inversamente proporcional a D^4.

El ajuste de la regresión ponderada de la biomasa B con respecto a D^2H con esta ponderación, obtenida mediante el comando:

```
m <- lm(Btot~I(dbh^2*haut),data=dat,weights=1/dat$dbh^4)
summary(m)
```

da:

```
                Estimate  Std. Error  t value  Pr(>|t|)
 (Intercept)   1.181e-03   2.288e-03    0.516     0.608
 I(dbh^2*haut) 2.742e-05   1.527e-06   17.957    <2e-16  ***
```

Un examen del resultado de este ajuste muestra que la intersección no es significativamente diferente de cero. Así pues tenemos que ajustar una nueva regresión ponderada de la biomasa B con respecto a D^2H sin intersección:

```
m <- lm(Btot~-1+I(dbh^2*haut),data=dat,weights=1/dat$dbh^4)
summary(m)
```

lo que da:

```
                Estimate  Std. Error  t value  Pr(>|t|)
 I(dbh^2*haut) 2.747e-05   1.511e-06    18.19    <2e-16  ***
```

El modelo se escribe entonces: $B = 2{,}747 \times 10^{-5} D^2 H$, con un R^2 de 0,8897 y una desviación estándar residual de $k = 0{,}0003513$ toneladas cm^{-2}. El modelo es altamente significativo (prueba de Fisher: $F_{1,41} = 330{,}8$, p-value $< 2{,}2 \times 10^{-16}$). Como este modelo fue ajustado directamente sobre los datos no transformados, cabe señalar que no hace falta retirar las observaciones con una biomasa nula (contrariamente a la Línea roja 8). La Figura 6.10A, obtenida con el comando:

```
plot(fitted(m),residuals(m)/dat$dbh^2,xlab="Valores predichos",ylab=
"Residuos ponderados")
```

muestra los residuos ponderados en función de los valores predichos. En comparación, la Figura 6.10B muestra los residuos ponderados en función de los valores predichos cuando la ponderación es demasiado pequeña (con pesos inversamente proporcionales a D^2):

```
m <- lm(Btot~-1+I(dbh^2*haut),data=dat,weights=1/dat$dbh^2)
plot(fitted(m),residuals(m)/dat$dbh,xlab="Valores predichos",ylab="Residuos ponderados")
```

mientras que la Figura 6.10C muestra los residuos ponderados en función de los valores predichos, si la ponderación hubiera sido demasiado grande (con pesos inversamente proporcionales a D^5):

```
m <- lm(Btot~-1+I(dbh^2*haut),data=dat,weights=1/dat$dbh^5)
plot(fitted(m),residuals(m)/dat$dbh^2.5,xlab="Valores predichos",ylab=
"Residuos ponderados")
```

Por tanto, el coeficiente $c = 2$ la ponderación se revela claramente como el adecuado.

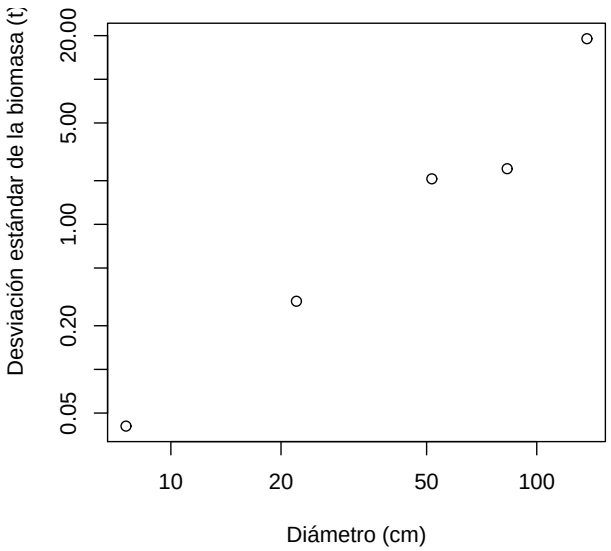


Figura 6.9 – *Desviación estándar de la biomasa calculada en cinco clases de diámetro en función del diámetro mediano de la clase (en coordenadas logarítmicas) para 42 árboles medidos en Ghana por Henry et al. (2010).*

12

Regresión polinomial ponderada entre B y D

El análisis exploratorio (Línea roja 2) demostró que la relación entre la biomasa y el diámetro es parabólica, con un aumento de la varianza de la biomasa con el diámetro. La transformación logarítmica permite linealizar la relación entre la biomasa y el diámetro pero también se puede buscar modelar directamente la relación entre la biomasa y el diámetro mediante una función parabólica:

$$B = a_0 + a_1 D + a_2 D^2 + \varepsilon$$

con

$$\mathrm{Var}(\varepsilon) \propto D^{2c}$$

En la Línea roja 11, vimos que el valor $c = 2$ del exponente convenía para modelar la desviación estándar condicional de la biomasa conociendo el diámetro. Entonces, ajustamos

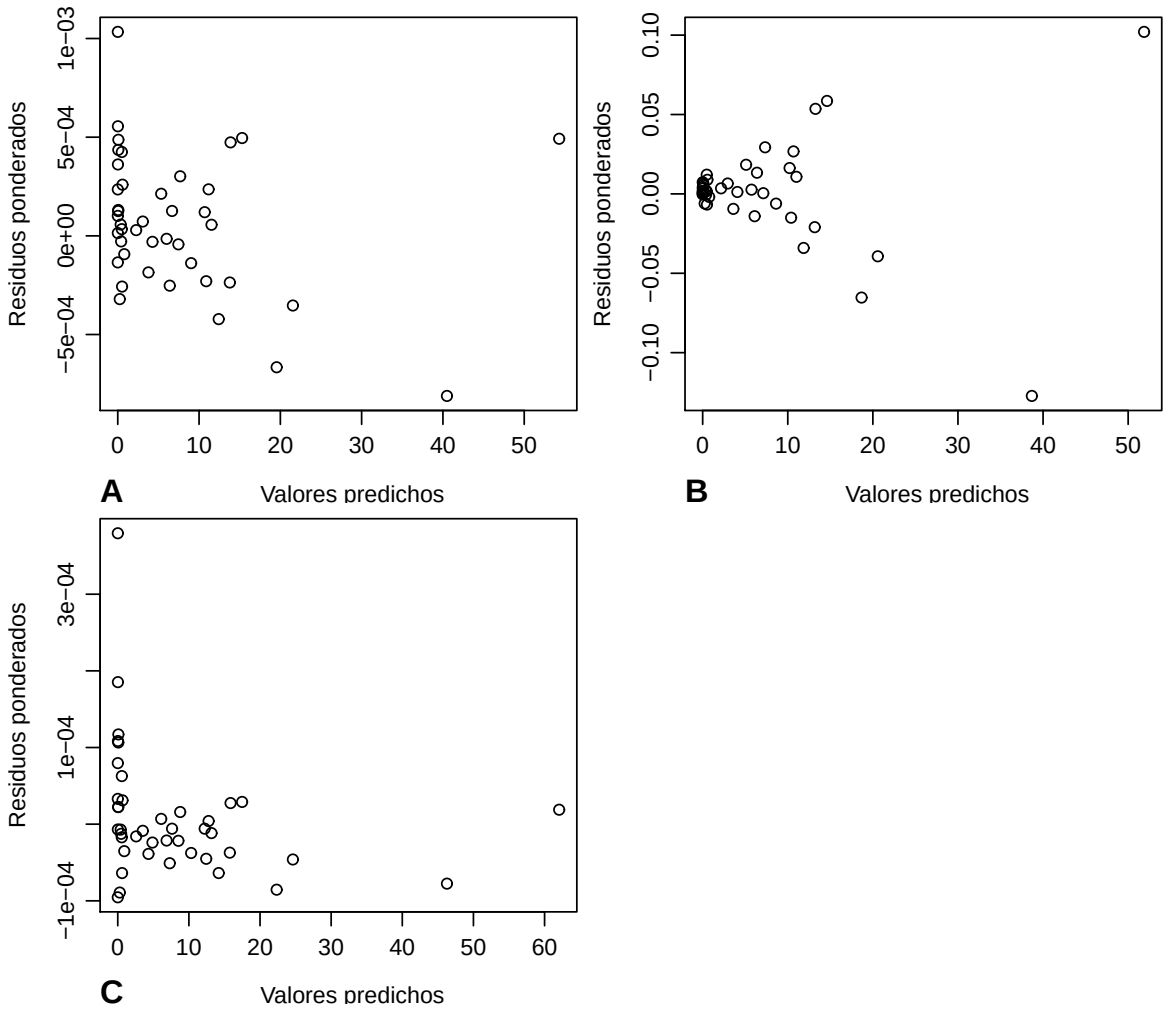


Figura 6.10 – *Gráfico de los residuos ponderados en función de los valores predichos para la regresión ponderada de la biomasa con respecto a D^2H para 42 árboles medidos en Ghana por Henry et al. (2010): (A) la ponderación es inversamente proporcional a D^4; (B) la ponderación es inversamente proporcional a D^2; (C) la ponderación es inversamente proporcional a D^5.*

la regresión múltiple mediante los mínimos cuadrados ponderados con una ponderación de las observaciones proporcional a $1/D^4$:

```
m <- lm(Btot~dbh+I(dbh^2),data=dat,weights=1/dat$dbh^4)
summary(m)
```

lo que da:

```
               Estimate  Std. Error  t value  Pr(>|t|)
(Intercept)   1.127e-02   6.356e-03    1.772   0.08415  .
dbh          -7.297e-03   2.140e-03   -3.409   0.00153  **
I(dbh^2)      1.215e-03   9.014e-05   13.478  2.93e-16  ***
```

con una desviación estándar residual $k = 0{,}0003882$ toneladas cm^{-2} y $R^2 = 0{,}8709$. La intersección resulta no ser significativamente diferente de cero. En consecuencia, ajustaremos nuevamente una función parabólica pero sin intersección:

```
m <- lm(Btot~-1+dbh+I(dbh^2),data=dat,weights=1/dat$dbh^4)
summary(m)
```

lo que da:

```
             Estimate  Std. Error  t value  Pr(>|t|)
dbh        -3.840e-03   9.047e-04   -4.245  0.000126  ***
I(dbh^2)    1.124e-03   7.599e-05   14.789    <2e-16  ***
```

con una desviación estándar residual $k = 0{,}0003985$ toneladas cm^{-2} y $R^2 = 0{,}8615$. El modelo es altamente significativo (prueba de Fisher: $F_{2,40} = 124{,}4$, p-value $= 2{,}2 \times 10^{-16}$) y se escribe: $B = -3{,}840 \times 10^{-3} D + 1{,}124 \times 10^{-3} D^2$. El gráfico 6.11 obtenido mediante el comando:

```
plot(fitted(m),residuals(m)/dat$dbh^2,xlab="Valores predichos",ylab=
"Residuos ponderados")
```

muestra los residuos ponderados en función de los valores predichos.

6.1.4. Regresión lineal con modelo de varianza

Una alternativa a la regresión ponderada consiste en plantear explícitamente un modelo para la varianza de los residuos. Igual que antes, es realista plantear que existe una variable explicativa (sin pérdida de generalidad, la primera) tal que la desviación estándar residual es una función de potencia de esta variable:

$$\mathrm{Var}(\varepsilon) = (kX_1^c)^2 \tag{6.11}$$

con $k > 0$ y $c \geq 0$. . El modelo se escribe entonces:

$$Y = a_0 + a_1 X_1 + a_2 X_2 + \ldots + a_p X_p + \varepsilon \tag{6.12}$$

con:

$$\varepsilon \sim \mathcal{N}(0,\, kX_1^c)$$

En cuanto a la forma, el modelo no se diferencia de la regresión ponderada. En cuanto al fondo, hay una diferencia fundamental: los coeficientes k y c son ahora parámetros del modelo por estimar, del mismo modo que los coeficientes a_0, a_1, …, a_p. Debido a estos parámetros k y c por estimar, el método de los mínimos cuadrados no puede usarse para

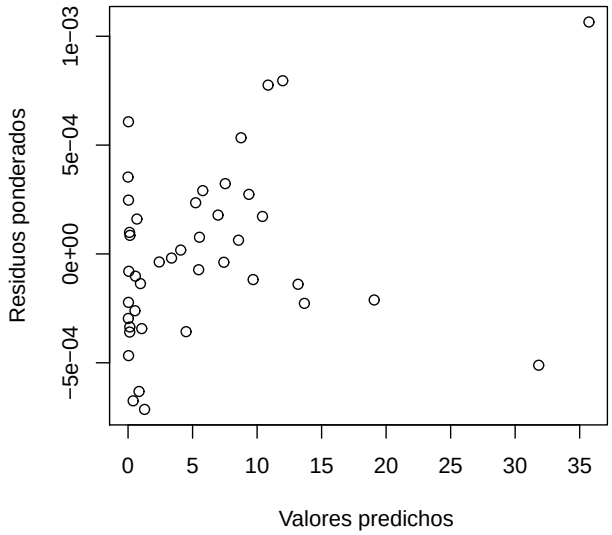


Figura 6.11 – *Gráfico de los residuos ponderados en función de los valores predichos para la regresión ponderada de la biomasa con respecto a D y D^2 para 42 árboles medidos en Ghana por Henry et al. (2010).*

estimar los coeficientes del modelo. Hay que usar otro método de estimación, a saber, el método de máxima verosimilitud. *En sentido estricto*, el modelo definido por (6.11) y (6.12) no corresponde al modelo lineal. Conceptualmente está mucho más próximo del modelo no lineal que veremos en la Sección 6.2. No entraremos en más detalles aquí sobre el modelo no lineal: el método de ajuste del modelo definido por (6.11) y (6.12) se presentará como un caso particular del modelo no lineal en la Sección 6.2.

13

Regresión lineal entre B y D^2H con modelo de varianza

Anticipándonos a la Sección 6.2, vamos a ajustar una regresión lineal de la biomasa con respecto a D^2H en especificando un modelo de potencia sobre la varianza residual:

$$B = a + bD^2H + \varepsilon$$

con

$$\mathrm{Var}(\varepsilon) = (kD^c)^2$$

Más adelante (§ 6.2) veremos que este modelo está ajustado mediante la máxima verosimilitud. Esta regresión es muy parecida en esencia a la regresión ponderada de la biomasa con respecto a D^2H efectuada anteriormente (Línea roja 11), con la única diferencia de que el exponente c usado para definir la ponderación de las observaciones ahora es un parámetro por estimar de pleno derecho y no un coeficiente dado *a priori*. La regresión lineal con modelo de varianza se ajusta del modo siguiente:

```
library(nlme)
start <- coef(lm(Btot~I(dbh^2*haut),data=dat))
names(start) <- c("a","b")
summary(nlme(Btot~a+b*dbh^2*haut, data=cbind(dat,g="a"), fixed=a+b~1, start=start,
groups=~g, weights=varPower(form=~dbh)))
```

y da (en la Sección 6.2 volveremos sobre el significado del objeto `start`):

```
            Value      Std.Error  DF    t-value  p-value
a   0.0012868020  0.0024211610  40   0.531481    0.598
b   0.0000273503  0.0000014999  40  18.234340    0.000
```

con un valor estimado del exponente $c = 1{,}977736$. Al igual que en la regresión no lineal ponderada (Línea roja 11), la intersección no es significativamente diferente de cero. Así pues, se reajusta el modelo sin intersección:

```
summary(nlme(Btot~b*dbh^2*haut, data=cbind(dat,g="a"), fixed=b~1, start=start["b"],
groups=~g, weights=varPower(form=~dbh)))
```

lo que da:

```
            Value    Std.Error  DF   t-value  p-value
b   2.740688e-05  1.4869e-06  41  18.43223        0
```

con un valor estimado del exponente $c = 1{,}980263$. Dicho valor es muy similar al evaluado para la regresión lineal ponderada ($c = 2$ en la Línea roja 11). El modelo ajustado se escribe entonces: $B = 2{,}740688 \times 10^{-5} D^2 H$, lo que es muy próximo al modelo ajustado por regresión lineal ponderada (Línea roja 11).

14

Regresión polinomial entre B y D con modelo de varianza

Anticipándonos a la Sección 6.2, vamos a ajustar una regresión múltiple de la biomasa con respecto a D y D^2 en especificando un modelo de potencia sobre la varianza residual:

$$B = a_0 + a_1 D + a_2 D^2 + \varepsilon$$

con

$$\mathrm{Var}(\varepsilon) = (kD^c)^2$$

Más adelante (§ 6.2) veremos que este modelo está ajustado por máxima verosimilitud. Esta regresión es muy parecida en esencia a la regresión polinomial de la biomasa con respecto a D y D^2 realizada antes (Línea roja 12), con la única diferencia de que el exponente c usado para definir la ponderación de las observaciones ahora es un parámetro por estimar de pleno derecho y ya no un coeficiente dado *a priori*. La regresión lineal con modelo de varianza se ajusta del modo siguiente:

```
library(nlme)
start <- coef(lm(Btot~dbh+I(dbh^2),data=dat))
names(start) <- c("a0","a1","a2")
summary(nlme(Btot~a0+a1*dbh+a2*dbh^2,data=cbind(dat,g="a"),fixed=a0+a1+a2~1,
start=start,groups=~g,weights=varPower(form=~dbh)))
```

y da (en la Sección 6.2 volveremos sobre el significado del objeto `start`):

```
             Value     Std.Error  DF    t-value  p-value
a0    0.009048498  0.005139129  39   1.760706   0.0861
a1   -0.006427411  0.001872346  39  -3.432812   0.0014
a2    0.001174388  0.000094063  39  12.485081   0.0000
```

con un valor estimado del exponente $c = 2{,}127509$. Como en la regresión polinomial ponderada (Línea roja 12), la intersección no es significativamente diferente de cero. Se reajusta entonces el modelo sin intersección:

```
summary(nlme(Btot~a1*dbh+a2*dbh^2,data=cbind(dat,g="a"),fixed=a1+a2~1,start=start[
c("a1","a2")],groups=~g,weights=varPower(form=~dbh)))
```

lo que da:

```
            Value      Std.Error  DF    t-value  p-value
a1  -0.003319456  0.0006891736  40  -4.816574        0
a2   0.001067068  0.0000759745  40  14.045082        0
```

con un valor estimado del exponente $c = 2{,}139967$. Este valor es muy similar al evaluado para la regresión polinomial ponderada ($c = 2$ en la Línea roja 12). El modelo ajustado se escribe entonces: $B = -3{,}319456 \times 10^{-3} D + 1{,}067068 \times 10^{-3} D^2$, lo que es muy próximo del modelo ajustado por regresión polinomial ponderada (Línea roja 12).

6.1.5. Transformación de variable

Retomemos el ejemplo de un modelos de biomasa de una entrada (en este caso, el diámetro) de tipo potencia:

$$B = aD^b \tag{6.13}$$

Ya vimos que se trata de un modelo no lineal dado que B depende en forma no lineal de los coeficientes a y b. Por el contrario, se puede linealizar este modelo aplicando la transformación logarítmica. La relación (6.13) es equivalente a: $\ln(B) = \ln(a) + b\ln(D)$, que se puede ver como una regresión lineal de la variable de respuesta $Y = \ln(B)$ con respecto a la variable explicativa $X = \ln(D)$. Se pueden entonces estimar los coeficientes a y b (o más bien $\ln(a)$ y b) del modelo de potencia (6.13) mediante regresión lineal sobre los datos transformados logarítmicamente. ¿Qué ocurre con el error residual? Si la regresión lineal en los datos transformados logarítmicamente es pertinente, esto significa que $\varepsilon = \ln(B) - \ln(a) - b\ln(D)$ corresponde a una distribución normal centrada y de desviación estándar constante σ. Si volvemos a los datos de partida utilizando la transformación exponencial (que es la transformación inversa de la transformación logarítmica), el error residual es el factor:

$$B = aD^b \times \varepsilon'$$

con $\varepsilon' = \exp(\varepsilon)$. Así pues, pasamos de un error aditivo en los datos transformados logarítmicamente a un error multiplicativo en los datos de partida. Además, si ε corresponde a una distribución normal centrada con una desviación estándar σ, entonces, por definición, $\varepsilon' = \exp(\varepsilon)$ corresponde a una distribución lognormal de parámetros 0 y σ:

$$\varepsilon' \underset{\text{i.i.d.}}{\sim} \mathcal{LN}(0, \sigma)$$

En contraste con ε cuya media es cero, la media de ε' no lo es sino que vale: $\mathrm{E}(\varepsilon') = \exp(\sigma^2/2)$. En el Capítulo 7 veremos las consecuencias de ello.

Hay dos cosas que debemos retener de este ejemplo:

1. cuando nos enfrentamos a una relación no lineal entre una variable de respuesta y una o varias variables explicativas, una transformación de variable puede permitir volver lineal esta relación;

2. la transformación de variable afecta no sólo la forma de la relación entre la o las variables explicativas y la variable de respuesta, sino también el error residual.

Con respecto al primer punto, la transformación de variables lleva a diferenciar dos enfoques para ajustar un modelo no lineal. Ante una relación no lineal entre una variable de respuesta y variables explicativas, el primer enfoque consiste en buscar una transformación que linealice esta relación, para acercarse al caso del modelo lineal. El segundo enfoque consiste en ajustar directamente el modelo no lineal, como lo veremos en la Sección 6.2. Cada enfoque tiene sus ventajas e inconvenientes. El modelo lineal presenta la ventaja de aportar un marco teórico relativamente simple y, sobre todo, las estimaciones de sus coeficientes son expresiones explícitas. El inconveniente es que la etapa de linealización del modelo introduce una dificultad adicional y que la transformación inversa, si no tenemos cuidado, puede introducir un sesgo de predicción (al que volveremos en el Capítulo 7). Además, no todos los modelos son linealizables. Por ejemplo, no existe ninguna transformación de variable que permita linealizar el modelo siguiente: $Y = a_0 + a_1 X + a_2 \exp(a_3 X)$.

Con respecto al segundo punto, a partir de ahora tendremos que distinguir la forma de la relación entre la variable de respuesta y las variables explicativas (se habla también de modelo para la media – sobreentendiéndose la media de la variable de respuesta Y), y la forma del modelo para el error residual (se habla también de modelo para la varianza – sobreentendiéndose la varianza de Y). La transformación de variable afecta a ambas simultáneamente. Todo el arte de la transformación de variable consiste en actuar en estos dos planos simultáneamente para hacer que el modelo se vuelva lineal con respecto a sus coeficientes y estabilizar la varianza de los residuos (es decir, volverla constante).

Transformaciones usuales de las variables

Aunque no haya límite teórico a las transformaciones de variable que se pueden usar, las transformaciones que pueden afectar los volúmenes o las biomasas son pocas. La transformación que resultará más usada para el ajuste de los modelos es la transformación logarítmica. Dado un modelo de potencia:

$$Y = a X_1^{b_1} X_2^{b_2} \times \ldots \times X_p^{b_p} \times \varepsilon$$

la transformación logarítmica consiste en remplazar la variable Y por su logaritmo: $Y' = \ln(Y)$, y cada una de las variables de respuesta por su logaritmo: $X'_j = \ln(X_j)$. El modelo resultante es:

$$Y' = a' + b_1 X'_1 + b_2 X'_2 + \ldots + b_p X'_p + \varepsilon' \tag{6.14}$$

con $\varepsilon' = \ln(\varepsilon)$. La transformación inversa es exponencial para el conjunto de variables (de respuesta y explicativas). En términos de error residual, la transformación logarítmica es adecuada si ε' tiene una distribución normal, o sea si el error ε es positivo y actúa en forma multiplicativa. Cabe señalar que para variables que puedan tener un valor cero, la transformación logarítmica plantea problemas. En ese caso, se usa la transformación $X' = \ln(X + 1)$ en vez de $X' = \ln(X)$ (o en forma más general $X' = \ln(X + \text{constante})$ si X puede tener valores negativos, como un crecimiento diamétrico, por ejemplo). Como ejemplo, los modelos de biomasa siguientes:

$$\begin{aligned} B &= a D^b \\ B &= a (D^2 H)^b \\ B &= a \rho^{b_1} D^{b_2} H^{b_3} \end{aligned}$$

pueden ajustarse mediante una regresión lineal luego de una transformación logarítmica de los datos.

Dado un modelo exponencial:

$$Y = a\exp(b_1X_1 + b_2X_2 + \ldots + b_pX_p) \times \varepsilon \tag{6.15}$$

la transformación adecuada consiste en remplazar la variable Y por su logaritmo: $Y' = \ln(Y)$, y en no transformar las variables de respuesta: $X'_j = X_j$. El modelo resultante es idéntico a (6.14). La transformación inversa es exponencial para la variable de respuesta y no hay cambios para las variables explicativas. En términos de error residual, esta transformación es adecuada si ε' tiene un distribución normal, o sea si el error ε es positivo y actúa de forma multiplicativa. Cabe señalar que, sin pérdida de generalidad, se pueden volver a parametrar los coeficientes del modelo exponencial (6.15) planteando $b'_j = \exp(b_j)$. Una forma de escribir estrictamente equivalente del modelo exponencial (6.15) es pues:

$$Y = ab_1'^{X_1}b_2'^{X_2} \times \ldots \times b_p'^{X_p} \times \varepsilon$$

Como ejemplo, el modelo de biomasa siguiente:

$$B = \exp\{a_0 + a_1\ln(D) + a_2[\ln(D)]^2 + a_3[\ln(D)]^3\}$$

puede ajustarse mediante regresión lineal luego de una transformación de variable de este tipo (con, en ese ejemplo $X_j = [\ln(D)]^j$).

La transformación de Box-Cox generaliza la transformación logarítmica. Es en realidad una familia de transformaciones indexada por un parámetro ξ. Dada una variable X, la transformación de Box-Cox X'_ξ es:

$$X'_\xi = \begin{cases} (X^\xi - 1)/\xi & (\xi \neq 0) \\ \ln(X) = \lim_{\xi\to 0}(X^\xi - 1)/\xi & (\xi = 0) \end{cases}$$

La transformación de Box-Cox permite convertir el dilema de la elección de una transformación de variable en uno de estimación de un parámetro ξ (Hoeting *et al.*, 1999).

Transformación de una variable particular

Las transformaciones de variable usuales cambian la forma de la relación entre la variable de respuesta y la variable explicativa. Cuando la nube de puntos (X_i, Y_i) de la variable de respuesta en función de la variable explicativa tiene la forma de una recta con heterocedasticidad, tal como se esquematiza en la Figura 6.12, es necesario aplicar una transformación de variable para estabilizar la varianza de Y, sin afectar no obstante el carácter lineal de la relación entre X y Y. El ejemplo presentado en 6.12 se da con bastante frecuencia cuando se ajusta una ecuación alométrica entre dos magnitudes que varían proporcionalmente (cf. por ejemplo Ngomanda *et al.*, 2012). El carácter lineal de la relación entre X e Y significa que el modelo es de forma:

$$Y = a + bX + \varepsilon \tag{6.16}$$

pero la heterocedasticidad significa que la varianza de ε no es constante, lo que impide ajustar una regresión lineal. Una transformación de variable en este caso consiste en remplazar Y por $Y' = Y/X$ y X por $X' = 1/X$. Dividiendo cada miembro de (6.16) por X, el modelo después de la transformación de la variable se convierte en:

$$Y' = aX' + b + \varepsilon' \tag{6.17}$$

con $\varepsilon' = \varepsilon/X$. El modelo transformado corresponde siempre a una relación lineal, con la salvedad de que la intersección a de la relación entre X e Y se convierte en la pendiente de la relación entre X' e Y', y recíprocamente la pendiente b de la relación entre X e Y se convierte en intersección de la relación entre X' e Y'. El modelo (6.17) podrá ajustarse por una regresión lineal simple si la varianza de ε' es constante. Como $\text{Var}(\varepsilon') = \sigma^2$ implica $\text{Var}(\varepsilon) = \sigma^2 X^2$, esto implica que la transformación de la variable es adecuada si la desviación estándar de ε es proporcional a X.

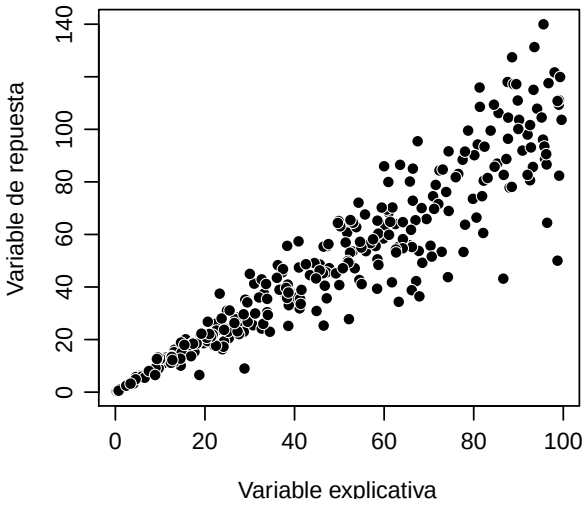


Figura 6.12 – *Relación lineal entre una variable explicativa (X) y una variable de respuesta (Y), con crecimiento de la variabilidad de Y cuando aumenta X (heterocedasticidad).*

En el modelo (6.17) que está ajustado por regresión lineal simple, la suma de los cuadrados de sus desviaciones vale:

$$\text{SCE}(a,\, b) = \sum_{i=1}^{n}(Y_i' - aX_i' - b)^2 = \sum_{i=1}^{n}(Y_i/X_i - a/X_i - b)^2 = \sum_{i=1}^{n} X_i^{-2}(Y_i - a - bX_i)^2$$

En esta última expresión, se reconoce la expresión de la suma de los cuadrados de las desviaciones para una regresión ponderada que utiliza pesos $w_i = X_i^{-2}$. De esta forma, la transformación de la variable $Y' = Y/X$ y $X' = 1/X$ es estrictamente idéntica a una regresión ponderada de peso $w = 1/X^2$.

15

Regresión lineal entre B/D^2 y H

En la Línea roja 11 vimos que un modelo de biomasa con dos entradas con respecto al diámetro y la altura era: $B = a + bD^2H + \varepsilon$ con $\text{Var}(\varepsilon) \propto D^4$. Al dividir cada miembro de la ecuación por D^2, obtenemos:

$$B/D^2 = a/D^2 + bH + \varepsilon'$$

con

$$\text{Var}(\varepsilon') = \sigma^2$$

De esta forma, la regresión de la variable de respuesta $Y = B/D^2$ con respecto a las dos variables explicativas $X_1 = 1/D^2$ y $X_2 = H$ verifica *a priori* priori las hipótesis de la regresión lineal múltiple. Esta regresión se ajusta mediante los mínimos cuadrados ordinarios. El ajuste de dicha regresión múltiple se logra mediante el comando:

```
summary(lm((Btot/dbh^2)~-1+I(1/dbh^2)+haut,data=dat))
```

da:

```
              Estimate  Std. Error  t value  Pr(>|t|)
I(1/dbh^2)  1.181e-03   2.288e-03    0.516     0.608
haut        2.742e-05   1.527e-06   17.957    <2e-16  ***
```

donde se pone de manifiesto que el coeficiente asociado a $X_1 = 1/D^2$ no es significativamente diferente de cero. Si volvemos a los datos de partida, eso significa simplemente que la intersección a no es significativamente diferente de cero, lo que ya habíamos diagnosticado en la Línea roja 11. Podemos retirar entonces X_1 y ajustar una regresión lineal simple de $Y = B/D^2$ con respecto a $X_2 = H$:

```
with(dat,plot(haut,Btot/dbh^2,xlab="Altura (m)",ylab="Biomasa/cuadrado del diámetro
(t/cm2)"))
m <- lm((Btot/dbh^2)~-1+haut,data=dat)
summary(m)
plot(m,which=1:2)
```

La nube de puntos de B/D^2 en función de H tiene efectivamente la forma de una recta con una varianza de B/D^2 que es aproximativamente constante (Figura 6.13). El ajuste de la regresión lineal simple da:

```
       Estimate  Std. Error  t value  Pr(>|t|)
haut  2.747e-05   1.511e-06    18.19    <2e-16  ***
```

con un R^2 de 0,8897 y una desviación estándar residual de 0,0003513 toneladas cm^{-2}. El modelo se escribe: $B/D^2 = 2{,}747 \times 10^{-5} H$, o sea, volviendo a las variables de partida: $B = 2{,}747 \times 10^{-5} D^2 H$. Hace falta verificar que este modelo es estrictamente idéntico a la regresión ponderada de B con respecto a D^2H realizada en la Línea roja 11 con una ponderación proporcional a $1/D^4$. El gráfico de los residuos en función de los valores predichos y el gráfico cuantil-cuantil de los residuos se muestran en la Figura 6.14.

16

Regresión lineal entre B/D^2 y $1/D$

En la Línea roja 12 vimos que un modelo polinomial de biomasa, con respecto al diámetro, era: $B = a_0 + a_1 D + a_2 D^2 + \varepsilon$ con $\mathrm{Var}(\varepsilon) \propto D^4$. Al dividir cada miembro de la ecuación por D^2, se obtiene:

$$B/D^2 = a_0/D^2 + a_1/D + a_2 + \varepsilon'$$

con

$$\mathrm{Var}(\varepsilon') = \sigma^2$$

Por tanto, la regresión de la variable de respuesta $Y = B/D^2$ con respecto a las dos variables explicativas $X_1 = 1/D^2$ y $X_2 = 1/D$ verifican *a priori* las hipótesis de la regresión lineal múltiple. Esta regresión se ajusta mediante los mínimos cuadrados ordinarios. El ajuste de esta regresión múltiple con el comando:

```
summary(lm((Btot/dbh^2)~I(1/dbh^2)+I(1/dbh),data=dat))
```

da:

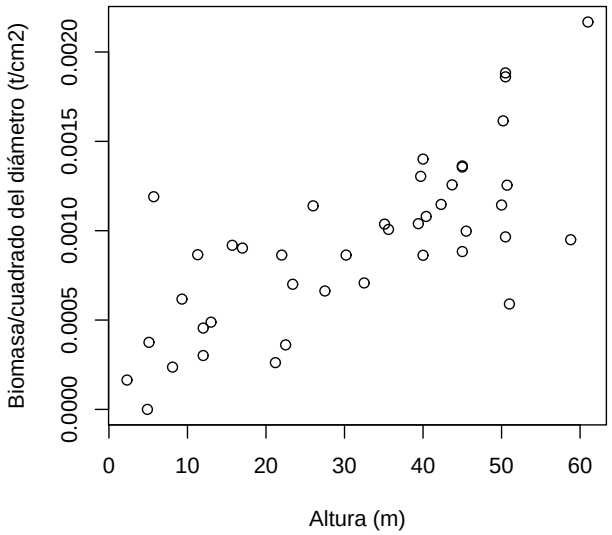


Figura 6.13 – *Nube de puntos de la biomasa dividida por el cuadrado del diámetro (toneladas cm^{-2}) en función de la altura (m) para 42 árboles medidos en Ghana por Henry et al. (2010).*

```
              Estimate  Std. Error  t value  Pr(>|t|)
(Intercept)   1.215e-03   9.014e-05   13.478  2.93e-16  ***
I(1/dbh^2)    1.127e-02   6.356e-03    1.772   0.08415  .
I(1/dbh)     -7.297e-03   2.140e-03   -3.409   0.00153  **
```

donde se pone de manifiesto que el coeficiente asociado a $X_1 = 1/D^2$ no es significativamente diferente de cero. Si volvemos a los datos de partida, eso significa simplemente que la intersección a_0 no es significativamente diferente de cero, lo que habíamos diagnosticado en la Línea roja 12. Por tanto, podemos retirar X_1 y ajustar una regresión lineal simple de $Y = B/D^2$ con respecto a $X_2 = 1/D$:

```
with(dat,plot(1/dbh,Btot/dbh^2,xlab="1/diámetro (/cm)",ylab="Biomasa/cuadrado del
diámetro (t/cm2)"))
m <- lm((Btot/dbh^2)~I(1/dbh),data=dat)
summary(m)
plot(m,which=1:2)
```

La nube de puntos de B/D^2 en función de $1/D$ tiene aproximativamente la forma de una recta con una varianza de B/D^2 que es aproximativamente constante (Figura 6.15). El ajuste de la regresión lineal simple da:

```
              Estimate  Std. Error  t value  Pr(>|t|)
(Intercept)   1.124e-03   7.599e-05   14.789    <2e-16  ***
I(1/dbh)     -3.840e-03   9.047e-04   -4.245  0.000126  ***
```

con un R^2 de 0,3106 y una desviación estándar residual de 0,0003985 toneladas cm^{-2}. El modelo se escribe: $B/D^2 = 1{,}124 \times 10^{-3} - 3{,}84 \times 10^{-3} D^{-1}$, o, volviendo a las variables de partida: $B = -3{,}84 \times 10^{-3} D + 1{,}124 \times 10^{-3} D^2$. Hace falta verificar que este modelo sea estrictamente idéntico a la regresión polinomial ponderada de B con respecto a D realizada con la Línea roja 12 con una ponderación proporcional a $1/D^4$. El gráfico de los residuos en función de los valores predichos y el gráfico cuantil-cuantil de los residuos se representan en la Figura 6.16.

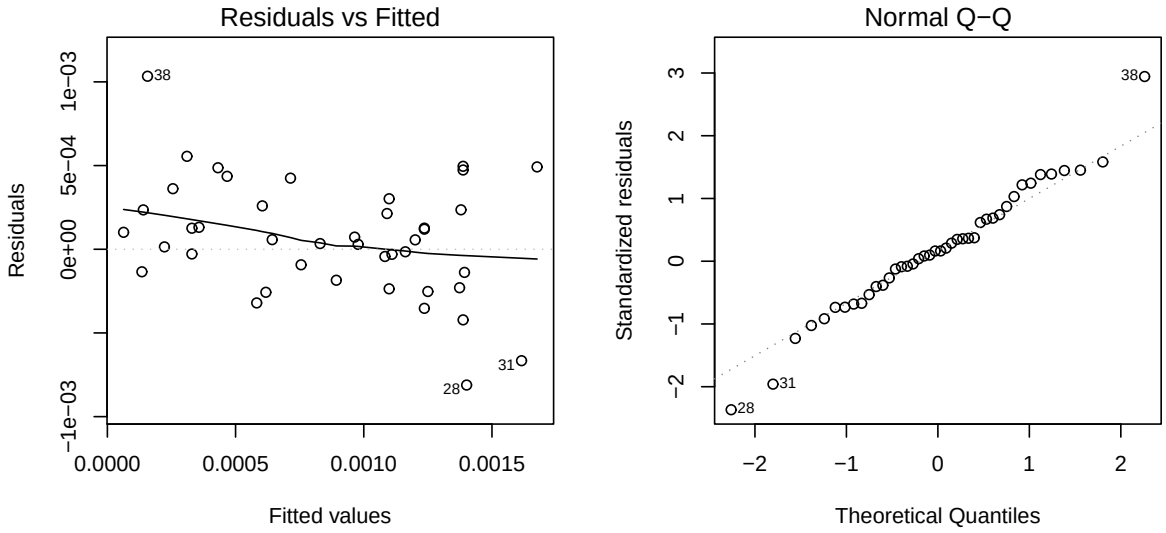


Figura 6.14 – *Gráfico de los residuos en función de los valores predichos (a la izquierda) y gráfico de cuantil-cuantil (a la derecha) de los residuos de la regresión lineal simple de B/D^2 con respecto a H ajustada a los 42 árboles medidos por Henry et al. (2010) en Ghana.*

6.2. Ajuste de un modelo no lineal

Abordemos ahora el caso más general del ajuste de un modelo no lineal. Ese modelo se escribe:

$$Y = f(X_1, \ldots, X_p; \theta) + \varepsilon$$

donde Y es la variable de respuesta, $X_1, \ldots, X_p$ son las variables explicativas, θ es el vector del conjunto de coeficientes del modelo, ε es el error residual, f es una función. Si f es lineal con respecto a los coeficientes θ, volvemos al modelo lineal estudiado anteriormente. Ya no elaboramos ninguna hipótesis *a priori* sobre la linealidad de la función f en relación con los coeficientes θ. Al igual que antes, suponemos que los residuos son independientes y están distribuidos según una distribución normal centrada. Por el contrario, no hay ninguna hipótesis *a priori* sobre su varianza. $\mathrm{E}(\varepsilon) = 0$ implica que $\mathrm{E}(Y) = f(X_1, \ldots, X_p; \theta)$. Por eso se dice que f define el modelo para la media (se sobreentiende: de Y). Planteemos:

$$\mathrm{Var}(\varepsilon) = g(X_1, \ldots, X_p; \vartheta)$$

donde g es una función y ϑ un conjunto de parámetros. Como $\mathrm{Var}(Y) = \mathrm{Var}(\varepsilon)$, decimos que g define el modelo para la varianza. La función g puede asumir formas diversas pero, para los datos de biomasa o de volumen, suele asumir la forma de una función de potencia de una variable que caracteriza el tamaño del árbol (típicamente su diámetro). Sin pérdida de generalidad, plantearemos que esta variable explicativa es X_1, y entonces:

$$g(X_1, \ldots, X_p; \vartheta) \equiv (kX_1^c)^2$$

con $\vartheta \equiv (k, c)$, $k > 0$ y $c \geq 0$.

La interpretación de los resultados del ajuste de un modelo no lineal es fundamentalmente la misma que para el modelo lineal. Además de las propiedades del modelo, la diferencia entre el modelo lineal y el modelo no lineal está asociada a la forma en que se estiman los coeficientes del modelo. Hay que distinguir dos tipos: (*i*) el exponente c está determinado *a priori*; (*ii*) el exponente c es un parámetro por estimar al igual que los otros parámetros del modelo.

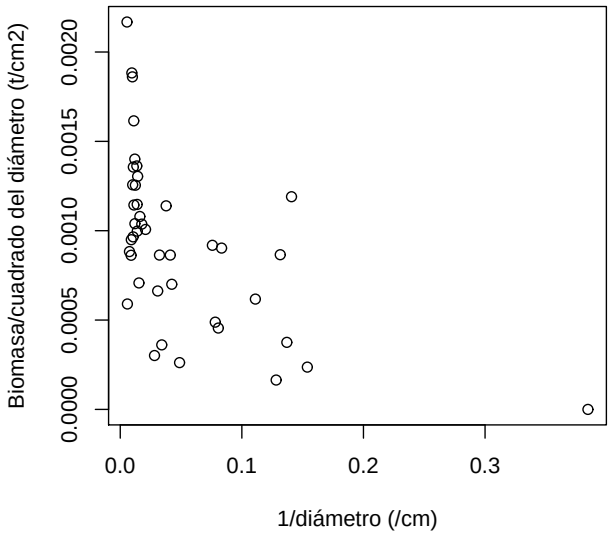


Figura 6.15 – *Nube de puntos de la biomasa dividida por el cuadrado del diámetro (toneladas* cm^{-2}*) en función del inverso el diámetro (*cm^{-1}*) para 42 árboles medidos en Ghana por Henry et al. (2010).*

6.2.1. Exponente conocido

Consideremos primero el caso en el que el exponente c del modelo para la varianza se conoce *a priori*. En ese caso, el método de los mínimos cuadrados puede usarse nuevamente para ajustar el modelo. La suma ponderada de los cuadrados de las desviaciones es:

$$\text{SCE}(\theta) = \sum_{i=1}^{n} w_i \, \varepsilon_i^2 = \sum_{i=1}^{n} w_i \, [Y_i - f(X_{i1}, \ldots, X_{ip}; \theta)]^2$$

donde los pesos son inversamente proporcionales a la varianza de los residuos:

$$w_i = \frac{1}{X_{i1}^{2c}} \propto \frac{1}{\text{Var}(\varepsilon_i)}$$

Al igual que antes, el estimador de los coeficientes del modelo corresponde al valor de θ que minimiza la suma de los cuadrados de las desviaciones ponderadas:

$$\hat{\theta} = \arg\min_{\theta} \text{SCE}(\theta) = \arg\min_{\theta} \left\{ \sum_{i=1}^{n} \frac{1}{X_{i1}^{2c}} [Y_i - f(X_{i1}, \ldots, X_{ip}; \theta)]^2 \right\}$$

En el caso particular en que los residuos tienen una varianza constante (es decir $c = 0$), el método de los mínimos cuadrados ponderados se simplifica en mínimos cuadrados ordinarios (todos los w_i son iguales a 1), pero el principio de los cálculos sigue siendo el mismo. El estimador de θ se obtiene resolviendo

$$\frac{\partial \text{SCE}}{\partial \theta}(\hat{\theta}) = 0 \tag{6.18}$$

con la restricción $(\partial^2 \text{SCE}/\partial \theta^2) > 0$ que garantiza que se trata realmente de un mínimo y no de un máximo. En el caso del modelo lineal, la resolución de (6.18) había dado una expresión explícita para el estimador $\hat{\theta}$. En el caso general del modelo no lineal, ya no es así: no hay una expresión explícita para $\hat{\theta}$. La minimización de la suma de los cuadrados de las desviaciones debe hacerse entonces mediante un algoritmo numérico. En la Sección 6.2.3 entraremos en más detalles sobre este punto.

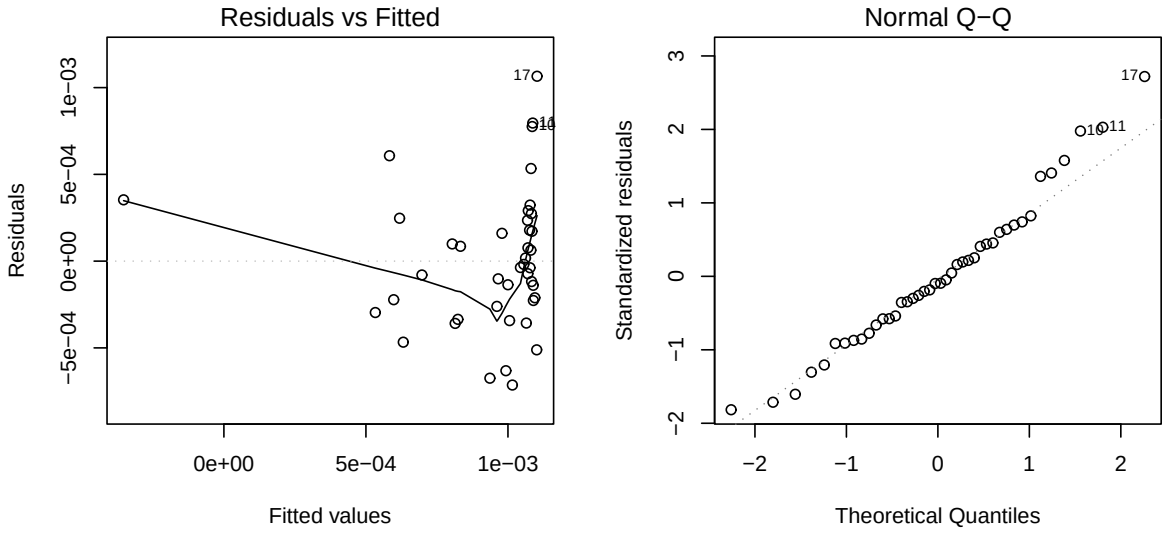


Figura 6.16 – *Gráfico de los residuos en función de los valores predichos (a la izquierda) y gráfico cuantile–cuantile (a la derecha) de los residuos de la regresión lineal simple de B/D^2 con respecto a $1/D$ ajustada a los 42 árboles medidos por Henry et al. (2010) en Ghana.*

Valor a priori del exponente

El valor a priori del exponente c se obtiene en el caso no lineal del mismo modo que en el caso lineal (cf. pág. 128): o bien por tanteo, o bien subdividiendo X_1 en clases y estimando la varianza de Y para cada clase, o bien minimizando el índice de Furnival (cf. pág. 161).

17

Regresión no lineal ponderada entre B y D

La exploración gráfica (Líneas rojas 2 y 5) demostró que la relación entre la biomasa B y el diámetro D era de tipo potencia, con un aumento de la varianza de la biomasa con el diámetro:

$$B = aD^b + \varepsilon$$

con

$$\mathrm{Var}(\varepsilon) \propto D^{2c}$$

Vimos en la Línea roja 11 que la desviación estándar condicional de la biomasa, conociendo el diámetro, era proporcional al cuadrado del diámetro: $c = 2$. Se puede entonces ajustar una regresión no lineal mediante los mínimos cuadrados ponderados usando una ponderación inversamente proporcional a D^4:

```
start <- coef(lm(log(Btot)~I(log(dbh)),data=dat[dat$Btot>0,]))
start[1] <- exp(start[1])
names(start) <- c("a","b")
m <- nls(Btot~a*dbh^b,data=dat,start=start,weights=1/dat$dbh^4)
summary(m)
```

El ajuste de la regresión no lineal se realiza con el comando `nls`, que pide los valores iniciales de los coeficientes. Dichos valores están contenidos en el objeto `start` y se calculan volviendo a transformar los coeficientes de regresión lineal en los datos transformados logarítmicamente. El resultado de ajustar la regresión no lineal mediante los mínimos cuadrados ponderados es:

```
     Estimate  Std. Error  t value  Pr(>|t|)
a  2.492e-04   7.893e-05    3.157   0.00303  **
b  2.346e+00   7.373e-02   31.824    <2e-16  ***
```

con una desviación estándar residual $k = 0{,}0003598$ toneladas cm^{-2}. El modelo se escribe pues: $B = 2{,}492 \times 10^{-4} D^{2{,}346}$. Volvamos a la regresión lineal ajustada a los datos transformados logarítmicamente (Línea roja 7), que se escribía: $\ln(B) = -8{,}42722 + 2{,}36104 \ln(D)$. Si volvemos ingenuamente a los datos de partida aplicando la función exponencial (en § 7.2.4 veremos por qué esto resulta ingenuo), el modelo se convierte en: $B = \exp(-8{,}42722) \times D^{2{,}36104} = 2{,}188 \times 10^{-4} D^{2{,}36104}$. El modelo ajustado por regresión no lineal y el modelo ajustado por regresión lineal en los datos transformados logarítmicamente resultan pues muy próximos.

18

Regresión no lineal ponderada entre B y D^2H

Ya ajustamos un modelo de potencia $B = a(D^2H)^b$ por regresión lineal simple en los datos transformados logarítmicamente (Línea roja 8). Ajustemos ahora ese modelo directamente a través de la regresión no lineal:

$$B = a(D^2H)^b + \varepsilon$$

con

$$\mathrm{Var}(\varepsilon) \propto D^{2c}$$

Para tener en cuenta la heterocedasticidad y considerando que la desviación estándar condicional de la biomasa conociendo el diámetro es proporcional a D^2 (Línea roja 11), podemos ajustar ese modelo no lineal mediante el método de los mínimos cuadrados ponderados, utilizando una ponderación inversamente proporcional a D^4:

```
start <- coef(lm(log(Btot)~I(log(dbh^2*haut)),data=dat[dat$Btot>0,]))
start[1] <- exp(start[1])
names(start) <- c("a","b")
m <- nls(Btot~a*(dbh^2*haut)^b,data=dat,start=start,weights=1/dat$dbh^4)
summary(m)
```

Al igual que antes (Línea roja 17), el comando `nls` pide los valores iniciales de los coeficientes y éstos se obtienen a partir de los coeficientes de la regresión múltiple en los datos transformados logarítmicamente. El resultado del ajuste es:

```
     Estimate  Std. Error  t value  Pr(>|t|)
a  7.885e-05   2.862e-05    2.755    0.0088  **
b  9.154e-01   2.957e-02   30.953    <2e-16  ***
```

con una desviación estándar residual $k = 0{,}0003325$ toneladas cm^{-2}. El modelo se escribe entonces: $B = 7{,}885 \times 10^{-5} (D^2H)^{0{,}9154}$. Volvamos a la regresión lineal ajustada en los datos transformados logarítmicamente (Línea roja 8), que se escribía: $\ln(B) = -8{,}99427 + 0{,}87238 \ln(D^2H)$. Si volvemos ingenuamente a los datos de partida aplicando la función exponencial, este modelo se convierte en: $B = \exp(-8{,}99427) \times D^{0{,}87238} = 1{,}241 \times 10^{-4} D^{0{,}87238}$. El modelo ajustado por regresión no lineal y el modelo ajustado por regresión lineal en los datos transformados logarítmicamente son entonces relativamente próximos.

19

Regresión no lineal ponderada entre B, D y H

Ya ajustamos un modelo de potencia $B = aD^{b_1}H^{b_2}$ por regresión múltiple en los datos transformados logarítmicamente (Línea roja 10). Ajustemos ahora ese modelo directamente por regresión no lineal:

$$B = aD^{b_1}H^{b_2} + \varepsilon$$

con

$$\mathrm{Var}(\varepsilon) \propto D^{2c}$$

Para tener en cuenta la heterocedasticidad y considerando que la desviación estándar condicional de la biomasa conociendo el diámetro es proporcional a D^2 (Línea roja 11), se puede ajustar ese modelo no lineal mediante el método de los mínimos cuadrados ponderados, usando una ponderación inversamente proporcional a D^4:

```
start <- coef(lm(log(Btot)~I(log(dbh))+I(log(haut)),data=dat[dat$Btot>0,]))
start[1] <- exp(start[1])
names(start) <- c("a","b1","b2")
m <- nls(Btot~a*dbh^b1*haut^b2,data=dat,start=start,weights=1/dat$dbh^4)
summary(m)
```

Al igual que antes (Línea roja 17), el comando `nls` pide los valores iniciales de los coeficientes y éstos se obtienen a partir de los coeficientes de la regresión múltiple de los datos transformados logarítmicamente. El resultado del ajuste es:

```
     Estimate  Std. Error  t value  Pr(>|t|)
a   1.003e-04   5.496e-05    1.824    0.0758  .
b1  1.923e+00   1.956e-01    9.833  4.12e-12  ***
b2  7.435e-01   3.298e-01    2.254    0.0299  *
```

con una desviación estándar residual $k = 0{,}0003356$ toneladas cm^{-2}. El modelo se escribe entonces: $B = 1{,}003 \times 10^{-4} D^{1{,}923} H^{0{,}7435}$. El modelo es similar al que había sido ajustado por regresión múltiple en los datos transformados logarítmicamente (Línea roja 10). El coeficiente a se estima, sin embargo, con menor precisión aquí que en el caso de la regresión múltiple en los datos transformados logarítmicamente.

6.2.2. Estimación del exponente

Consideremos ahora el caso en que hay que estimar el exponente c al mismo tiempo que los otros parámetros del modelo. Esto incluye la regresión lineal con modelo de varianza que habíamos evocado en la Sección 6.1.4. El método de los mínimos cuadrados ya no resulta válido en este caso. Así pues tenemos que usar otro método de ajuste: el método de máxima verosimilitud. La verosimilitud de una observación $(X_{i1}, \ldots, X_{ip}, Y_i)$ es la densidad de probabilidad de observar $(X_{i1}, \ldots, X_{ip}, Y_i)$ en el modelo especificado. La densidad de probabilidad de la distribución normal de esperanza μ y de desviación estándar σ es:

$$\phi(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right]$$

Como Y_i está distribuido según una distribución normal de esperanza $f(X_{i1}, \ldots, X_{ip}; \theta)$ y de desviación estándar kX_{i1}^c, la verosimilitud de la i-ésima observación es:

$$\frac{1}{kX_{i1}^c\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{Y_i - f(X_1, \ldots, X_p; \theta)}{kX_{i1}^c}\right)^2\right]$$

Como las observaciones son independientes, su verosimilitud conjunta es el producto de las verosimilitudes de cada una de las observaciones. La verosimilitud de la muestra de n observaciones es por tanto:

$$\begin{aligned}\ell(\theta,\, k,\, c) &= \prod_{i=1}^{n} \frac{1}{kX_{i1}^{c}\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{Y_i - f(X_1,\, \ldots,\, X_p;\theta)}{kX_{i1}^{c}}\right)^2\right] \\ &= \frac{1}{(k\sqrt{2\pi})^n} \frac{1}{(\prod_{i=1}^{n} X_{i1})^c} \exp\left[-\frac{1}{2}\sum_{i=1}^{n}\left(\frac{Y_i - f(X_1,\, \ldots,\, X_p;\theta)}{kX_{i1}^{c}}\right)^2\right]\end{aligned} \tag{6.19}$$

Dicha verosimilitud es considerada como una función de los parámetros θ, k y c.

Los valores de los parámetros θ, k y c serán mucho mejores cuantas más probabilidades haya de obtener las observaciones con el modelo correspondiente a dichos valores de parámetros. En otras palabras, los mejores valores de los parámetros θ, k y c son los que maximizan la verosimilitud de las observaciones. El estimador correspondiente es, por definición, el estimador de la máxima verosimilitud y se escribe:

$$(\hat{\theta},\, \hat{k},\, \hat{c}) = \arg \max_{(\theta,\, k,\, c)} \ell(\theta,\, k,\, c) = \arg \max_{(\theta,\, k,\, c)} \ln[\ell(\theta,\, k,\, c)]$$

donde la última igualdad se deriva del hecho de que una función y su logarítmica alcanzan su máximo para los mismos valores de su argumento. El logaritmo de verosimilitud, que llamamos log-verosimilitud y que escribimos como $\mathcal{L}$, es más fácil de calcular que la verosimilitud y, por eso, para nuestros cálculos, lo que se trata de maximizar es la log-verosimilitud. En este caso, la log-verosimilitud se escribe:

$$\begin{aligned}\mathcal{L}(\theta,\, k,\, c) &= \ln[\ell(\theta,\, k,\, c)] \\ &= -n\ln(k\sqrt{2\pi}) - c\sum_{i=1}^{n}\ln(X_{i1}) - \frac{1}{2}\sum_{i=1}^{n}\left(\frac{Y_i - f(X_1,\, \ldots,\, X_p;\theta)}{kX_{i1}^{c}}\right)^2 \\ &= -\frac{1}{2}\sum_{i=1}^{n}\left[\left(\frac{Y_i - f(X_1,\, \ldots,\, X_p;\theta)}{kX_{i1}^{c}}\right)^2 + \ln(2\pi) + \ln(k^2 X_i^{2c})\right]\end{aligned} \tag{6.20}$$

Para obtener los estimadores de la máxima verosimilitud de los parámetros, habría que calcular las derivadas parciales de la log-verosimilitud con respecto a esos parámetros y buscar los valores en que se anulan (asegurándose al mismo tiempo de que las segundas derivadas son realmente negativas). En general, no hay una solución analítica a este problema. Al igual que antes, para la suma de los cuadrados de las desviaciones, habrá que recurrir a un algoritmo numérico para maximizar la log-verosimilitud.

Se puede demostrar que el método de máxima verosimilitud lleva a un estimador de los coeficientes que es mejor asintóticamente (es decir, cuando el número n de observaciones tiende al infinito). Podemos demostrar también que en el caso del modelo lineal, el estimador de los mínimos cuadrados y el estimador de máxima verosimilitud son iguales.

20

Regresión no lineal entre B y D con modelo de varianza

Volvamos a la regresión no lineal entre la biomasa y el diámetro (cf. Línea roja 17) pero considerando ahora que el exponente c del modelo para la varianza es un parámetro por estimar como los otros. El modelo se escribe del mismo modo que antes (Línea roja 17):

$$B = aD^b + \varepsilon$$

con

$$\mathrm{Var}(\varepsilon) = (kD^c)^2$$

pero se ajusta por el método de máxima verosimilitud:

```
start <- coef(lm(log(Btot)~I(log(dbh)),data=dat[dat$Btot>0,]))
start[1] <- exp(start[1])
names(start) <- c("a","b")
library(nlme)
m <- nlme(Btot~a*dbh^b, data=cbind(dat,g="a"), fixed=a+b~1, start=start, groups=~g,
weights=varPower(form=~dbh))
summary(m)
```

El ajuste se hace mediante el comando **nlme**[1], que, al igual que el comando **nls** (Línea roja 17) requiere los valores iniciales de los coeficientes (dados por el comando **start**). Dichos valores iniciales se calculan como en la Línea roja 17. El resultado del ajuste es:

```
      Value   Std.Error  DF   t-value  p-value
a 0.0002445 0.00007136  40   3.42568   0.0014
b 2.3510500 0.06947401  40 33.84071    0.0000
```

con un valor estimado del exponente $c = 2{,}090814$. Dicho valor estimado es muy próximo del valor evaluado por la regresión no lineal ponderada ($c = 2$, cf. Línea roja 11). El modelo ajustado se escribe entonces: $B = 2{,}445 \times 10^{-4} D^{2{,}35105}$, lo que es muy próximo del modelo ajustado por regresión no lineal ponderada (Línea roja 17).

21

Regresión no lineal entre B y D^2H con modelo de varianza

Retomemos la regresión no lineal entre la biomasa y D^2H (cf. Línea roja 18) pero considerando ahora que el exponente c del modelo para la varianza es un parámetro por estimar como los otros. El modelo se escribe del mismo modo que el anterior (Línea roja 18):

$$B = a(D^2H)^b + \varepsilon$$

con

$$\mathrm{Var}(\varepsilon) = (kD^c)^2$$

pero se ajusta por el método de máxima verosimilitud:

```
start <- coef(lm(log(Btot)~I(log(dbh^2*haut)),data=dat[dat$Btot>0,]))
start[1] <- exp(start[1])
names(start) <- c("a","b")
library(nlme)
m <- nlme(Btot~a*(dbh^2*haut)^b,data=cbind(dat,g="a"),fixed=a+b~1,start=start,
groups=~g,weights=varPower(form=~dbh))
summary(m)
```

[1] El comando **nlme** sirve en realidad para ajustar los modelos no lineales con efecto mixto. El comando **nlreg** ajusta los modelos no lineales con modelo de varianza, pero hemos obtenido resultados anormales con este comando (versión 3.1-96), lo que explica que hayamos preferido aquí usar **nlme**, aun cuando no haya efecto mixto en los modelos considerados aquí.

El ajuste se realiza mediante el comando `nlme`, la que, al igual que `nls` (Línea roja 17) requiere los valores iniciales de los coeficientes. Dichos valores iniciales `start` se calculan como en la Línea roja 17. El resultado del ajuste es:

```
        Value     Std.Error  DF    t-value  p-value
a   0.0000819  0.000028528  40    2.87214   0.0065
b   0.9122144  0.028627821  40   31.86461   0.0000
```

con un valor estimado del exponente $c = 2{,}042586$. Dicho valor estimado es muy próximo al valor evaluado por la regresión no lineal ponderada ($c = 2$, cf. Línea roja 11). El modelo ajustado se escribe entonces: $B = 8{,}19 \times 10^{-5}(D^2H)^{0{,}9122144}$, lo que es muy próximo del modelo ajustado por regresión no lineal ponderada (Línea roja 18).

22

Regresión no lineal entre B, D y H con modelo de varianza

Retomemos la regresión no lineal entre la biomasa, el diámetro y la altura (cf. Línea roja 19):

$$B = aD^{b_1}H^{b_2} + \varepsilon$$

con

$$\mathrm{Var}(\varepsilon) = (kD^c)^2$$

pero considerando ahora que el exponente c del modelo para la varianza es un parámetro por estimar como los demás. El ajuste por la máxima verosimilitud:

```
library(nlme)
start <- coef(lm(log(Btot)~I(log(dbh))+I(log(haut)),data=dat[dat$Btot>0,]))
start[1] <- exp(start[1])
names(start) <- c("a","b1","b2")
m <- nlme(Btot~a*dbh^b1*haut^b2,data=cbind(dat,g="a"),fixed=a+b1+b2~1,
start=start,groups=~g,weights=varPower(form=~dbh))
summary(m)
```

requiere, al igual que antes, que se den los valores iniciales de los coeficientes (por medio del comando `start`). El ajuste da:

```
         Value  Std.Error  DF   t-value  p-value
a   0.0001109  0.0000566  39  1.959869   0.0572
b1  1.9434876  0.1947994  39  9.976866   0.0000
b2  0.6926256  0.3211766  39  2.156526   0.0373
```

con un valor estimado del exponente $c = 2{,}055553$. Este valor estimado es muy próximo al valor evaluado para la regresión no lineal ponderada ($c = 2$, cf. Línea roja 11). El modelo ajustado se escribe entonces: $B = 1{,}109 \times 10^{-4}D^{1{,}9434876}H^{0{,}6926256}$, lo que es muy próximo del modelo ajustado por regresión no lineal ponderada (Línea roja 19).

23

Regresión no lineal entre B y un polinomio de $\ln(D)$

Antes (Línea roja 9), por regresión múltiple un modelo entre $\ln(B)$ y un polinomio de $\ln(D)$. Si volvemos a las variables de partida, el modelo se escribe:

$$B = \exp\{a_0 + a_1 \ln(D) + a_2[\ln(D)]^2 + \ldots + a_p[\ln(D)]^p\} + \varepsilon$$

con

$$\text{Var}(\varepsilon) = (kD^c)^2$$

Ahora vamos a ajustar este modelo no lineal directamente por la máxima verosimilitud (de forma tal que el exponente c se estime al mismo tiempo que los otros parámetros del modelo). Para un polinomio de grado 3, el ajuste se obtiene mediante:

```
library(nlme)
start <- coef(lm(log(Btot)~I(log(dbh))+I(log(dbh)^2)+I(log(dbh)^3),data=dat[
dat$Btot>0,]))
start[1] <- exp(start[1])
names(start) <- paste("a",0:3,sep="")
m <- nlme(Btot~exp(a0+a1*log(dbh)+a2*log(dbh)^2+a3*log(dbh)^3),data=cbind(dat,
g="a"),fixed=a0+a1+a2+a3~1,start=start,groups=~g,weights=varPower(form=~dbh))
summary(m)
```

y el resultado del ajuste es:

```
      Value  Std.Error  DF    t-value  p-value
a0 -8.983801  2.2927006  38 -3.918436   0.0004
a1  2.939020  2.1073819  38  1.394631   0.1712
a2 -0.158585  0.6172529  38 -0.256921   0.7986
a3  0.013461  0.0581339  38  0.231547   0.8181
```

Con un valor estimado del exponente $c = 2{,}099938$. Encontramos un resultado muy parecido al que había sido obtenido por regresión múltiple en los datos transformados logarítmicamente (Línea roja 9).

6.2.3. Optimización numérica

Hay que recurrir a un algoritmo de optimización numérica para minimizar la suma de los cuadrados de las desviaciones (cuando se conoce el exponente c) o para maximizar la log-verosimilitud (cuando debe estimarse el exponente c). Maximizar la log-verosimilitud equivale a minimizar lo opuesto de la log-verosimilitud, con lo cual, a continuación sólo se considerará el problema de la minimización de una función en un espacio multidimensional. Existen muchísimos algoritmos de optimización (Press *et al.*, 2007, Capítulo 10) y aquí el objetivo no es enumerarlos. Lo que importa saber a estas alturas es que dichos algoritmos son iterativos y exigen un valor de partida de los parámetros. A partir de este valor inicial y en cada iteración, el algoritmo se desplaza en el espacio de los parámetros buscando minimizar la función objetivo (a saber, la suma de los cuadrados de las desviaciones o menos la log-verosimilitud). Se puede representar la función objetivo como una hipersuperficie en el espacio de los parámetros (Figura 6.17). Cada posición en ese espacio corresponde a un valor de los parámetros. Una protuberancia en esa superficie corresponde a un máximo local de la función objetivo, mientras que una concavidad de la superficie corresponde a un mínimo local. El objetivo es encontrar el mínimo global, es decir, la concavidad más profunda. La posición de esta concavidad corresponde al valor estimado de los parámetros. Si el algoritmo indica la posición de una concavidad que no es la más profunda, la estimación de los parámetros es falsa.

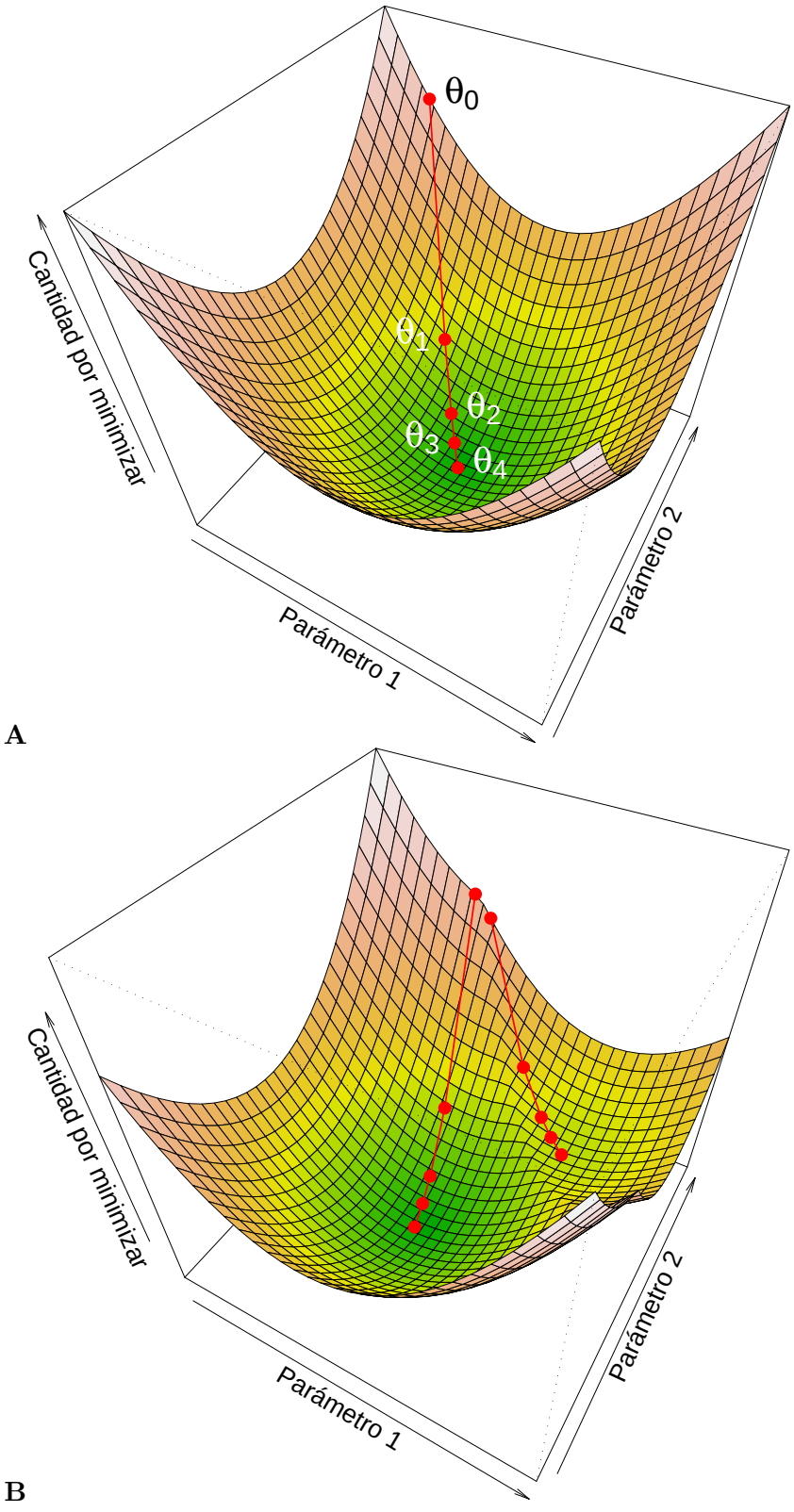


Figura 6.17 – *Representación de la función objetivo (p.ej., la cantidad por minimizar) como una superficie en el espacio de los parámetros. Cada posición en ese espacio corresponde a un valor de los parámetros. Los valores sucesivos θ_1, θ_2, ..., de los parámetros se obtienen a partir de un valor inicial θ_0 al descender por la superficie según la mayor pendiente. (A) La superficie tiene una única cuenca. (B) La superficie tiene varias cuencas.*

Algoritmo de descenso

El algoritmo de optimización más simple consiste en calcular las posiciones sucesivas, es decir, los valores sucesivos de los parámetros descendiendo la superficie definida mediante la función objetivo según su línea de mayor pendiente (Figura 6.17A). Este algoritmo conduce a una concavidad de la superficie pero nada nos indica que esa concavidad sea la más profunda. En efecto, la superficie puede tener varias cuencas con varias concavidades. Según la posición de partida, el algoritmo convergerá en una concavidad o en otra (Figura 6.17B). Más aún, dos posiciones iniciales muy próximas, cada una de un lado distinto de la línea de cresta que separa las dos cuencas, llevarán a dos concavidades diferentes, es decir, a estimaciones distintas de los parámetros. El único caso en que este algoritmo da la buena estimación de los parámetros independientemente del valor inicial de los mismos es cuando la superficie tiene una concavidad única, es decir, cuando la función objetivo es convexa. Este es el caso especialmente para el modelo lineal pero no suele ser cierto para el modelo no lineal.

Mejora de los algoritmos en caso de mínimos locales

Se dispone de algoritmos más sutiles que el de descenso según la mayor pendiente. Por ejemplo, se puede dar la posibilidad de volver a salir de una concavidad en la cual haya convergido temporalmente el algoritmo para explorar si no hay una concavidad más profunda en los alrededores. No obstante, ningún algoritmo, ni siquiera el más sutil, ofrece la certeza de que haya convergido realmente en la concavidad más profunda. Así pues, cualquier algoritmo de optimización numérica (i) puede ser atrapado por un mínimo local en vez de convergir en el mínimo global, y (ii) es sensible a la posición de partida indicada, que determina parcialmente la posición final en la que convergirá el algoritmo.

Si volvemos al problema que nos interesa, esto significa (i) que el ajuste de un modelo no lineal podrá dar estimaciones erróneas de los parámetros y (ii) que la elección de los valores iniciales de los parámetros para el algoritmo de optimización es un asunto delicado. Aquí reside el principal inconveniente del ajuste de un modelo no lineal. Para limitar este inconveniente, habrá que escoger cuidadosamente el valor inicial de los parámetros y, sobre todo, someter a prueba varios de ellos.

Elección del valor inicial de los parámetros

Cuando el modelo f para la media puede transformarse en una relación lineal entre la variable de respuesta Y y las variables explicativas $X_1, \ldots, X_p$, puede obtenerse un valor de partida de los coeficientes ajustando una regresión lineal a las variables transformadas sin tener en cuenta la heterocedasticidad eventual de los residuos. Tomemos el ejemplo de un modelo de biomasa de tipo potencia:

$$B = aD^{b_1}H^{b_2}\rho^{b_3} + \varepsilon \tag{6.21}$$

con

$$\varepsilon \underset{\text{i.i.d.}}{\sim} \mathcal{N}(0,\ kD^c)$$

El modelo de potencia para la esperanza de B puede ser linealizado al transformar las variables logarítmicamente: $\ln(B) = a' + b_1\ln(D) + b_2\ln(H) + b_3\ln(\rho)$. Sin embargo, esta transformación no es compatible con la aditividad de los errores en el modelo (6.21). En otras palabras, la regresión múltiple de la variable de respuesta $\ln(B)$ con respecto a las variables explicativas $\ln(D)$, $\ln(H)$ y $\ln(\rho)$:

$$\ln(B) = a' + b_1\ln(D) + b_2\ln(H) + b_3\ln(\rho) + \varepsilon' \tag{6.22}$$

con $\varepsilon' \sim \mathcal{N}(0, \sigma)$, no es un modelo equivalente a (6.21), aunque los residuos ε' de este modelo tengan una varianza constante. Incluso si los modelos (6.21) y (6.22) no son matemáticamente equivalentes, los coeficientes de (6.22) estimados por regresión múltiple pueden servir de valores iniciales para el algoritmo numérico que estima los coeficientes de (6.21). Si anotamos como $x^{(0)}$ el valor inicial del parámetro x para el algoritmo de optimización numérica, tendremos entonces:

$$a^{(0)} = \exp(\hat{a}'), \quad b_i^{(0)} = \hat{b}_i, \quad k^{(0)} = \hat{\sigma}, \quad c^{(0)} = 0$$

A veces el modelo para la media no es linealizable. Por ejemplo, el siguiente modelo parametrado que se usa para los árboles en plantación (Saint-André *et al.*, 2005):

$$B = a + [b_0 + b_1 T + b_2 \exp(-b_3 T)] D^2 H + \varepsilon$$

donde T es la edad de la plantación y $\varepsilon \sim \mathcal{N}(0, kD^c)$, tiene un modelo para la media que no es linealizable. En este caso, los valores iniciales de los parámetros tendrán que elegirse de forma empírica. En este ejemplo preciso, se podría tomar:

$$a^{(0)} = \hat{a}, \quad b_0^{(0)} + b_2^{(0)} = \hat{b}_0, \quad b_1^{(0)} = \hat{b}_1, \quad b_3^{(0)} = 0, \quad k^{(0)} = \hat{\sigma}, \quad c^{(0)} = 0$$

donde $\hat{a}$, $\hat{b}_0$, $\hat{b}_1$ y $\hat{\sigma}$ son los valores estimados de los coeficientes y la desviación estándar residual de la regresión múltiple de B con respecto a D^2H y D^2HT.

La elección de los valores iniciales de los parámetros no nos exime de probar varios valores iniciales. Cuando ajustamos un modelo no lineal con un algoritmo de optimización numérica, es fundamental probar varios valores iniciales de los parámetros para garantizar la estabilidad de las estimaciones.

6.3. Selección de variables y modelos

Cuando queremos construir un modelo de volumen o de biomasa, la exploración gráfica de los datos (Capítulo 5) suele dar varias formas posibles del modelo. Se pueden ajustar todos los modelos potencialmente interesantes. Pero, al final, entre todos los modelos ajustados, ¿cuál elegir y recomendar al usuario? La selección de variables y la selección de modelos tiene por objeto determinar cuál es la "mejor" expresión posible del modelo entre todas aquellas que fueron ajustadas.

6.3.1. Selección de variables

Tomemos el ejemplo de un modelo de biomasa que queremos construir a partir de un conjunto de datos que incluyen el diámetro de los árboles, su altura y la densidad específica de la madera. Si trabajamos sobre los datos transformados logarítmicamente y según las variables incluidas en el modelo, se podrán ajustar los modelos siguientes:

$$\begin{aligned}
\ln(B) &= a_0 + a_1 \ln(D) + \varepsilon \\
\ln(B) &= a_0 + a_2 \ln(H) + \varepsilon \\
\ln(B) &= a_0 + a_3 \ln(\rho) + \varepsilon \\
\ln(B) &= a_0 + a_1 \ln(D) + a_2 \ln(H) + \varepsilon \\
\ln(B) &= a_0 + a_1 \ln(D) + a_3 \ln(\rho) + \varepsilon \\
\ln(B) &= a_0 + a_2 \ln(H) + a_3 \ln(\rho) + \varepsilon \\
\ln(B) &= a_0 + a_1 \ln(D) + a_2 \ln(H) + a_3 \ln(\rho) + \varepsilon
\end{aligned}$$

Llamamos modelo *completo* al modelo que incluye todas las variables explicativas disponibles (el último de la lista anterior). Todos los otros modelos pueden considerarse como subconjuntos del modelo completo en los cuales ciertas variables explicativas fueron utilizadas y otras dejadas de lado. La selección de variables pretende elegir, entre las variables explicativas de un modelo completo, las que hay que retener y aquellas que hay que descartar porque aportan poco a la predicción de la variable de respuesta. En otras palabras, en este ejemplo, la selección de variables consistiría en elegir el mejor modelo entre los siete previstos para $\ln(B)$.

Dadas p variables explicativas X_1, X_2, ..., X_p, hay $2^p - 1$ modelos que incluyen todo o parte de dichas variables. La selección de variables consiste en elegir la "mejor" combinación de variables explicativas entre todas las disponibles. Esto significa, ante todo, que existe un criterio que permite evaluar la calidad de un modelo. Ya vimos (pág. 120) que R^2 es un mal criterio para evaluar la calidad de un modelo con respecto a otro, ya que aumenta automáticamente con el número de variables explicativas, independientemente de qué tanto aporten realmente información para la predicción de la variable de respuesta. Un criterio mejor para seleccionar las variables explicativas es el estimador de la varianza residual que está asociado a R^2 a través de la relación:

$$\hat{\sigma}^2 = \frac{n}{n-p-1}(1-R^2)\,S_Y^2$$

donde S_Y^2 es la varianza empírica de la variable de respuesta.

La búsqueda de la mejor combinación de variables explicativas puede hacerse de varias maneras. Si p no es demasiado elevado, se puede pasar revista a los $2^p - 1$ modelos posibles en forma exhaustiva. Si p es demasiado elevado, puede usarse un método paso a paso de selección de variables. Los métodos paso a paso proceden por eliminación sucesiva o agregado sucesivo de variables explicativas. El método descendente consiste en eliminar la variable menos significativa entre las p. Se vuelve a calcular entonces la regresión y se recomienza hasta que se satisfaga un criterio de detención (por ejemplo, todos los coeficientes del modelo son significativamente diferentes de cero). El método ascendente actúa en sentido inverso: se parte de la mejor regresión con una variable y se agregan, una por una, las variables que hacen avanzar más el R^2, hasta que se satisfaga el criterio de detención.

El método llamado *stepwise* es un perfeccionamiento que consiste en efectuar además en cada paso pruebas de significatividad de tipo Fisher para evitar introducir una variable no significativa y eliminar eventualmente variables ya introducidas que no serían más informativas, teniendo en cuenta la última variable seleccionada. El algoritmo se detiene cuando ya no se pueden agregar ni quitar variables. Los distintos métodos de selección paso a paso no dan siempre el mismo resultado, por lo que parece mejor el método "stepwise". Sin embargo, no nos protegen de una eliminación repentina de variables realmente significativas, lo que podría sesgar los resultados. Cabe además recordar que si sabemos (por motivos biológicos) que una variable debe figurar en un modelo (la densidad específica de la madera, por ejemplo), no debe rechazarse porque una prueba estadística la declare no significativa (debido al riesgo de cometer un error tipo II).

24

Selección de variables

Hagamos una selección de variables $\ln(D)$, $[\ln(D)]^2$, $[\ln(D)]^3$, $\ln(H)$ para predecir el logaritmo de la biomasa. El modelo completo se escribe entonces:

$$\ln(B) = a_0 + a_1\ln(D) + a_2[\ln(D)]^2 + a_3[\ln(D)]^3 + a_4\ln(H) + \varepsilon$$

con

$$\mathrm{Var}(\varepsilon) = \sigma^2$$

La selección de variable en R se realiza con el comando `step` aplicado al modelo completo ajustado

```
m <- lm(log(Btot)~I(log(dbh))+I(log(dbh)^2)+I(log(dbh)^3)+I(log(haut)),data=dat[
dat$Btot>0,])
summary(step(m))
```

lo que da:

```
               Estimate  Std. Error  t value  Pr(>|t|)
 (Intercept)    -6.50202     0.35999  -18.062    <2e-16  ***
 I(log(dbh)^2)   0.23756     0.01972   12.044  1.53e-14  ***
 I(log(haut))    1.01874     0.17950    5.675  1.59e-06  ***
```

Las variables seleccionadas son pues $[\ln(D)]^2$ y $\ln(H)$. El modelo retenido finalmente se escribe: $\ln(B) = -6{,}50202 + 0{,}23756[\ln(D)]^2 + 1{,}01874\ln(H)$, con una desviación estándar residual de 0,3994 y $R^2 = 0{,}974$.

6.3.2. Selección de modelos

Dados dos modelos concurrentes que predicen la misma variable de respuesta excepto por una transformación de variable, ¿cuál escoger? Para responder a esta pregunta hay que considerar varias posibilidades.

Modelos anidados

El caso más sencillo es cuando ambos modelos por comparar son anidados. Un modelo está *anidado* en otro si ambos predicen la misma variable de respuesta y si se puede pasar del segundo al primero suprimiendo una o varias variables explicativas. Por ejemplo, el modelo de biomasa $B = a_0 + a_1 D + \varepsilon$ está anidado en $B = a_0 + a_1 D + a_2 D^2 H + \varepsilon$ porque pasa del segundo al primero al suprimir $D^2 H$ de las variables explicativas. Del mismo modo, el modelo $B = a_0 + a_1 D^2 H + \varepsilon$ está anidado en $B = a_0 + a_1 D + a_2 D^2 H + \varepsilon$ porque pasa del segundo al primero al suprimir D de las variables explicativas. Por el contrario, el modelo $B = a_0 + a_1 D + \varepsilon$ no está anidado en $B = a_0 + a_2 D^2 H + \varepsilon$.

Consideremos p como el número de variables explicativas del modelo completo y $p' < p$ el número de variables explicativas del modelo anidado. Sin pérdida de generalidad, se puede escribir el modelo completo de la siguiente forma:

$$Y = f(X_1, \ldots, X_{p'}, X_{p'+1}, \ldots, X_p; \theta_0, \theta_1) + \varepsilon \tag{6.23}$$

donde (θ_0, θ_1) es el vector de los coeficientes asociados al modelo completo y θ_0 es el vector de los coeficientes asociados al modelo anidado, que se obtiene planteando $\theta_1 = \mathbf{0}$. En particular en el caso del modelo lineal, el modelo completo se obtiene como la suma del modelo anidado y de los términos adicionales:

$$\underbrace{\underbrace{Y = a_0 + a_1 X_1 + \ldots + a_{p'} X_{p'}}_{\text{modelo anidado}} + a_{p'+1} X_{p'+1} + \ldots + a_p X_p + \varepsilon}_{\text{modelo completo}} \tag{6.24}$$

con $\theta_0 = (a_0, \ldots, a_{p'})$ y $\theta_1 = (a_{p'+1}, \ldots, a_p)$.

En el caso de los modelos anidados, se puede someter a prueba un modelo comparándolo con el otro mediante un test estadístico. La hipótesis nula de este test es $\theta_1 = \mathbf{0}$, es decir: los términos adicionales no son significativos, lo que también puede formularse como: el modelo anidado es mejor que el modelo completo. Si el p-value de este test resulta inferior al umbral de significancia (típicamente 5 %), entonces se rechaza la hipótesis nula, es decir que el modelo completo es mejor. Por el contrario, si el p-value es superior al umbral de significancia, el modelo anidado se considera mejor.

En el caso del modelo lineal (6.24), el estadístico de la prueba es una razón de los cuadrados medios que, bajo la hipótesis nula, sigue la distribución de Fisher. Por lo demás, se trata del mismo tipo de pruebas que la usada para comprobar el carácter significativo global de una regresión múltiple o la que se usó en el método "stepwise" de selección de variables. En el caso general del modelo no lineal (6.23), la estadística de la prueba es una razón de verosimilitud en la que $-2log$(relación de verosimilitud) sigue, bajo la hipótesis nula, una distribución de χ^2.

25

Prueba de modelos anidados: $\ln(D)$

En la Línea roja 24, la variable $[\ln(D)]^2$ fue seleccionada con $\ln(H)$ como variables explicativas de $\ln(B)$, pero no de $\ln(D)$. El modelo $\ln(B) = a_0 + a_1 \ln(D) + a_2[\ln(D)]^2 + a_4 \ln(H)$, que incluye el término adicional $\ln(D)$, puede compararse al modelo $\ln(B) = a_0 + a_2[\ln(D)]^2 + a_4 \ln(H)$, usando la prueba de modelos anidados. El comando de R que permite probar un modelo anidado es `anova`, con, como primer argumento, el modelo anidado y, como segundo argumento, el modelo completo:

```
comp <- lm(log(Btot)~I(log(dbh))+I(log(dbh)^2)+I(log(haut)),data=dat[dat$Btot>0,])
nest <- lm(log(Btot)~I(log(dbh)^2)+I(log(haut)),data=dat[dat$Btot>0,])
anova(nest,comp)
```

El resultado de la prueba es el siguiente:

```
  Res.Df     RSS  Df  Sum of Sq      F  Pr(>F)
1     38  6.0605
2     37  5.8964   1    0.16407 1.0295  0.3169
```

El p-value vale 0,3169 y es pues superior a 5 %. El modelo anidado (sin $\ln(D)$) se selecciona entonces en detrimento del modelo completo.

26

Prueba de modelos anidados: $\ln(H)$

En la Línea roja 7, se obtuvo el modelo $\ln(B) = -8{,}42722 + 2{,}36104 \ln(D)$ mientras que en la Línea roja 10, se obtuvo el modelo $\ln(B) = -8{,}9050 + 1{,}8654 \ln(D) + 0{,}7083 \ln(H)$ Al estar el primero anidado en el segundo, se puede probar cuál es el mejor. El comando:

```
comp <- lm(log(Btot)~I(log(dbh))+I(log(haut)),data=dat[dat$Btot>0,])
nest <- lm(log(Btot)~I(log(dbh)),data=dat[dat$Btot>0,])
anova(nest,comp)
```

da:

```
  Res.Df     RSS  Df  Sum of Sq      F    Pr(>F)
1     39  8.3236
2     38  6.4014   1     1.9222 11.410  0.001698  **
```

Al ser el p-value inferior a 5 %, el modelo completo (que incluye $\ln(H)$ como variable explicativa) se selecciona en detrimento del modelo anidado.

Modelos con la misma variable de respuesta

Si queremos comparar dos modelos que tienen la misma variable de respuesta pero que no están anidados, ya no se puede usar el test estadístico. Por ejemplo, no se puede utilizar la prueba presentada anteriormente para comparar $B = a_0 + a_1 D + \varepsilon$ y $B = a_0 + a_2 D^2 H + \varepsilon$. En este caso, se usará un criterio de información (Bozdogan, 1987; Burnham & Anderson, 2002, 2004). Existen varios, adaptados a distintos contextos. Los más usados son el criterio de información bayesiano ("Bayesian information criterion" o BIC) y sobre todo el criterio de información de Akaike (1974) ("Akaike information criterion" o AIC). El AIC se expresa como:

$$\text{AIC} = -2 \ln \ell(\hat{\theta}) + 2q$$

donde $\ell(\hat{\theta})$ es la verosimilitud del modelo, es decir, la verosimilitud de la muestra para los valores estimados de los parámetros del modelo (cf. ecuación 6.19), y q es el número de parámetros libres estimados. En particular, en el caso de una regresión múltiple con respecto a p variables explicativas, $q = p + 1$ (o sea, los p coeficientes asociados a las p variables explicativas más la intersección). El coeficiente -2 ante la log-verosimilitud en la expresión del AIC es idéntico al usado para el estadístico de prueba de la razón de verosimilitud en el caso de los modelos anidados. Dados dos modelos con el mismo número de parámetros, el mejor modelo será el que tenga la mayor verosimilitud, es decir, el que tenga el AIC menor. A igualdad de verosimilitud, el mejor modelo será el que tenga menos parámetros (según el principio de parsimonia o navaja de Occam), o sea, una vez más el que tiene el AIC menor. Al final de cuentas, el modelo mejor será el que tenga el menor valor de AIC.

El BIC es una expresión parecida al AIC pero con un término de penalización de parámetros mayor:

$$\text{BIC} = -2 \ln \ell(\hat{\theta}) + q \ln(n)$$

donde n es el número de observaciones. Una vez más, aquí también el mejor modelo será aquel con el menor valor del BIC. En el caso del ajuste de modelos de volumen o de biomasa, se usará más bien el AIC que el BIC como criterio de selección de modelos.

27

Selección de modelos con B como variable de respuesta

Los siguientes modelos con B como variable de respuesta fueron ajustados:

- Línea roja 12 o 16: $B = -3{,}840 \times 10^{-3} D + 1{,}124 \times 10^{-3} D^2$
- Línea roja 14: $B = -3{,}319456 \times 10^{-3} D + 1{,}067068 \times 10^{-3} D^2$
- Línea roja 17: $B = 2{,}492 \times 10^{-4} D^{2{,}346}$
- Línea roja 20: $B = 2{,}445 \times 10^{-4} D^{2{,}35105}$
- Línea roja 11 o 15: $B = 2{,}747 \times 10^{-5} D^2 H$

- Línea roja 13: $B = 2{,}740688 \times 10^{-5} D^2 H$
- Línea roja 18: $B = 7{,}885 \times 10^{-5} (D^2 H)^{0{,}9154}$
- Línea roja 21: $B = 8{,}19 \times 10^{-5} (D^2 H)^{0{,}9122144}$
- Línea roja 19: $B = 1{,}003 \times 10^{-4} D^{1{,}923} H^{0{,}7435}$
- Línea roja 22: $B = 1{,}109 \times 10^{-4} D^{1{,}9434876} H^{0{,}6926256}$

Los modelos de las líneas rojas 12, 14, 11 y 13 se ajustan mediante regresión lineal mientras que los otros se ajustan por regresión no lineal. Hay cinco formas diferentes de modelos y, para cada uno, dos modos de ajustes según una regresión ponderada por el método de los mínimos cuadrados ponderados (Líneas rojas 12, 17, 11, 18 y 19) o según una regresión con un modelo de varianza por el método de máxima verosimilitud (Líneas rojas 14, 20, 13, 21 y 22). La Figura 6.18 compara las predicciones de estos diferentes modelos. Consideremos a `m` como uno de los modelos ajustados que tiene el diámetro como única entrada. El gráfico de las predicciones para este modelo se obtiene como sigue:

```
with(dat,plot(dbh,Btot,xlab="Diámetro (cm)",ylab="Biomasa (t)"))
D <- seq(par("usr")[1],par("usr")[2],length=200)
lines(D,predict(m,newdata=data.frame(dbh=D)),col="red")
```

Para un modelo `m` que tenga el diámetro y la altura como entradas, las predicciones se obtienen como sigue:

```
D <- seq(0,180,length=20)
H <- seq(0,61,length=20)
B <- matrix(predict(m,newdata=expand.grid(dbh=D,haut=H)),length(D))
```

y el gráfico de la superficie de respuesta de la biomasa en función del diámetro y de la altura se obtiene mediante:

```
M <- persp(D,H,B,xlab="Diámetro (cm)",ylab="Altura (m)",zlab="Biomasa (t)",
ticktype="detailed")
points(trans3d(dat$dbh,dat$haut,dat$Btot,M))
```

Dado un modelo ajustado `m`, su AIC se calcula mediante el comando:

```
AIC(m)
```

Para los 10 modelos enumerados anteriormente, los valores de los AIC se dan en el Cuadro 6.1. Dicho Cuadro pone de manifiesto un problema que presentan varios software estadísticos, incluido R: cuando se maximiza la log-verosimilitud (6.20), cualquier término constante (tal como $-n \ln(2\pi)/2$) no desempeñan un papel. La constante que se usa para calcular la log-verosimilitud y, en consecuencia, el AIC, es pues una cuestión de convención, y se utilizan diferentes constantes según los cálculos. En el Cuadro 6.1, se ve pues que los valores de AIC de los modelos ajustados por el comando `nls` son claramente muy superiores a aquellos de los otros modelos: : no se trata de que esos modelos sean peores que los otros sino, simplemente, que el comando `nls` utiliza otra constante distinta de los otros para el cálculo de la log-verosimilitud. Hay que tener en cuenta que al usar R sólo hay que comparar valores del AIC para los modelos que fueron ajustados por medio del mismo comando.

En este caso, si se comparan los dos modelos que fueron ajustados con el comando `lm`, el mejor (es decir aquel con el AIC menor) es el que tiene D^2H como variable explicativa (Línea roja 11). Si comparamos los cinco modelos ajustados con el comando `nlme`, el mejor también es aquel que tiene D^2H como variable explicativa (Línea roja 13). Si comparamos

los cinco modelos ajustados con el comando `nls`, el mejor es aquel que tiene D^2H como variable explicativa (Línea roja 18). Independientemente del método de ajuste, se puede sacar la conclusión que el modelo de biomasa que usa D^2H como variable explicativa es el mejor.

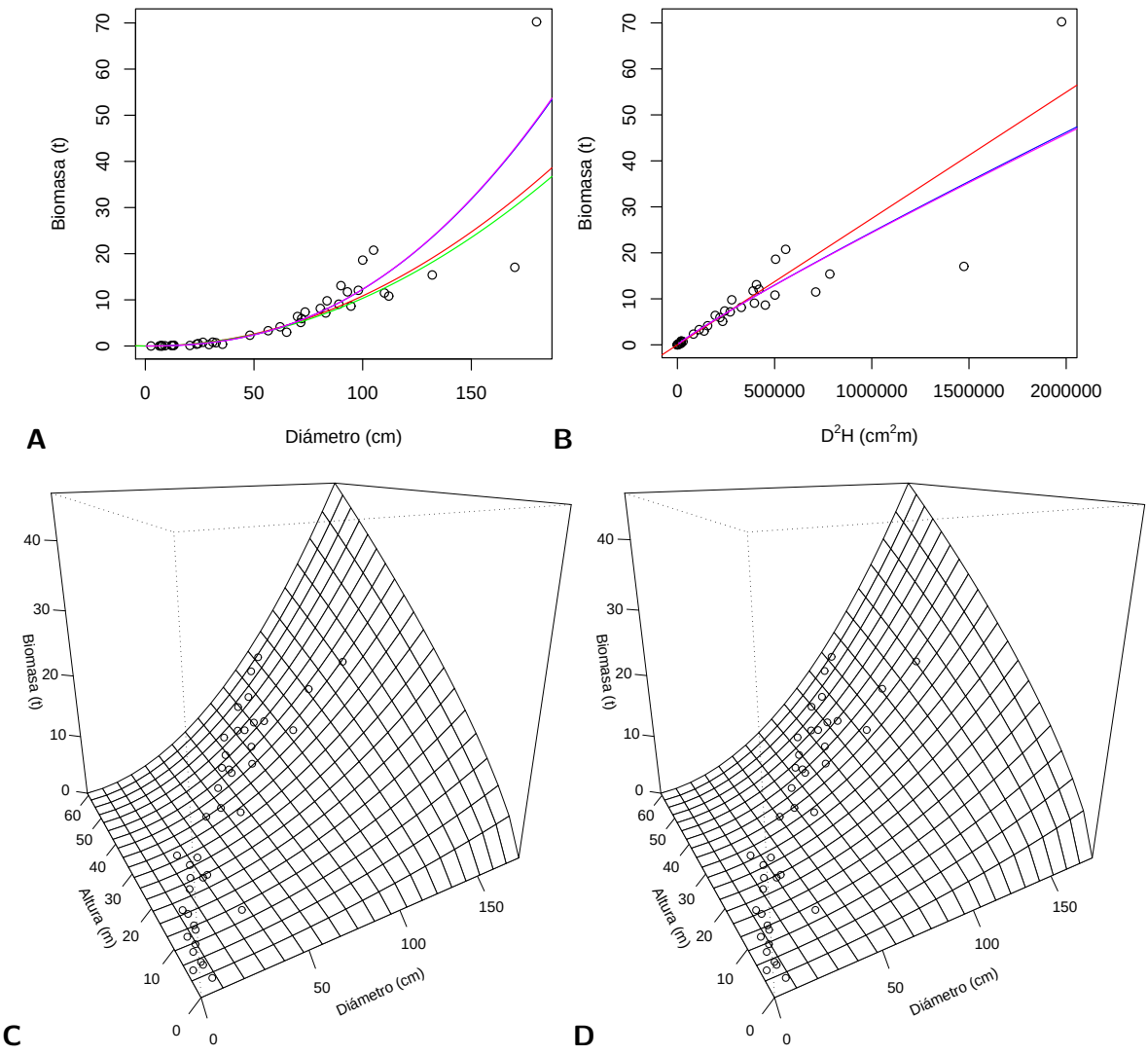


Figura 6.18 – *Predicciones de la biomasa mediante diferentes modelos ajustados a los datos de 42 árboles medidos en Ghana por Henry et al. (2010). Los datos están representados por los puntos: (A) modelos que tienen el diámetro como única entrada, que corresponden a las líneas rojas 12 (rojo), 14 (verde), 17 (azul) y 20 (violeta). (B) modelos que tienen D^2H como única variables explicativa, que corresponden a las líneas rojas 11 (rojo), 13 (verde), 18 (azul) y 21 (violeta). (C) modelo que corresponde a la Línea roja 19. (D) modelo que corresponde a la Línea roja 22.*

Cuadro 6.1 – *Valor del AIC para 10 modelos de biomasa ajustados a los datos de 42 árboles medidos en Ghana por Henry et al. (2010). Estos 10 modelos predicen la biomasa directamente.*

Linea roja	Entrada	Método* ajuste	Comando R	AIC
12	D	MCP	`lm`	76,71133
14	D	MV	`nlme`	83,09157
17	D	MCP	`nls`	24 809,75727
20	D	MV	`nlme`	75,00927
11	D^2H	MCP	`lm`	65,15002
13	D^2H	MV	`nlme`	69,09644
18	D^2H	MCP	`nls`	24 797,53706
21	D^2H	MV	`nlme`	69,24482
19	D, H	MCP	`nls`	24 802,91248
22	D, H	MV	`nlme`	76,80204

*MCP = mínimos cuadrados ponderados, MV = máxima verosimilitud

28

Selección de modelos con $\ln(B)$ como variable de respuesta

Los modelos siguientes con $\ln(B)$ como variable de respuesta fueron ajustados:

- Línea roja 7 ou 9: $\ln(B) = -8{,}42722 + 2{,}36104\ln(D)$
- Línea roja 8: $\ln(B) = -8{,}99427 + 0{,}87238\ln(D^2H)$
- Línea roja 10: $\ln(B) = -8{,}9050 + 1{,}8654\ln(D) + 0{,}7083\ln(H)$
- Línea roja 24: $\ln(B) = -6{,}50202 + 0{,}23756[\ln(D)]^2 + 1{,}01874\ln(H)$

Todos estos modelos fueron ajustados usando regresión lineal por el método de mínimos cuadrados ordinarios. El trazado de las predicciones en escala logarítmica para un modelo `m` que depende solamente del diámetro se obtiene mediante el comando siguiente:

```
with(dat,plot(dbh,Btot,xlab="Diámetro (cm)",ylab="Biomasa (t)",log="xy"))
D <- 10^par("usr")[1:2]
lines(D,exp(predict(m1,newdata=data.frame(dbh=D))))
```

Para un modelo dependiente al mismo tiempo del diámetro y de la altura, el comando para un gráfico en escala logarítmica será:

```
D <- exp(seq(log(1),log(180),length=20))
H <- exp(seq(log(1),log(61),length=20))
B <- matrix(predict(m,newdata=expand.grid(dbh=D,haut=H)),length(D))
M <- persp(log(D),log(H),B,xlab="log(Diámetro) (cm)",ylab="log(Altura) (m)",zlab=
"log(Biomasa) (t)",ticktype="detailed")
points(trans3d(log(dat$dbh),log(dat$haut),log(dat$Btot),M))
```

La Figura 6.19 muestra la predicción de $\ln(B)$ según los cuatro modelos. Dado un modelo ajustado `m`, su AIC se calcula mediante el comando:

```
AIC(m)
```

El Cuadro 6.2 da el AIC para los cuatro modelos. Al haber sido ajustados con la misma instrucción `lm`, los valores del AIC son directamente comparables. El mejor modelo, es decir aquel con el AIC menor, resulta ser el cuarto (modelo dela Línea roja 24). Tomaremos nota también de que la clasificación de los modelos según el AIC es completamente coherente con las pruebas de modelos anidados realizadas anteriormente líneas rojas 25 y 26).

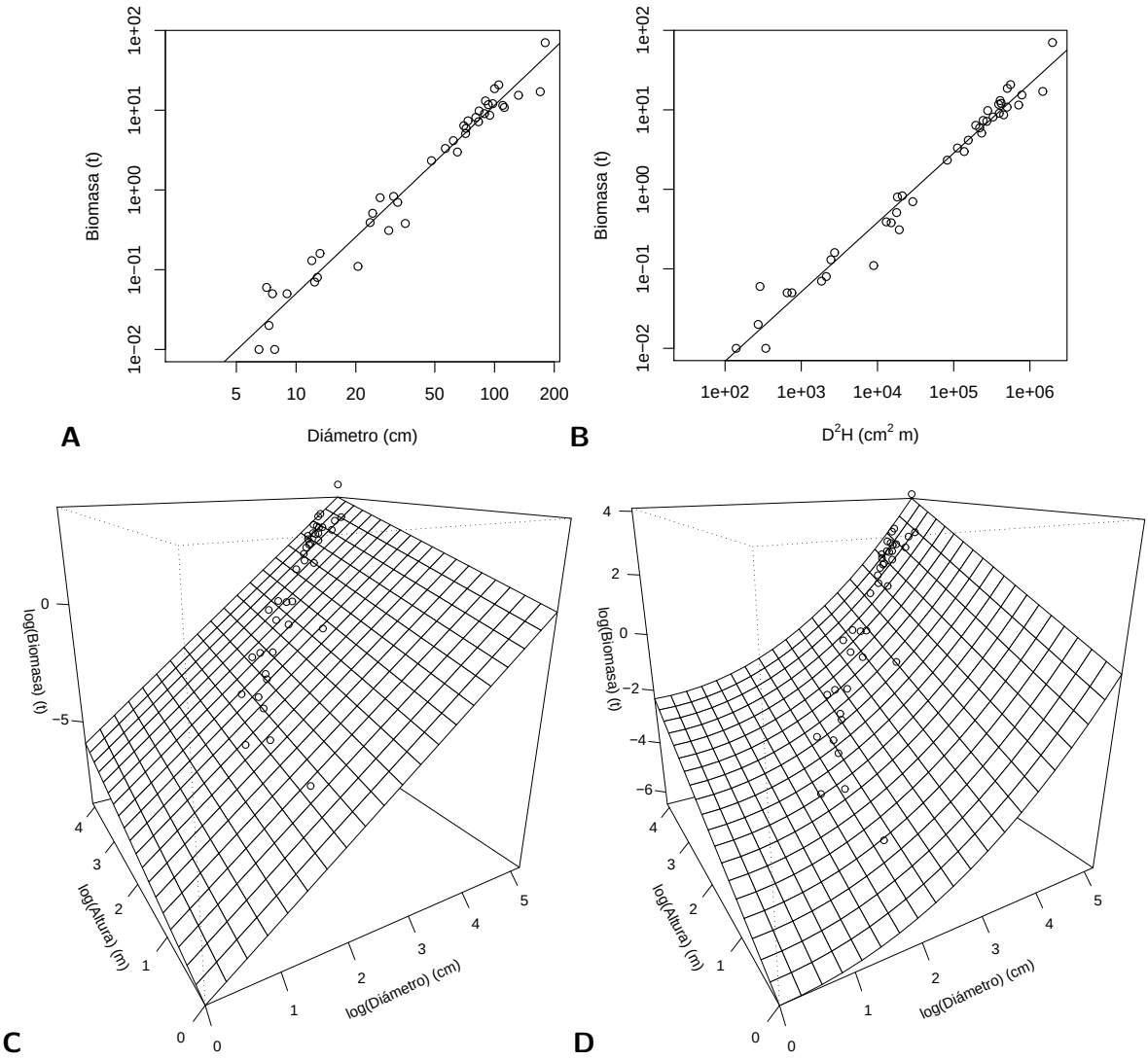


Figura 6.19 – *Predicciones de la biomasa mediante diferentes modelos ajustados a los datos de 42 árboles medidos en Ghana por Henry et al. (2010). Los datos están representados por los puntos. (A) modelo dela Línea roja 7. (B) modelo dela Línea roja 8. (C) modelo dela Línea roja 10. (D) modelo dela Línea roja 24.*

Cuadro 6.2 – *Valor del AIC para cuatro modelos de biomasa ajustadas a los datos de los 42 árboles medidos en Ghana por Henry et al. (2010). Esos cuatro modelos predicen el logaritmo de la biomasa y están todos ajustados según una regresión lineal por el método de los mínimos cuadrados ordinarios (MCO).*

línea roja	Entrada	Método ajuste	Comando R	AIC
7	D	MCO	`lm`	56,97923
8	D^2H	MCO	`lm`	46,87780
10	D, H	MCO	`lm`	48,21367
24	D, H	MCO	`lm`	45,96998

Modelos con variables de respuesta diferentes

El caso más general es cuando se quiere comparar dos modelos que no tienen la misma variable de respuesta porque una está transformada a partir de la otra. Por ejemplo, los modelos $B = aD^b + \varepsilon$ e $\ln(B) = a + b\ln(D) + \varepsilon$ predicen ambos la biomasa pero la variable de respuesta es B en un caso e $\ln(B)$ en el otro. En esta situación, no se pueden usar los criterios de información (AIC o BIC) para comparar los modelos. Sin embargo, en este caso puede usarse el índice de Furnival (1961) para comparar los modelos. Aquél con el valor menor del índice de Furnival será considerado como el mejor (Parresol, 1999).

El índice de Furnival está definido únicamente para un modelo cuyo error residual ε tenga una varianza que asumimos es constante: $\mathrm{Var}(\varepsilon) = \sigma^2$. En cambio no impone ninguna restricción en la forma de la transformación de la variable que une la variable de respuesta Y modelada a la variable de interés (volumen o biomasa). Consideremos el caso de un modelo de biomasa (la transposición a una modelo de volumen es inmediata) y sea ψ esta transformación de variable: $Y = \psi(B)$. El índice de Furnival se define mediante:

$$F = \frac{\hat{\sigma}}{\sqrt[n]{\prod_{i=1}^{n} \psi'(B_i)}} = \exp\Big(-\frac{1}{n}\sum_{i=1}^{n} \ln[\psi'(B_i)]\Big)\,\hat{\sigma}$$

donde $\hat{\sigma}$ es la estimación de la desviación estándar residual del modelo ajustado y B_i es la biomasa del i-ésimo árbol medido. Cuando no hay transformación de variables, ψ es la función de identidad y el índice de Furnival F es entonces igual a la desviación estándar residual $\hat{\sigma}$. La transformación de variables más frecuente es la logarítmica: $\psi(B) = \ln(B)$ y $\psi'(B) = 1/B$, en cuyo caso el índice de Furnival es:

$$F_{\ln} = \hat{\sigma}\,\sqrt[n]{\prod_{i=1}^{n} B_i} = \exp\Big(\frac{1}{n}\sum_{i=1}^{n} \ln(B_i)\Big)\,\hat{\sigma}$$

Para las regresiones lineales cuya varianza residual se asume que es proporcional a una potencia de una variable explicativa X_1, un truco permite definir el índice de Furnival. En efecto, la regresión lineal

$$Y = a_0 + a_1X_1 + a_2X_2 + \ldots + a_pX_p + \varepsilon \tag{6.25}$$

con $\mathrm{Var}(\varepsilon) = (kX_1^c)^2$ es estrictamente equivalente a la regresión lineal (cf. pág. 137):

$$Y' = a_0X_1^{-c} + a_1X_1^{1-c} + a_2X_2X_1^{-c} + \ldots + a_pX_pX_1^{-c} + \varepsilon' \tag{6.26}$$

con $Y' = YX_1^{-c}$, $\varepsilon' = \varepsilon X_1^{-c}$ y $\mathrm{Var}(\varepsilon') = k^2$. El índice de Furnival está definido para el modelo (6.26) por tener una varianza residual. Por extensión, se define el índice de Furnival del modelo (6.25) como el índice de Furnival del modelo (6.26). Si $Y = \psi(B)$, entonces $Y' = X_1^{-c}\psi(B)$, de modo que el índice de Furnival es ahora igual a:

$$F = \frac{\hat{k}}{\sqrt[n]{\prod_{i=1}^{n} X_{i1}^{-c}\psi'(B_i)}} = \exp\left(\frac{1}{n}\sum_{i=1}^{n}\{c\ln(X_{i1}) - \ln[\psi'(B_i)]\}\right)\hat{k}$$

Así se demuestra que el índice de Furnival también puede usarse para seleccionar el valor del exponente c en una regresión ponderada (cf. pág.128).

6.3.3. ¿Qué método de ajuste elegir?

Volvamos a la forma de ajustar un modelo de volumen o de biomasa. Con frecuencia, se presentan varias soluciones para ajustar un modelo. Consideremos, por ejemplo, el modelo de biomasa

$$B = a\rho^{b_1}D^{b_2}H^{b_3} + \varepsilon$$

con

$$\varepsilon \sim \mathcal{N}(0,\, kD^c)$$

Este modelo podrá ajustarse como un modelo no lineal (*i*) por el método de mínimo cuadrados ponderados (c fijado a priori) o (*ii*) por el método de máxima verosimilitud (c no fijado a priori). Si aplicamos la transformación logarítmica a los datos, podremos (*iii*) ajustar la regresión múltiple

$$\ln(B) = a' + b_1\ln(\rho) + b_2\ln(D) + b_3\ln(H) + \varepsilon$$

con

$$\varepsilon \sim \mathcal{N}(0,\, \sigma)$$

De este modo, para el mismo modelo que predice la biomasa como una potencia de las variables explicativas, tenemos tres métodos de ajuste. Los métodos (*i*), (*ii*) y (*iii*) se basan en hipótesis diferentes para la estructura de los errores residuales: error aditivo con respecto a B en los casos (*i*) y (*ii*), error multiplicativo con respecto a B en el caso (*iii*). Sin embargo, ambos tipos de error pueden reflejar la heterocedasticidad de los datos, de modo que los métodos de ajuste (*i*), (*ii*) y (*iii*) tienen posibilidades todos ellos de ser válidos.

Como otro ejemplo, consideremos el modelo de biomasa:

$$B = \exp\{a_0 + a_1\ln(D) + a_2[\ln(D)]^2 + a_3[\ln(D)]^3 + a_4\ln(\rho)\} + \varepsilon$$

con

$$\varepsilon \sim \mathcal{N}(0,\, kD^c)$$

Aquí también se podrá (*i*) ajustar un modelo no lineal mediante el método de mínimos cuadrados (especificando c a priori), (*ii*) ajustar un modelo no lineal con el método de máxima verosimilitud (estimando c), o (*iii*) ajustar una regresión múltiple con los datos transformados logarítmicamente:

$$\ln(B) = a_0 + a_1\ln(D) + a_2[\ln(D)]^2 + a_3[\ln(D)]^3 + a_4\ln(\rho) + \varepsilon$$

con

$$\varepsilon \sim \mathcal{N}(0,\, \sigma)$$

Aquí también la estructura de los errores no es la misma en los tres casos pero todos pueden reflejar la heterocedasticidad de la biomasa.

Con mucha frecuencia los distintos métodos de ajuste darán resultados muy parecidos en términos de predicción. Si surgiera una duda sobre el método de ajuste más adecuado, se podrían usar los métodos de selección de modelos para zanjar la cuestión. En la práctica, la elección de un método de ajuste resultará más bien de la importancia que se conceda a las ventajas e inconvenientes respectivos de cada método. La regresión múltiple tiene el inconveniente de imponer restricciones sobre la forma de los residuos y de tener menos flexibilidad en la forma del modelo para la media. Como ventaja, ofrece una expresión explícita de los estimadores de los coeficientes del modelo; no hay riesgo de tener estimaciones erróneas de los coeficientes. El modelo no lineal presenta la ventaja de no plantear ninguna restricción sobre el modelo para la media o para la varianza. Como inconveniente, no hay expresión explícita de los estimadores de parámetros: hay pues un riesgo de tener estimaciones erróneas de los parámetros.

29

Métodos de ajuste del modelo de potencia

Vimos tres formas de ajustar el modelo de potencia $B = aD^b$:

1. con una regresión lineal simple con los datos transformados logarítmicamente (Línea roja 7): $\ln(B) = -8{,}42722 + 2{,}36104 \ln(D)$, sea $B = 2{,}18829 \times 10^{-4} D^{2{,}36104}$ si se aplica "ingenuamente" la transformación exponencial inversa;

2. con una regresión no lineal ponderada (Línea roja 17): $B = 2{,}492 \times 10^{-4} D^{2{,}346}$;

3. con una regresión no lineal con modelo sobre la varianza (Línea roja 20): $B = 2{,}445 \times 10^{-4} D^{2{,}35105}$.

La Figura 6.20 compara las predicciones de estos tres ajustes del mismo modelo, lo que muestra que las diferencias son mínimas, muy por debajo de la precisión de las predicciones, como lo veremos más adelante (§ 7.2).

6.4. Factores de estratificación y agregación

Hasta ahora hemos considerado que el conjunto de datos usado para ajustar un modelo de volumen o de biomasa era homogéneo. En realidad, el conjunto de datos puede ser el resultado de mediciones efectuadas en condiciones diversas o puede resultar de la fusión de varios juegos de datos distintos. En general se utilizan covariables para describir esta heterogeneidad del conjunto de datos. Por ejemplo, una covariable podrá indicar el tipo de bosque en el que se hicieron las mediciones (bosque latifoliado, semicaducifolio, siempre verde, etc.) o el tipo de suelo o el año de la plantación (si se trata de una plantación), etc. Para los conjuntos de datos pluriespecíficos, una covariable muy importante es la especie del árbol. En un primer momento, todas las covariables que pueden explicar la heterogeneidad de un conjunto de datos serán consideradas como variables cualitativas (o factores). Las categorías de estos factores definen los estratos. Un conjunto de datos bien constituido habrá tenido que dar lugar a muestreos en función de los estratos identificados previamente (cf. § 2.2.3). ¿Cómo tomar en cuenta estas covariables cualitativas al construir un modelo

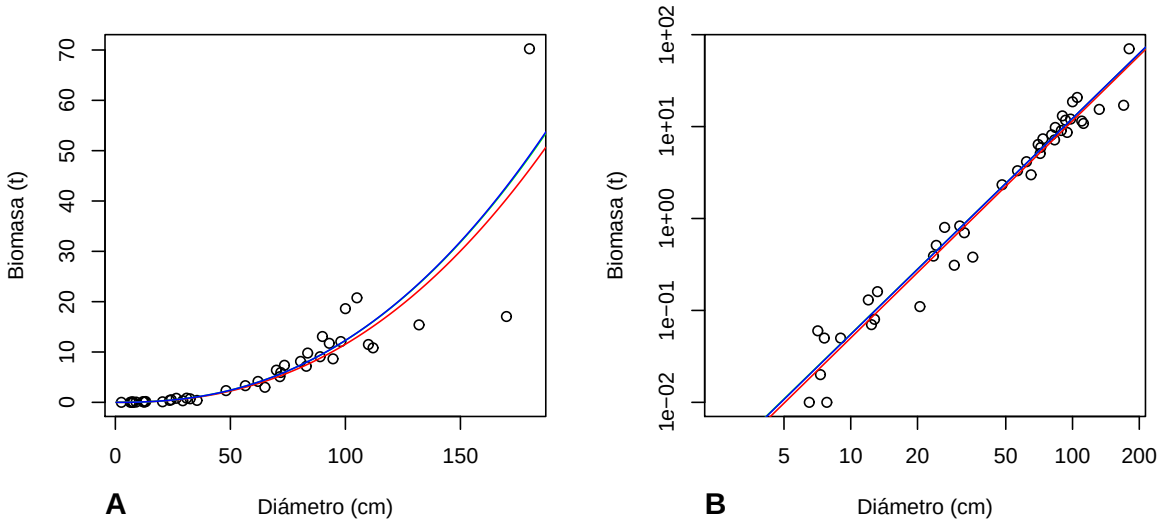


Figura 6.20 – *Predicciones de la biomasa para el mismo modelo de potencia ajustada de tres formas diferentes a los datos de 42 árboles medidos en Ghana por Henry et al. (2010). Los datos están representados por los puntos. En rojo, el ajuste mediante la regresión lineal con los datos transformados logarítmicamente (Línea roja 7). En verde (prácticamente superpuesto con el azul), el ajuste por la regresión no lineal ponderada (Línea roja 17). En azul, el ajuste por regresión no lineal con modelo de varianza (Línea roja 20). (A) Sin transformaciones de los datos. (B) En escala logarítmica.*

de volumen o de biomasa? ¿Es válido analizar el juego de datos en su totalidad o bien hay que analizar los subconjuntos de datos correspondientes a cada estrato por separado? Éstas son las preguntas que vamos a abordar ahora (§ 6.4.1).

Además, las mediciones de biomasa se hacen por separado para cada parte del árbol (cf. Capítulo 3). Para cada árbol de la muestra, además de la estimación de su biomasa total, hay una estimación de su biomasa foliar, de la biomasa de su tronco, de sus ramas gruesas, de sus ramillas, etc. ¿Cómo tener en cuenta estos diferentes compartimentos al construir los modelos de biomasa? También abordaremos esta cuestión luego (§ 6.4.2).

6.4.1. Estratificación de los datos

Consideremos en adelante que hay covariables cualitativas que estratifican el conjunto de datos según S estratos. Cada estrato corresponde a un cruce de modalidades de covariables cualitativas (en un contexto de diseños experimentales hablaríamos de *tratamiento* más que de estrato) y no consideraremos cada una de las covariables cualitativas por separado. Por ejemplo, si hay una covariable que indica el tipo de bosque con tres modalidades (supongamos, bosque latifoliado, bosque semicaducifolio y bosque siempre verde) y otra covariable que indica el tipo de suelo con tres modalidades (digamos, arenoso, arcilloso y limoso), el cruce de ambas da $S = 3 \times 3 = 9$ estratos (bosque latifoliado en suelo arenoso, bosque latifoliado en suelo arcilloso, etc.). No intentaremos analizar el efecto del tipo de bosque por separado ni tampoco el efecto del tipo de suelo por separado. Además, si ciertas combinaciones de modalidades de covariables no están representadas en el conjunto de datos, el número de estratos disminuirá en consecuencia. Por ejemplo, si no hay bosque perenne en terrenos limosos, el número de estratos será $S = 8$ en vez de 9.

Frente a una estratificación del conjunto de datos, una estrategia posible consistiría en

ajustar un modelo por separado para cada estrato. En el caso de la regresión múltiple, eso se escribiría:

$$Y_s = a_{0s} + a_{1s}X_{1s} + a_{2s}X_{2s} + \ldots + a_{ps}X_{ps} + \varepsilon_s$$

con

$$\varepsilon_s \underset{\text{i.i.d.}}{\sim} \mathcal{N}(0,\, \sigma_s)$$

donde $(Y_s,\, X_{1s},\, \ldots,\, X_{ps})$ designa una observación relativa al estrato s, para $s = 1,\, \ldots,\, S$. Hay entonces $S \times (p+1)$ coeficientes por estimar. Una estrategia alternativa consiste en analizar el conjunto de datos en su globalidad, ajustando un modelo de tipo:

$$Y_s = a_{0s} + a_{1s}X_{1s} + a_{2s}X_{2s} + \ldots + a_{ps}X_{ps} + \varepsilon \tag{6.27}$$

con

$$\varepsilon \underset{\text{i.i.d.}}{\sim} \mathcal{N}(0,\, \sigma)$$

La forma de escribir el modelo sólo se diferencia en la estructura del error. Este tipo de modelo se llama análisis de *covarianza*. Éste parte del supuesto que todos los residuos tienen la misma varianza, no sólo dentro de cada estrato sino también entre un estrato y otro. El análisis de covarianza permite demostrar si hay un efecto del estrato en la variable de respuesta, único o en interacción con cada una de las variables explicativas $X_1,\, \ldots,\, X_p$. Someter a prueba el efecto principal de la estratificación equivale a demostrar la hipótesis nula $a_{01} = a_{02} = \ldots = a_{0S}$. El estadístico de prueba es una razón entre los cuadrados medios que, bajo la hipótesis nula, sigue una distribución de Fisher. Someter a prueba el efecto de la interacción entre la estratificación y la j-ésima variable explicativa equivale a demostrar la hipótesis nula $a_{j1} = a_{j2} = \ldots = a_{jS}$. Al igual que antes, la estadística de prueba es una razón entre los cuadrados medios que, bajo la hipótesis nula, sigue una distribución de Fisher.

El interés de someter a prueba estos efectos es que, cada vez que uno de ellos resulta ser no significativo, se pueden remplazar los S coeficientes $a_{j1},\, a_{j2},\, \ldots,\, a_{jS}$ por estimar por un único coeficiente común a_j. Imaginemos, por ejemplo, que en el análisis de covarianza (6.27), el efecto principal del estrato no sea significativo y que tampoco lo sea la interacción entre el estrato y las p' primeras variables explicativas (con $p' < p$). En ese caso, el modelo por ajustar se escribe:

$$Y_s = a_0 + a_1X_{1s} + \ldots + a_{p'}X_{p's} + a_{p'+1,s}X_{p'+1,s} + \ldots + a_{ps}X_{ps} + \varepsilon$$

con $\varepsilon \sim \mathcal{N}(0,\, \sigma)$. Este modelo incluye "solo" a $p' + 1 + (p - p')S$ coeficientes por estimar, en vez de los $(p+1)S$ coeficientes, si ajustáramos un modelo por separado para cada estrato. Al servir el conjunto de observaciones para estimar los coeficientes comunes $a_0,\, \ldots,\, a_{p'}$, éstos se estimarán con más precisión que si hubiéramos ajustado un modelo separado para cada estrato.

Este principio de análisis de covarianza se aplica directamente también al caso de un modelo no lineal. Allí también se podrá comprobar si los coeficientes son o no significativamente diferentes entre los estratos para, eventualmente, estimar un coeficiente común a todos los estratos.

30

Modelo específico de biomasa

En la Línea roja 8, ajustamos por regresión lineal simple a los datos transformados logarítmicamente un modelo de potencia que usa D^2H como variable explicativa: $\ln(B) =$

$a + b\ln(D^2H)$. Ahora podemos integrar la información sobre la especie en este modelo para probar si los coeficientes a y b difieren de una especie a otra. El modelo corresponde a un análisis de covarianza:

$$\ln(B_s) = a_s + b_s \ln(D_s^2 H_s) + \varepsilon$$

con

$$\mathrm{Var}(\varepsilon) = \sigma^2$$

donde el índice s designa la especie. El ajuste de este modelo se logra con el comando:

```
m <- lm(log(Btot)~especie*I(log(dbh^2*haut)),data=dat[dat$Btot>0,])
```

Para probar si los coeficientes a y b difieren de una especie a otra, se usa el comando

```
anova(m)
```

que da:

```
                           Df  Sum Sq Mean Sq  F value    Pr(>F)
especie                    15 117.667   7.844  98.4396 1.647e-13 ***
I(log(dbh^2*haut))          1 112.689 112.689 1414.1228  <2.2e-16 ***
especie:I(log(dbh^2*haut))  7   0.942   0.135    1.6879    0.1785
Residuals                  17   1.355   0.080
```

El primer renglón del Cuadro verifica si hay un efecto especie, es decir, si la intersección a_s difiere de una especia a otra. La hipótesis nula de esta prueba es que no hay diferencia entre las especies: $a_1 = a_2 = \ldots = a_S$, donde $S = 16$ es el número de especies. Es estadístico de prueba está dado en la columna"F value". El p-value de la prueba es inferior aquí a 5 %, así que podemos concluir que la intersección del modelo es significativamente diferente entre especies. El segundo renglón del Cuadro comprueba si hay un efecto de la variable D^2H, es decir, si la pendiente media asociada a dicha variable es significativamente diferente de cero. El tercer renglón del Cuadro verifica si la interacción pendiente-especie es significativa, es decir, si la pendiente b_s difiere de una especie a otra. La hipótesis nula es que no hay diferencias entre especies: $b_1 = b_2 = \ldots = b_S$. El p-value de 0,1785 es pues superior a 5 %: por ende, no hay diferencia significativa de pendiente entre las especies.

Por lo tanto tenemos que ajustar el modelo siguiente:

$$\ln(B_s) = a_s + b\ln(D_s^2 H_s) + \varepsilon \tag{6.28}$$

que considera que la pendiente b es la misma para todas las especies. El comando es:

```
m <- lm(log(Btot)~especie+I(log(dbh^2*haut)),data=dat[dat$Btot>0,])
anova(m)
```

y da:

```
                  Df  Sum Sq Mean Sq F value    Pr(>F)
especie           15 117.667   7.844   81.99  <2.2e-16 ***
I(log(dbh^2*haut))  1 112.689 112.689 1177.81  <2.2e-16 ***
Residuals          24   2.296   0.096
```

Los coeficientes del modelo se obtienen mediante el comando:

```
summary(m)
```

lo que da:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	-9.00359	0.45144	-19.944	<2e-16	***
especieAubrevillea kerstingii	-0.54634	0.43784	-1.248	0.2241	
especieCecropia peltata	-0.77688	0.36261	-2.142	0.0425	*
especieCeiba pentandra	-0.70841	0.38048	-1.862	0.0749	.
especieCola nitida	-0.46428	0.44476	-1.044	0.3069	
especieDaniellia thurifera	0.04685	0.46413	0.101	0.9204	
especieDialium aubrevilliei	-0.15626	0.43757	-0.357	0.7241	
especieDrypetes chevalieri	0.04953	0.45395	0.109	0.9140	
especieGarcinia epunctata	1.09645	0.47318	2.317	0.0293	*
especieGuarea cedrata	-0.45255	0.38460	-1.177	0.2509	
especieHeritiera utilis	-0.26865	0.32663	-0.822	0.4189	
especieNauclea diderrichii	-0.55464	0.35759	-1.551	0.1340	
especieNesogordonia papaverifera	-0.47817	0.44335	-1.079	0.2915	
especiePiptadeniastrum africanum	-0.17956	0.35718	-0.503	0.6197	
especieStrombosia glaucescens	0.06333	0.39597	0.160	0.8743	
especieTieghemella heckelii	-0.09104	0.33908	-0.268	0.7906	
I(log(dbh^2*haut))	0.89985	0.02622	34.319	<2e-16	***

El último renglón de este Cuadro da el valor de la pendiente: $b = 0{,}89985$. . Los renglones anteriores dan las intersecciones para las 16 especies. Por convención, el software R actúa del modo siguiente para especificar dichos valores: el primer renglón del Cuadro da la intersección para la primera especie según el orden alfabético. Al ser la primera especie en ese orden *Afzelia bella*, la intersección para *Afzelia bella* es $a_1 = -9{,}00359$. Los renglones siguientes dan la *diferencia* $a_s - a_1$ entre la intersección para la especie indicada y la intersección de *Afzelia bella*. Por tanto, la intersección para *Aubrevillea kerstingii* es: $a_2 = a_1 - 0{,}54634 = -9{,}00359 - 0{,}54634 = -9{,}54993$. En definitiva, la expresión específica del modelo es:

$$\ln(B) = 0{,}89985\ln(D^2H) - \begin{cases} 9{,}00359 & \text{para } \textit{Afzelia bella} \\ 9{,}54993 & \text{para } \textit{Aubrevillea kerstingii} \\ 9{,}78047 & \text{para } \textit{Cecropia peltata} \\ 9{,}71200 & \text{para } \textit{Ceiba pentandra} \\ 9{,}46786 & \text{para } \textit{Cola nitida} \\ 8{,}95674 & \text{para } \textit{Daniellia thurifera} \\ 9{,}15985 & \text{para } \textit{Dialium aubrevilliei} \\ 8{,}95406 & \text{para } \textit{Drypetes chevalieri} \\ 7{,}90713 & \text{para } \textit{Garcinia epunctata} \\ 9{,}45614 & \text{para } \textit{Guarea cedrata} \\ 9{,}27223 & \text{para } \textit{Heritiera utilis} \\ 9{,}55823 & \text{para } \textit{Nauclea diderrichii} \\ 9{,}48176 & \text{para } \textit{Nesogordonia papaverifera} \\ 9{,}18315 & \text{para } \textit{Piptadeniastrum africanum} \\ 8{,}94026 & \text{para } \textit{Strombosia glaucescens} \\ 9{,}09462 & \text{para } \textit{Tieghemella heckelii} \end{cases}$$

A diámetro y altura iguales, la especie que tiene la mayor biomasa es *Garcinia epunctata* mientras que aquella con la biomasa menor es *Cecropia peltata*. La desviación estándar residual del modelo es $\hat{\sigma} = 0{,}3093$ y $R^2 = 0{,}9901$.

Caso de una covariable numérica

Hasta ahora sólo hemos considerado que las covariables que definen la estratificación eran factores cualitativos. En ciertos casos, dichas covariables pueden ser también interpretadas como variables numéricas. Tomemos como ejemplo un modelo de biomasa para plantaciones (Saint-André *et al.*, 2005). El año en que se hizo la plantación (o, lo que viene a ser lo mismo, la edad de los árboles) podría usarse como covariable de estratificación. Ese año o esa edad pueden verse indiferentemente como variables cualitativas (cohortes de árboles con la misma edad) o como variables numéricas. Más generalmente, toda variable numérica puede ser vista como una variable cualitativa si se la subdivide en clases. En el caso de la edad, podríamos pues considerar las plantaciones entre 0 y 5 años como un estrato, aquellas entre 5 y 10 años como otro, las plantaciones entre 10 y 20 años como un tercer estrato, etc. La ventaja de subdividir una covariable numérica Z en clases y considerarla como una variable cualitativa es que eso permite modelar la relación entre Z y la variable de respuesta Y sin imponer a priori la forma de esta relación. En el extremo opuesto, cuando consideramos Z como una variable numérica, estamos obligados a plantear a priori cierta forma de relación entre Y y Z (una relación lineal, polinomial, exponencial o de potencia, etc.). El inconveniente de subdividir Z en clases y considerar esta covariable como cualitativa es que la subdivisión introduce un elemento de arbitrariedad. Además el modelo de covarianza que usa las clases de Z (covariables cualitativas) tendrá generalmente más parámetros por estimar que el modelo que considera Z como una covariable numérica.

En el modelado se suele jugar con la dualidad de interpretación de las variables numéricas. Cuando una covariable Z es numérica (como la edad de los árboles), recomendamos en ese caso proceder en dos etapas (como se explicó en el § 5.1.1):

1. considerar Z como una variable cualitativa (después de subdividirla en clases, de ser necesario) y ajustar un modelo de covarianza, lo que permitirá visualizar la forma de la relación entre Z y los coeficientes del modelo;

2. modelar esta relación mediante una expresión adecuada y volver al ajuste de un modelo lineal o no lineal, considerando Z como una variable numérica.

Para retomar el ejemplo de la edad de los árboles en la plantación: supongamos que la edad de Z fue subdividida en S clases de edad. La primera etapa consistiría normalmente en un análisis de covarianza (suponiendo que el modelo haya podido ser linealizado):

$$Y_s = a_{0s} + a_{1s}X_{1s} + a_{2s}X_{2s} + \ldots + a_{ps}X_{ps} + \varepsilon$$

con $s = 1, \ldots, S$. Siendo Z_s la edad mediana de la clase de edad s. Luego graficaríamos la nube de puntos de a_{0s} en función de Z_s, la nube de puntos de a_{1s} en función de Z_s, ..., la nube de puntos de a_{ps} en función de Z_s. Para cada nube de puntos, buscaríamos la forma de la relación que se ajusta a esa nube de puntos. Imaginemos que a_{0s} varía en forma lineal en función de Z_s, que a_{1s} varía en forma exponencial en función de Z_s, que a_{2s} varía en forma de potencia en función de Z_s, y que los coeficientes a_{3s} a a_{ps} no varían en función de Z_s (lo que además puede demostrarse formalmente). En este caso particular tendríamos que ajustar en una segunda etapa el modelo no lineal siguiente:

$$Y = \underbrace{b_0 + b_1 Z}_{a_{0s}} + \underbrace{b_2 \exp(-b_3 Z)}_{a_{1s}} X_1 + \underbrace{b_4 Z^{b_5}}_{a_{2s}} X_2 + a_3 X_3 + \ldots + a_p X_p + \varepsilon$$

donde la edad Z se considera ahora como una variable numérica. Un modelo de este tipo con una covariable explicativa numérica se llama modelo *parametrado* (por la edad, en este caso).

Las covariables *ordinales* merecen una observación particular. Una variable ordinal es una variable cualitativa que define un orden. El mes del año es una variable cualitativa que establece un ordencronológico. El tipo de suelo a lo largo de un gradiente de fertilidad de suelos también es una variable ordinal. Las variables ordinales se tratan generalmente como si fueran variables cualitativas de pleno derecho pero, en ese caso, se pierde la información de orden que aportan. Una alternativa consiste en numerar las modalidades ordenadas de la variable ordinal por valores enteros y considerar luego la variable ordinal como una variable numérica. Por ejemplo, en el caso de los meses del año, se podría poner enero = 1, febrero = 2, etc. Este enfoque sólo tiene sentido si las desviaciones entre los enteros reflejan bien las desviaciones entre las modalidades de la variable ordinal. Por ejemplo, si pusimos 1 = enero 2011 hasta 12 = diciembre 2011, pondremos 1 = enero 2012 si la respuesta es estacional cíclica, mientras que pondremos 13 = enero 2012 si la respuesta presenta una tendencia continua. En el caso de los tres tipos de suelo a lo largo de un gradiente de fertilidad, ponderemos 1 = el suelo más pobre, 2 = el suelo de fertilidad intermedia y 3 = el suelo más rico, si pensamos que la diferencia de fertilidad entre ambos suelos induce una respuesta proporcional a dicha diferencia, pero pondremos 1 = el suelo más pobre, 4 = el suelo de fertilidad intermedia y 9 = el suelo más rico, si pensamos que la respuesta es proporcional al cuadrado de la diferencia de fertilidad.

Caso particular de las especies

En el caso de conjuntos de datos pluriespecíficos, la especie es una covariable de estratificación que merece una atención especial. Si el conjunto de datos conlleva pocas especies (menos de 10 aproximadamente) y que hay suficientes observaciones por especie (cf. § 2.2.1), ésta podría considerarse como una covariable de estratificación cualquiera. En ese caso tendremos que desglosar el modelo en S modelos específicos o reagruparlos en función de la similitud alométrica de las especies.

Cuando el conjunto de datos contiene muchas especies o si algunas especies tienen pocas observaciones, es difícil tratar la especie como una covariable de estratificación. En este caso una solución sería usar los *rasgos funcionales* de las especies. Dichos rasgos se definen aquí, en forma un poco imprecisamente, como variables numéricas que caracterizan la especie (Díaz & Cabido, 1997; Rösch *et al.*, 1997; Lavorel & Garnier, 2002; véanse Violle *et al.*, 2007 una definición más rigurosa). El rasgo más usado en el caso de los modelos de biomasa es la densidad de la madera. Si decidimos usar rasgos funcionales para representar las especies, éstos actúan como variables explicativas del modelo de igual modo que las variables explicativas que caracterizan el árbol, como su diámetro o su altura. Un modelo de potencia monoespecífico para biomasa, con una entrada (con respecto al diámetro), que en su forma linealizada se escribe:

$$\ln(B) = a_0 + a_1 \ln(D) + \varepsilon$$

en el caso pluriespecífico se convertirá en modelo de biomasa de dos entradas:

$$\ln(B) = a_0 + a_1 \ln(D) + a_2 \ln(\rho) + \varepsilon$$

si decidimos usar la densidad de la madera ρ para representar el efecto específico.

31

Modelo de biomasa que depende de la densidad específica de la madera

En la Línea roja 30, la información sobre la especie se tuvo en cuenta en el modelo $\ln(B) = a + b \ln(D^2 H)$ mediante una covariable cualitativa. Ahora podemos tratar de captar

esa información a través de la densidad específica de la madera ρ. El modelo ajustado es pues:

$$\ln(B) = a_0 + a_1 \ln(D^2 H) + a_2 \ln(\rho) + \varepsilon \tag{6.29}$$

con

$$\text{Var}(\varepsilon) = \sigma^2$$

Como la densidad de la madera se midió en el conjunto de datos para cada individuo y ahora hay que comenzar por calcular la densidad media de la madera para cada especie:

```
dm <- tapply(dat$dens,dat$especie,mean)
dat <- cbind(dat,dmoy=dm[as.character(dat$especie)])
```

El conjunto de datos `dat` contiene ahora una variable adicional `dmoy` que da la densidad específica de la madera. El modelo se ajusta mediante el comando:

```
m <- lm(log(Btot)~I(log(dbh^2*haut))+I(log(dmoy)),data=dat[dat$Btot>0,])
summary(m)
```

lo que da:

```
                   Estimate  Std. Error  t value  Pr(>|t|)
 (Intercept)       -8.38900     0.26452  -31.714    <2e-16  ***
 I(log(dbh^2*haut))  0.85715     0.02031   42.205    <2e-16  ***
 I(log(dmoy))        0.72864     0.17720    4.112  0.000202  ***
```

con una desviación estándar residual de 0,3442 y $R^2 = 0{,}9806$. El modelo se escribe: $\ln(B) = -8{,}38900 + 0{,}85715 \ln(D^2 H) + 0{,}72864 \ln(\rho)$. ¿Es mejor tener en cuenta la especie por medio de la densidad de la madera como acabamos de hacerlo o bien construir modelos específicos como lo habíamos hecho con la Línea roja 30? Para responder a esta pregunta, podemos comparar el modelo (6.28) al (6.29) usando el AIC:

```
AIC(m)
```

lo que da AIC = 34,17859 para el modelo específico (6.28) y AIC = 33,78733 para el modelo (6.29) que utiliza la densidad de la madera. Es preferible usar esta última opción, aunque la diferencia de AIC es pequeña.

Para tener en cuenta las variaciones de densidad de la madera dentro de un mismo árbol, es posible analizar las variaciones inter e intraespecíficas más que usar una densidad media basada en la hipótesis de que la densidad de la madera es la misma en la médula que en la corteza o desde la parte baja hacia la parte alta de los árboles (véase el Capítulo 1). La densidad de la madera puede modelarse tomando en cuenta factores como la especie, el grupo funcional, la dimensión del árbol, la posición radial y vertical en el árbol. Se puede efectuar primera comparación usando un análisis de varianza de Friedman, luego la Prueba HSD (Diferencia Honestamente Significativa) de Tukey. Éstas permiten distinguir las variables que influyen más en la densidad de la madera. A continuación podemos modelar usando dichas variables (Henry *et al.*, 2010).

32

Modelo de biomasa que depende de la densidad individual de la madera

En la Línea roja 31, la densidad de la madera ρ se definió a nivel de la especie calculando la media de las densidades individuales para los árboles de una misma especie. Ajustemos

ahora un modelo de biomasa basado en la medición individual de la densidad de la madera para tener en cuenta la variabilidad entre individuos de la densidad dentro de la especie. El modelo ajustado es:

$$\ln(B) = a_0 + a_1 \ln(D) + a_2 \ln(\rho) + \varepsilon$$

con

$$\mathrm{Var}(\varepsilon) = \sigma^2$$

donde ρ es aquí, a diferencia de lo que ocurre en la Línea roja 31, la medición *individual* de la densidad de la madera. El modelo se ajusta mediante el comando:

```
m <- lm(log(Btot)~I(log(dbh))+I(log(dens)),data=dat[dat$Btot>0,])
summary(m)
```

lo que da:

```
               Estimate  Std. Error  t value  Pr(>|t|)
 (Intercept)   -7.76644     0.20618  -37.668    <2e-16  ***
 I(log(dbh))    2.35272     0.04812   48.889    <2e-16  ***
 I(log(dens))   1.00717     0.14053    7.167  1.46e-08  ***
```

con una desviación estándar residual de 0,3052 y $R^2 = 0{,}9848$. El modelo se escribe: $\ln(B) = -7{,}76644 + 2{,}35272 \ln(D) + 1{,}00717 \ln(\rho)$. Según este modelo, la biomasa depende de la densidad individual por el término $\rho^{1{,}00717}$, es decir, prácticamente ρ. En comparación, el modelo (6.29) dependía de la densidad específica de la madera por el término $\rho^{0{,}72864}$. Desde un punto de vista biológico, el exponente 1,00717 es más satisfactorio que el exponente 0,72864 puesto que significa que la biomasa es el producto de un volumen (que depende únicamente de las dimensiones del árbol) y de una densidad. La diferencia entre ambos valores del exponente puede atribuirse a las variaciones en la densidad de la madera entre los individuos de una especie. Sin embargo, el modelo basado en la densidad individual de la madera no tiene ninguna utilidad práctica porque implica que haría falta medir la densidad de la madera de todo árbol del que quisiéramos predecir la biomasa.

6.4.2. Partes del árbol

La biomasa de los árboles se pesa por separado para cada compartimento (tocón, tronco, ramas gruesas, ramillas, follaje, etc.). La biomasa epigea es la suma de todas esas partes. El procedimiento que hemos presentado para ajustar un modelo podría seguirse para cada compartimento por separado. De ese modo construiríamos un modelo para la biomasa foliar, uno para la biomasa de las ramas gruesas, etc. Esta forma de proceder integra la estratificación del conjunto de datos. Por tanto, ajustaremos primero un modelo para cada compartimento y cada estrato; luego, en función de las diferencias encontradas entre estratos, podremos agregar los estratos y/o parametrar el modelo por cada compartimento para todos los estratos. Sin embargo, esto no termina allí. Se pueden seguir integrando los datos para orientarse hacia un número menor de modelos más integradores.

Aditividad de los compartimentos

Al ser la biomasa epigea total la suma de las biomasas de las partes, se podría pensar que el mejor modelo para predecir la biomasa epigea es la suma de los modelos que predicen la biomasa de cada compartimento. En realidad, debido a las correlaciones que existen entre las biomasas de las distintas partes, no es así (Cunia & Briggs, 1984, 1985a; Parresol, 1999). Además, ciertas familias de modelos no son estables respecto a la adición. Es lo que ocurre,

en particular, con los modelos de potencia: la suma de dos funciones de potencia no es una función de potencia. Si hemos ajustado un modelo de potencia para cada parte del árbol:

$$\begin{aligned} B^{\text{tocón}} &= a_1 D^{b_1} \\ B^{\text{tronco}} &= a_2 D^{b_2} \\ B^{\text{ramas grandes}} &= a_3 D^{b_3} \\ B^{\text{ramillas}} &= a_4 D^{b_4} \\ B^{\text{follaje}} &= a_5 D^{b_5} \end{aligned}$$

la suma $B^{\text{apical}} = B^{\text{tocón}} + B^{\text{tronco}} + B^{\text{ramas grandes}} + B^{\text{ramillas}} + B^{\text{follaje}} = \sum_{m=1}^{5} a_m \, D^{b_m}$ no es una función de potencia del diámetro. Los modelos polinomiales, por el contrario, son estables respecto a la adición.

Ajuste de un modelo multivariado

Para tener en cuenta las correlaciones que existen entre las biomasas de sus compartimentos, se pueden ajustar simultáneamente los modelos relativos a las distintas partes del árbol en vez de hacerlo por separado. Esta última etapa en la integración del modelo necesita una redefinición de la variable de respuesta. Como queremos predecir simultáneamente las biomasas de las distintas partes, ya no se trata de una variable de respuesta sino de un *vector* de respuesta $\mathbf{Y}$. La longitud de dicho vector es igual al número M de compartimentos. Por ejemplo, si la variable de respuesta es la biomasa,

$$\mathbf{Y} = \begin{bmatrix} B^{\text{apical}} \\ B^{\text{tocón}} \\ B^{\text{tronco}} \\ B^{\text{ramas grandes}} \\ B^{\text{ramillas}} \\ B^{\text{follaje}} \end{bmatrix}$$

Si la variable de respuesta es el logaritmo de la biomasa,

$$\mathbf{Y} = \begin{bmatrix} \ln(B^{\text{apical}}) \\ \ln(B^{\text{tocón}}) \\ \ln(B^{\text{tronco}}) \\ \ln(B^{\text{ramas grandes}}) \\ \ln(B^{\text{ramillas}}) \\ \ln(B^{\text{follaje}}) \end{bmatrix}$$

Supongamos que Y_m la variable de respuesta del m-ésimo compartimento (con $m = 1, \ldots, M$). Sin pérdida de generalidad podemos considerar que todos los compartimentos tienen el mismo conjunto $X_1, X_2, \ldots, X_p$ de variables explicativas. Si una variable no interviene en la predicción de un compartimento, bastará con fijar el coeficiente en cero. Un modelo que predice un vector de respuesta en vez de una variable de respuesta es un modelo multivariado. Una observación para el ajuste de un modelo multivariado consiste en un vector $(Y_1, \ldots, Y_M, X_1, \ldots, X_p)$ de longitud $M + p$. El residuo de un modelo multivariado es un vector $\boldsymbol{\varepsilon}$ de longitud M, igual a la diferencia entre el vector de respuesta observado y el vector de respuesta predicho.

La expresión de un modelo M-variado sólo difiere de los M modelos univariados correspondientes a cada compartimento en la estructura del error residual; la estructura del

modelo para la media no cambia. Tomemos el caso general de un modelo no lineal. Si los M modelos univariados son:

$$Y_m = f_m(X_1, \ldots, X_p; \theta_m) + \varepsilon_m \tag{6.30}$$

para $m = 1, \ldots, M$, entonces el modelo multivariado se escribe:

$$\mathbf{Y} = \mathbf{F}(X_1, \ldots, X_p; \boldsymbol{\theta}) + \boldsymbol{\varepsilon}$$

donde $\mathbf{Y} = {}^{\mathrm{t}}[Y_1, \ldots, Y_M]$, $\boldsymbol{\theta} = {}^{\mathrm{t}}[\theta_1, \ldots, \theta_M]$, y

$$\mathbf{F}(X_1, \ldots, X_p; \boldsymbol{\theta}) = \begin{bmatrix} f_1(X_1, \ldots, X_p; \theta_1) \\ \vdots \\ f_m(X_1, \ldots, X_p; \theta_m) \\ \vdots \\ f_M(X_1, \ldots, X_p; \theta_M) \end{bmatrix} \tag{6.31}$$

El vector residual $\boldsymbol{\varepsilon}$ sigue ahora una distribución multinormal centrada, de matriz de varianza-covarianza:

$$\mathrm{Var}(\boldsymbol{\varepsilon}) \equiv \boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1^2 & \zeta_{12} & \cdots & \zeta_{1M} \\ \zeta_{21} & \sigma_2^2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \zeta_{M-1,M} \\ \zeta_{M1} & \cdots & \zeta_{M,M-1} & \sigma_M^2 \end{bmatrix}$$

La matriz $\boldsymbol{\Sigma}$ es una matriz simétrica con M filas y Mcolumnas, tal que $\sigma_m^2 = \mathrm{Var}(\varepsilon_m)$ es la varianza residual de la biomasa del m-ésimo compartimento y $\zeta_{ml} = \zeta_{lm}$ es la covarianza residual entre la biomasa del m-ésimo compartimento y aquella del l-ésimo compartimento. Como en el caso univariado, se supone que dos residuos que corresponden a dos observaciones diferentes, son independientes: $\boldsymbol{\varepsilon}_i$ es independiente de $\boldsymbol{\varepsilon}_j$ para $i \neq j$. La diferencia viene del hecho de que ya no se supone que los diferentes compartimentos son independientes unos de otros. El ajuste de un modelo multivariado como (6.31) se hace según los mismos principios que los modelos univariados (6.30). Si la matriz de varianza-covarianza $\boldsymbol{\Sigma}$ fuera diagonal (es decir, $\zeta_{ml} = 0, \forall m, l$), entonces el ajuste del modelo multivariado (6.31) sería equivalente al ajuste separado de los M modelos univariados (6.30). En el caso de un modelo lineal, los valores estimados de los coeficientes $\theta_1, \theta_2, \ldots, \theta_M$ resultantes del ajuste del modelo lineal M-variado, son idénticos a los valores obtenidos por los ajustes separados de los M modelos lineales univariados (siempre y cuando se mantengan las mismas variables explicativas X_1, $\ldots$, X_p en todos los casos) (Muller & Stewart, 2006, capitulo 3). No obstante, las pruebas de significancia asociadas a los coeficientes no dan los mismos resultados en ambos casos. Si los distintos compartimentos están suficientemente correlacionados entre sí, el ajuste simultáneo de todos los compartimentos mediante el modelo multivariado (6.31) llegará a una estimación más precisa de los coeficientes del modelo, es decir, a predicciones más precisas de la biomasa.

Armonización de un modelo

En ciertos casos, especialmente en el contexto de la dendroenergía, se desea predecir la biomasa seca del tronco en diferentes diámetros de corte. Por ejemplo, queremos predecir al mismo tiempo la biomasa total B del tronco, la biomasa B_7 del tronco hasta un diámetro de 7 cm en el extremo fino, y la biomasa B_{10} del tronco hasta un diámetro de 10 cm en el

extremo fino. Podríamos considerar entonces el tronco entero, el tronco hasta el corte de 7 cm y el tronco hasta el corte de 10 cm como tres compartimentos diferentes y aplicar los mismos principios de ajuste que los presentados en el párrafo anterior. En realidad, el problema es más complejo ya que, a diferencia de los compartimentos tronco y follaje que son distintos, aquéllos definidos por diferentes diámetros de corte se encajan unos dentro de otros: $B = B_7+$ biomasa del segmento que va del diámetro de 7 cm al extremo fino, y $B_7 = B_{10}+$ biomasa del segmento que va del diámetro de corte de 10 cm a 7 cm. De ese modo, el modelo multivariado que predice el vector (B, B_7, B_{10}) debe hacer que $B > B_7 > B_{10}$ se de en todo el ámbito de validez del modelo. El proceso que consiste en obligar al modelo multivariado a que prediga las biomasas de los diferentes compartimentos verificando al mismo tiempo la lógica de su anidación se denomina *armonización* de un modelo (Parresol, 1999). Jacobs & Cunia (1980) y Cunia & Briggs (1985b) propusieron soluciones a este problema en forma de ecuaciones que vinculan los coeficientes de los modelos de los distintos compartimentos. Hace falta ajustar entonces un modelo M-variado (si hay M diámetros de corte) cerciorándose que los coeficientes $\theta_1, \ldots, \theta_M$ corresponden a los M diámetros de corte satisfagan cierto número de ecuaciones que los vinculan. Cuando se estiman los coeficientes del modelo multivariado mediante la máxima verosimilitud, su estimación numérica se reduce a un problema de optimización con restricciones.

En el caso de la predicción del volumen o de la biomasa de un fuste, una alternativa a los modelos de volumen o de biomasa es la integración del perfil de tronco (Parresol & Thomas, 1989; Parresol, 1999). Tomemos $P(h)$ como el perfil de un tronco, es decir una curva que da la superficie de la sección transversal del tronco en función de la altura h a partir del suelo. La h representa también la longitud recorrida cuando seguimos el tronco desde su extremo más grueso hasta el más fino. (Maguire & Batista, 1996; Dean & Roxburgh, 2006; Metcalf *et al.*, 2009). Si la sección del fuste tiene una forma aproximadamente circular, el diámetro del árbol en la altura h puede ser calculado como: $D(h) = \sqrt{4P(h)/\pi}$. La biomasa del tronco hasta el diámetro de corte D se calcula integrando el perfil de tronco desde el suelo $(h = 0)$ hasta la altura $P^{-1}(\frac{\pi}{4}D^2)$ corresponde a dicho diámetro:

$$B_D = \int_0^{P^{-1}(\frac{\pi}{4}D^2)} \rho(h)\ P(h)\,\mathrm{d}h$$

donde $\rho(h)$ es la densidad de la madera a la altura h. El volumen del fuste hasta el diámetro de corte D se calcula del mismo modo, con la única diferencia de que ρ es remplazado por 1. El enfoque por perfil de tronco presenta la ventaja de que la armonización del modelo es automática. No obstante, se trata de un enfoque diferente desde el punto de vista conceptual del correspondiente a los modelos de volumen y de biomasa, con problemas de ajuste específicos (Fang & Bailey, 1999; Parresol, 1999), y que exceden el marco del presente manual. Cabe señalar que para los árboles muy grandes, para los cuáles la medición directa de la biomasa es prácticamente imposible, el enfoque del perfil de tronco ofrece una alternativa pertinente (Van Pelt, 2001; Dean *et al.*, 2003; Dean, 2003; Dean & Roxburgh, 2006; Sillett *et al.*, 2010).

7

Utilización y predicción

Una vez que se ha ajustado el modelo de volumen o de biomasa, hay diversos usos posibles para esas predicciones. Lo más frecuente será predecir el volumen o la biomasa de los árboles para los cuáles no se efectuaron esas mediciones. Se trata aquí de la *predicción* propiamente dicha (§ 7.2–7.4). A veces, el volumen o la biomasa de los árboles también habrán sido medidos además de las variables de entrada del modelo. Cuando se dispone de un conjunto de datos *independientes* del utilizado para el ajuste del modelo, y que contiene al mismo tiempo la variable de respuesta y las variables explicativas del modelo, es posible hacer una *validación* del mismo (§ 7.1). Cuando los criterios de validación se aplican al mismo conjunto de datos que sirvió para la calibración del modelo, se habla de *verificación* del modelo. No insistiremos sobre la verificación del modelo puesto que ya está implícita en el análisis de los residuos del modelo ajustado. Por último, cuando se dispone de modelos que existían antes de ajustar uno nuevo, se pueden comparar también los modelos o combinarlos (§ 7.5).

Ámbito de validez del modelo

Antes de usar cualquier modelo hay que cerciorarse que las características del árbol cuyo volumen o biomasa queremos predecir estén dentro del *ámbito de validez* del modelo (Rykiel, 1996). Si un modelo de volumen o de biomasa fue ajustado para árboles de diámetro comprendido entre $D_{\text{mín}}$ y $D_{\text{máx}}$, en principio no es posible usar ese modelo para predecir el volumen o la biomasa de un árbol de diámetro inferior a $D_{\text{mín}}$ o superior a $D_{\text{máx}}$. Lo mismo es válido para todas las entradas del modelo. Sin embargo, no todos los modelos están sujetos a los mismos errores cuando se los extrapola fuera de su ámbito de validez. Los modelos de potencia siguen siendo, en general, extrapolables con una buena fiabilidad fuera de su ámbito de validez porque estas relaciones de potencia se basan en un modelo alométrico fractal que es invariante a todas las escalas (Zianis & Mencuccini, 2004). Por el contrario, los modelos de tipo polinomial presentan con frecuencia comportamientos anormales fuera de su ámbito de validez (valores predichos negativos, por ejemplo), y mucho más aún a medida que aumenta el grado del polinomio.

7.1. Validación de un modelo

La validación de un modelo consiste en comparar sus predicciones con las observaciones independientes usadas para el ajuste de dicho modelo (Rykiel, 1996). Consideremos a $(Y_i', X_{i1}', \ldots, X_{ip}')$ con $i = 1, \ldots, n'$ como un conjunto de datos de n' observaciones independiente del usado para el ajuste de un modelo f, donde $X_{i1}', \ldots, X_{ip}'$ son las variables explicativas Y_i' es la variable de respuesta, es decir, el volumen o la biomasa, o una transformada de una de esas dos cantidades. Consideremos

$$\hat{Y}_i' = f(X_{i1}', \ldots, X_{ip}'; \hat{\theta})$$

el valor predicho de la variable de respuesta para la i-ésima observación, donde $\hat{\theta}$ son los valores estimados para los parámetros del modelo. La validación consiste en comparar los valores predichos $\hat{Y}_i'$ a los valores observados Y_i'.

7.1.1. Criterios de validación

Varios criterios, que son el equivalente de aquellos utilizados para evaluar la calidad del ajuste de un modelo, pueden usarse para comparar las predicciones a las observaciones (Schlaegel, 1982; Parresol, 1999; Tedeschi, 2006), en especial:

- el sesgo: $\sum_{i=1}^{n'} |Y_i' - \hat{Y}_i'|$
- la suma de los cuadrados de los residuos: $\text{SCE} = \sum_{i=1}^{n'} (Y_i' - \hat{Y}_i')^2$
- la varianza residual: $s^2 = \text{SCE}/(n' - p)$
- el error residual ajustado: $\text{SCE}/(n' - 2p)$
- el R^2 de regresión: $R^2 = 1 - s^2/\text{Var}(Y')$
- el criterio de información de Akaike: $\text{AIC} = n' \ln(s^2) + n' \ln(1 - p/n') + 2p$

donde $\text{Var}(Y')$ es la varianza empírica de Y' y p es el número de parámetros libremente estimado del modelo. Los dos primeros criterios corresponden a dos normas diferentes de la diferencia entre el vector $(Y_1', \ldots, Y_{n'}')$ de las observaciones y el vector $(\hat{Y}_1', \ldots, \hat{Y}_{n'}')$ de las predicciones: norma L^1 para el sesgo y norma L^2 para la suma de los cuadrados de las diferencias. Cualquier otra norma sería igualmente válida. Los tres últimos criterios involucran el número de parámetros usados en el modelo y, en consecuencia, son más adecuados cuando se trata de comparar diferentes modelos.

7.1.2. Validación cruzada

Cuando no se dispone de un conjunto de datos independiente, se tiene la tentación de dividir el conjunto de datos de calibración en dos subconjuntos de datos: uno para el ajuste del modelo y el otro para la validación del mismo. Dado que los conjuntos de datos de volumen o de biomasa son costosos y suelen ser de tamaño limitado, no recomendamos esta práctica cuando se construyan modelos de volumen o de biomasa. Los que recomendamos en este caso es una *validación cruzada* (Efron & Tibshirani, 1993, chapitre 17).

La validación cruzada "K veces" consiste en dividir el conjunto de datos en K partes más o menos iguales y usar cada parte una vez como conjunto de datos de validación, ajustándose el modelo en función de las $K - 1$ partes restantes. El seudoalgoritmo de validación cruzada "K veces" es el siguiente:

1. Dividir el conjunto de datos $\mathcal{S}_n \equiv \{(Y_i, X_{i1}, \ldots, X_{ip}): i = 1, \ldots, n\}$ en K subconjuntos de datos $\mathcal{S}_n^{(1)}, \ldots, \mathcal{S}_n^{(K)}$ de tamaños aproximadamente iguales (es decir, con aproximadamente n/K observaciones en cada subconjunto de datos, cuyo total da n).

2. Para k que va de 1 a K:

   a) ajustar el modelo a partir del conjunto de datos privado de su k-ésima parte, *es decir* a partir de $\mathcal{S}_n \backslash \mathcal{S}_n^{(k)} = \mathcal{S}_n^{(1)} \cup \ldots \cup \mathcal{S}_n^{(k-1)} \cup \mathcal{S}_n^{(k+1)} \cup \ldots \cup \mathcal{S}_n^{(K)}$;

   b) calcular un criterio de validación (cf. § 7.1.1) de dicho modelo ajustado tomando la parte restante $\mathcal{S}_n^{(k)}$ como conjunto de datos de validación; o sea, C_k el valor de ese criterio calculado para $\mathcal{S}_n^{(k)}$.

3. Calcular el promedio $(\sum_{k=1}^{K} C_k)/K$ de los K criterios de validación así calculados.

La ausencia de superposición entre los conjuntos de datos usados para el ajuste del modelo y aquellos utilizados para calcular el criterio de validación garantiza la validez de esta práctica. La validación cruzada exige más cálculos que una validación simple pero tiene la ventaja de aprovechar todas las observaciones disponibles para el ajuste del modelo.

Un caso particular de validación cruzada "K vezes" se da cuando K es igual al número n de observaciones disponibles en el conjunto de datos. Este método se llama también validación cruzada "dejando uno de lado" ("leave-one-out") y, desde el punto de vista conceptual, es similar a la técnica conocida como Jackknife (Efron & Tibshirani, 1993). El principio consiste en ajustar el modelo a partir de $n-1$ observaciones y en calcular el error residual para la observación dejada de lado. Se usa en análisis de residuos para cuantificar la influencia de las observaciones (en especial, es la base de cálculo de la distancia de Cook, cf. Saporta, 1990).

7.2. Predicción del volumen o de la biomasa de un árbol

La predicción con la ayuda de un modelo f consiste en calcular, para valores dados de las variables explicativas $X_1, \ldots, X_p$, el valor predicho $\hat{Y}$ por el modelo de la variable de respuesta. Una predicción no se detiene en el cálculo de

$$\hat{Y} = f(X_1, \ldots, X_p; \hat{\theta})$$

En efecto, el estimador $\hat{\theta}$ de los parámetros del modelo es un vector aleatorio cuya distribución se deriva de la distribución de las observaciones utilizadas para ajustar el modelo. Cualquier predicción $\hat{Y}$ del modelo resulta ella misma una variable aleatoria cuya distribución se desprende de distribución de las observaciones utilizadas para ajustar el modelo. Para expresar esta variabilidad intrínseca de la predicción, le asignaremos un indicador de incertidumbre como la desviación estándar de la predicción o su intervalo de confianza al 95 %.

Existen varios intervalos de confianza, según se prediga el volumen o la biomasa de un árbol tomado al azar en el rodal, o de un árbol promedio del rodal. Detallaremos las expresiones analíticas de dichos intervalos de confianza primero en el ejemplo del modelo lineal (§ 7.2.1), y, luego, en el ejemplo del modelo no lineal (§ 7.2.2). Las expresiones aproximadas pero más simples de calcular de estos intervalos de confianza se presentarán luego (§ 7.2.3), antes de interesarnos en el caso de las variables transformadas (§ 7.2.4).

7.2.1. Predicción: caso del modelo lineal

Predicción mediante una regresión lineal simple

Consideremos $\hat{a}$ como la intersección estimada para una regresión lineal, y $\hat{b}$ su pendiente estimada. La predicción $\hat{Y}$ de la variable de respuesta puede escribirse de dos formas distintas:

$$\hat{Y} = \hat{a} + \hat{b}X \tag{7.1}$$
$$\hat{Y} = \hat{a} + \hat{b}X + \varepsilon \tag{7.2}$$

En ambos casos, la esperanza de $\hat{Y}$ es la misma puesto que $\mathrm{E}(\varepsilon) = 0$. Por el contrario, la varianza de $\hat{Y}$ no es la misma en ambos casos: es más elevada en la segunda escritura que en la primera. La interpretación asociada a ambas escrituras es la siguiente. Supongamos que la variable explicativa X es el diámetro a la altura del pecho y la variable de respuesta Y la biomasa. El número de árboles en todo el bosque con un diámetro X dado (aproximado, que representa la precisión de la medición) es inconmensurable. Si pudiéramos medir la biomasa de todos esos árboles que tienen el mismo diámetro, encontraríamos valores variables, que oscilarían alrededor de cierto valor promedio. Cuando se trata de predecir esta biomasa promedio (sobreentendiéndose, promedio del conjunto de árboles existentes que tienen el diámetro X), la ecuación (7.1) de la predicción es válida. Por el contrario, si intentamos predecir la biomasa de un árbol tomado al azar entre el conjunto de árboles con diámetro X, la ecuación (7.2) de la predicción es válida. La variabilidad de la predicción es mayor para (7.2) que para (7.1) dado que, además de la variabilidad de la predicción de la biomasa media, en el segundo caso se suman a esto las diferencias de biomasa entre árboles.

Esto significa que hay dos formas de calcular un intervalo de confianza para una predicción. Hay un intervalo de confianza para la predicción del promedio de Y, y un intervalo de confianza para la predicción de un individuo tomado al azar de la población sobre la cual se calculó la media de Y. El segundo intervalo de confianza es más amplio que el primero.

En el caso de una regresión lineal simple, se puede demostrar (Saporta, 1990, p.373–374) que el intervalo de confianza en el umbral α para la predicción (7.1) de la media es:

$$\hat{a} + \hat{b}X \pm t_{n-2}\,\hat{\sigma}\sqrt{\frac{1}{n} + \frac{(X - \bar{X})^2}{nS_X^2}} \tag{7.3}$$

mientras que el intervalo de confianza en el umbral α para la predicción (7.2) de un árbol tomado al azar es:

$$\hat{a} + \hat{b}X \pm t_{n-2}\,\hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(X - \bar{X})^2}{nS_X^2}} \tag{7.4}$$

donde t_{n-2} es el cuantile $1 - \alpha/2$ de una distribución de t de Student a $n - 2$ grados de libertad, $\bar{X} = (\sum_{i=1}^{n} X_i)/n$ es la media de los valores observados de X en el conjunto de datos que sirvieron para ajustar el modelo, y $S_X^2 = [\sum_{i=1}^{n}(X_i - \bar{X})^2]/n$ es la varianza empírica de los valores observados de X en el conjunto de datos que sirvió para ajustar el modelo.

Estas expresiones suscitan varias observaciones. La primera es que la diferencia entre los límites del intervalo de confianza (7.4) para un árbol tomado al azar y los límites del intervalo de confianza (7.3) para un árbol promedio es del orden de $t_{n-2}\hat{\sigma}$. Esta diferencia refleja la diferencia entre las ecuaciones (7.2) y (7.1), que depende del término residual ε cuya desviación estándar es $\hat{\sigma}$.

La segunda es que la amplitud del intervalo de confianza no es constante sino que varía con X. El intervalo de confianza es más estrecho cuando $X = \bar{X}$ se amplía cuando X se aleja de $\bar{X}$.

La tercera observación es que para calcular el intervalo de confianza de una predicción en función de una regresión lineal, hay que disponer, si no se tienen los datos originales que sirvieron para ajustar el modelo, por lo menos de la media $\bar{X}$ de la variable explicativa y de su desviación estándar empírica S_X. Si los datos originales que sirvieron para el ajuste del modelo ya no están disponibles y si los valores de $\bar{X}$ y S_X no se documentaron, no se podrá calcular el intervalo de confianza en forma exacta.

33

Intervalo de confianza de $\ln(B)$ predicho por $\ln(D)$

Retomemos la regresión lineal simple entre $\ln(B)$ y $\ln(D)$ que fue ajustada en la Línea roja 7. Consideremos `m` el objeto que contiene el modelo ajustado (cf. Línea roja 7). Los intervalos de confianza pueden calcularse con el comando `predict`. Por ejemplo, para un árbol de diámetro 20 cm, el intervalo de confianza con una incertidumbre del 95 % para el árbol promedio se obtiene mediante el comando:

```
predict(m,newdata=data.frame(dbh=20),interval="confidence",level=0.95)
```

lo que da:

```
          fit        lwr        upr
1  -1.354183  -1.533487  -1.174879
```

De este modo, el modelo predice $\ln(B) = -1{,}354183$ con un intervalo de confianza de 95 % que va de $-1{,}533487$ a $-1{,}174879$. Para un árbol de 20 cm tomado al azar, el intervalo de confianza se obtiene con el comando:

```
predict(m,newdata=data.frame(dbh=20),interval="prediction",level=0.95)
```

lo que da:

```
          fit        lwr         upr
1  -1.354183  -2.305672  -0.4026948
```

La Figura 7.1 muestra los intervalos de confianza en todo el intervalo de datos.

Predicción mediante una regresión múltiple

Los principios de la predicción, expuestos en el caso de la regresión lineal, se aplican inmediatamente a la regresión múltiple. Hay dos expresiones del intervalo de confianza: una para la predicción del árbol medio y otra para la predicción de un árbol tomado al azar.

En el caso de una regresión múltiple de coeficientes estimados $\hat{\mathbf{a}} = {}^{\mathrm{t}}[\hat{a}_0, \hat{a}_1, \hat{a}_2, \ldots, \hat{a}_p]$, el valor predicho $\hat{Y}$ de la variable de respuesta para un árbol cuyas variables explicativas son $\mathbf{x} = {}^{\mathrm{t}}[1, X_1, X_2, \ldots, X_p]$, es:

$$\hat{Y} = {}^{\mathrm{t}}\mathbf{x}\,\hat{\mathbf{a}}$$

y el intervalo de confianza en el umbral α de esta predicción es (Saporta, 1990, p.387):

- para la predicción del árbol medio:

$$ {}^{\mathrm{t}}\mathbf{x}\,\hat{\mathbf{a}} \pm t_{n-p-1}\,\hat{\sigma}\sqrt{{}^{\mathrm{t}}\mathbf{x}({}^{\mathrm{t}}\mathbf{X}\mathbf{X})^{-1}\mathbf{x}} \tag{7.5}$$

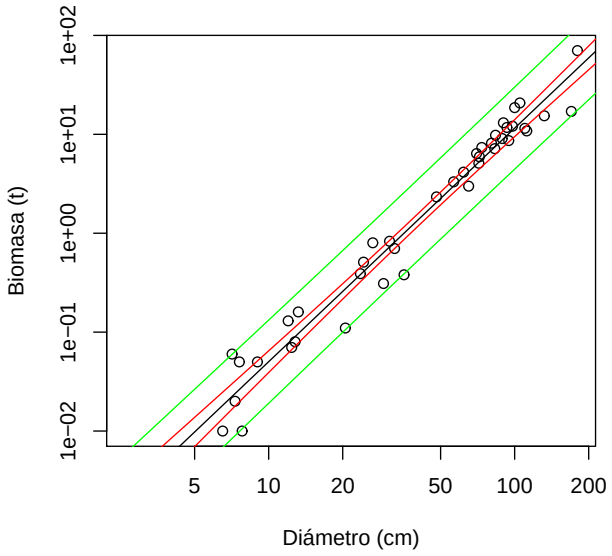


Figura 7.1 – *Datos de biomasa en función del diámetro (en escala logarítmica) para 42 árboles medidos en Ghana por Henry et al. (2010) (puntos), predicción (línea negra) de la regresión lineal simple de* $\ln(B)$ *con respecto a* $\ln(D)$*, e intervalos de confianza de esta predicción para un árbol tomado al azar (línea verde) y para el árbol medio (Línea roja).*

- para la predicción de un árbol tomado al azar:

$$ {}^{t}\mathbf{x}\,\hat{\mathbf{a}} \pm t_{n-p-1}\,\hat{\sigma}\sqrt{1 + {}^{t}\mathbf{x}({}^{t}\mathbf{X}\mathbf{X})^{-1}\mathbf{x}} \tag{7.6} $$

donde $\mathbf{X}$ es la matriz de diseño construida a partir de los datos que sirvieron para ajustar la regresión múltiple. Para calcular el intervalo de confianza de las predicciones hay que conocer los datos originales que sirvieron para el ajuste del modelo o, por lo menos, la matriz $({}^{t}\mathbf{X}\mathbf{X})^{-1}$. Cabe señalar que la varianza de las predicciones en el caso (7.6) de un árbol tomado al azar se compone de dos términos: un término $\hat{\sigma}^2$ que representa el error residual y un término $\hat{\sigma}^2\ {}^{t}\mathbf{x}({}^{t}\mathbf{X}\mathbf{X})^{-1}\mathbf{x}$ que representa la variabilidad inducida mediante la estimación de los coeficientes del modelo. En el caso de la estimación del árbol promedio, el primer término desaparece y sólo queda el segundo.

34

Intervalo de confianza de $\ln(B)$ **predicho por** $\ln(D)$ **y** $\ln(H)$

Retomemos la regresión lineal múltiple entre $\ln(B)$, $\ln(D)$ y $\ln(H)$ que fue ajustada en la lína roja 10. Consideremos `m` como el objeto que contiene el modelo ajustado (cf. Línea roja 10). Los intervalos de confianza pueden calcularse con el comando `predict`. Por ejemplo, para un árbol de diámetro 20 cm y de altura 20 m, el intervalo de confianza con una incertidumbre del 95 % para el árbol promedio se obtiene mediante el comando:

```
predict(m,newdata=data.frame(dbh=20,haut=20),interval="confidence",level=0.95)
```

lo que da:

```
         fit        lwr        upr
1  -1.195004  -1.380798  -1.009211
```

De esta forma el modelo predicho $\ln(B) = -1{,}195004$ con un intervalo de confianza del 95 % va de $-1{,}380798$ a $-1{,}009211$. Para un árbol de 20 cm de diámetro y de 20 m de altura, tomado al azar, el intervalo de confianza se obtiene mediante el comando:

```
predict(m,newdata=data.frame(dbh=20,haut=20),interval="prediction",level=0.95)
```

lo que da:

```
          fit        lwr         upr
1  -1.195004  -2.046408  -0.3436006
```

7.2.2. Predicción: caso de un modelo no lineal

En el caso general de un modelo no lineal tal como el definido por

$$Y = f(X_1, \ldots, X_p; \theta) + \varepsilon$$

con

$$\varepsilon \sim \mathcal{N}(0,\ kX_1^c)$$

no hay expresión explícita exacta de los intervalos de confianza de las predicciones, como ocurre con el modelo lineal. No obstante, el δ-método permite obtener una expresión aproximada (y asintóticamente exacta) de los intervalos de confianza (Serfling, 1980). Al igual que antes, hay dos intervalos de confianza:

- intervalo de confianza para la predicción del árbol promedio:

$$f(X_1, \ldots, X_p; \hat{\theta}) \pm t_{n-q} \sqrt{{}^{\mathrm{t}}[\mathrm{d}_\theta f(\hat{\theta})]\ \hat{\mathbf{\Sigma}}_{\hat{\theta}}\ [\mathrm{d}_\theta f(\hat{\theta})]} \tag{7.7}$$

- intervalo de confianza para la predicción de un árbol tomado al azar:

$$f(X_1, \ldots, X_p; \hat{\theta}) \pm t_{n-q} \sqrt{\hat{k}^2 X_1^{2\hat{c}} + {}^{\mathrm{t}}[\mathrm{d}_\theta f(\hat{\theta})]\ \hat{\mathbf{\Sigma}}_{\hat{\theta}}\ [\mathrm{d}_\theta f(\hat{\theta})]} \tag{7.8}$$

donde q es el número de coeficientes del modelo (es decir, la longitud del vector θ), $\mathrm{d}_\theta f(\hat{\theta})$ es el valor en $\theta = \hat{\theta}$ de la diferencial de f con respecto a los coeficientes del modelo, y $\hat{\mathbf{\Sigma}}_{\hat{\theta}}$ es una estimación en $\theta = \hat{\theta}$ de la matriz de varianza-covarianza $\mathbf{\Sigma}_\theta$ del estimador de θ. La diferencial de f con respecto a los coeficientes del modelo es el vector de longitud q:

$$\mathrm{d}_\theta f(\theta) = {}^{\mathrm{t}}\left[\left(\frac{\partial f(X_1, \ldots, X_p; \theta)}{\partial \theta_1}\right), \ \ldots, \ \left(\frac{\partial f(X_1, \ldots, X_p; \theta)}{\partial \theta_q}\right)\right]$$

donde θ_i es le i-ésimo elemento del vector θ, En el caso del estimador de máxima verosimilitud de θ, se puede demostrar que, asintóticamente, cuando $n \to \infty$ (Saporta, 1990, p.301):

$$\mathbf{\Sigma}_\theta \underset{n\to\infty}{\sim} \mathbf{I}_n(\theta)^{-1} = \frac{1}{n}\, \mathbf{I}_1(\theta)^{-1}$$

donde $\mathbf{I}_n(\theta)$ es la matriz de la información de Fisher aportada por una muestra de tamaño n sobre el vector de parámetros θ. Esta matriz de información de Fisher tiene q lineas y q columnas y se calcula a partir de la segunda derivada de la log-verosimilitud de la muestra:

$$\mathbf{I}_n(\theta) = -\mathrm{E}\left[\frac{\partial^2 \mathcal{L}(\theta)}{\partial \theta^2}\right]$$

Una estimación aproximada de la matriz de varianza-covarianza de los parámetros es pues:

$$\hat{\mathbf{\Sigma}}_{\hat{\theta}} = -\left[\left(\frac{\partial^2 \mathcal{L}(\theta)}{\partial \theta^2}\right)\Big|_{\theta=\hat{\theta}}\right]^{-1}$$

En la práctica, el algoritmo que optimiza numéricamente la log-verosimilitud de la muestra da, al mismo tiempo, una estimación numérica de la segunda derivada ($\partial^2\mathcal{L}/\partial\theta^2$). Así obtenemos de inmediato una estimación numérica de $\hat{\mathbf{\Sigma}}_{\hat{\theta}}$.

Al igual que antes, la varianza de las predicciones en el caso (7.8) de un árbol tomado al azar se compone de dos términos: un término $(\hat{k}X_1^{\hat{c}})^2$ que representa el error residual y un término ${}^{\mathrm{t}}[\mathrm{d}_\theta f(\hat{\theta})]\ \hat{\mathbf{\Sigma}}_{\hat{\theta}}\ [\mathrm{d}_\theta f(\hat{\theta})]$ que representa la variabilidad inducida por la estimación de los coeficientes del modelo. En el caso de la estimación del árbol promedio, el primer término desaparece y sólo queda el segundo.

7.2.3. Intervalos de confianza aproximados

El cálculo exacto de los intervalos de confianza de las predicciones exige información (matriz de $\mathbf{X}$ en el caso del modelo lineal, matriz de varianza-covarianza $\hat{\mathbf{\Sigma}}_{\hat{\theta}}$ en el caso del no lineal) que muy raramente se indica en las publicaciones relativas a los modelos de volumen o biomasa. Con mucha frecuencia, las publicaciones sólo indican el número n de observaciones usadas para ajustar el modelo y la desviación estándar residual $\hat{\sigma}$ (caso lineal) o $\hat{k}$ y $\hat{c}$ (caso no lineal). A veces, esa información básica sobre el ajuste ni siquiera se da. Cuando no se suministran $\mathbf{X}$ (caso del modelo lineal) ni $\hat{\mathbf{\Sigma}}_{\hat{\theta}}$ (caso del modelo no lineal), no es posible usar las fórmulas anteriores para calcular los intervalos de confianza. En ese caso, se utilizará un método aproximado.

Error residual solo

Con mucha frecuencia, sólo se da la desviación estándar residual $\hat{\sigma}$ (caso lineal) o $\hat{k}$ y $\hat{c}$ (caso no lineal). En ese caso, se podrá construir un intervalo de confianza aproximado en el umbral α:

- en el caso de una regresión lineal:

$$(a_0 + a_1 X_1 + \ldots + a_p X_p) \pm q_{1-\alpha/2}\,\hat{\sigma} \tag{7.9}$$

- en el caso de una regresión no lineal:

$$f(X_1, \ldots, X_p; \theta) \pm q_{1-\alpha/2}\,\hat{k}X_1^{\hat{c}} \tag{7.10}$$

donde $q_{1-\alpha/2}$ es el cuantile $1-\alpha/2$ de la distribución normal estándar. Este intervalo de confianza es una retranscripción directa de la relación $Y = a_0 + a_1 X_1 + \ldots + a_p X_p + \varepsilon$ con $\varepsilon \sim \mathcal{N}(0, \hat{\sigma})$ (caso lineal) o $Y = f(X_1, \ldots, X_p; \theta) + \varepsilon$ con $\varepsilon \sim \mathcal{N}(0, \hat{k}X_1^{\hat{c}})$ (caso no lineal), donde se escribieron a propósito los coeficientes del modelo sin acento circunflejo para destacar que aquí se trata de magnitudes fijas. Estas relaciones suponen pues implícitamente que los coeficientes del modelo se conocen exactamente y que la única fuente de variabilidad es el error residual. En otras palabras, la interpretación de estos intervalos de confianza aproximados es la siguiente: los intervalos de confianza (7.9) (caso lineal) y (7.10) (caso no lineal) son los que se obtendrían para la predicción de un árbol *tomado al azar* si el tamaño de la muestra fuera *infinito*. Esto se verificará, en efecto, cuando $n \to \infty$, t_{n-p-1} tiende hacia $q_{1-\alpha/2}$ y la matriz $({}^{\mathrm{t}}\mathbf{X}\mathbf{X})^{-1}$ en (7.6) tiende hacia la matriz nula (en la que todos los coeficientes valen cero). Por lo tanto, el intervalo de confianza (7.9) es realmente el límite del intervalo de confianza (7.6) cuando $n \to \infty$. Lo mismo se aplica para (7.8) y (7.10).

Intervalo de confianza para el árbol promedio

Cuando se da una estimación $\hat{\mathbf{\Sigma}}$ de la matriz de varianza-covarianza de los parámetros, un intervalo de confianza en el umbral α de la predicción para el árbol promedio es:

- en el caso del modelo lineal:

$$(\hat{a}_0 + \hat{a}_1 X_1 + \ldots + \hat{a}_p X_p) \pm q_{1-\alpha/2}\sqrt{{}^{\mathrm{t}}\mathbf{x}\hat{\mathbf{\Sigma}}\mathbf{x}} \tag{7.11}$$

  donde $\mathbf{x}$ es el vector ${}^{\mathrm{t}}[X_1, \ldots, X_p]$,

- en el caso del modelo no lineal:

$$f(X_1, \ldots, X_p; \hat{\theta}) \pm q_{1-\alpha/2}\sqrt{{}^{\mathrm{t}}[\mathrm{d}_\theta f(\hat{\theta})]\ \hat{\mathbf{\Sigma}}\ [\mathrm{d}_\theta f(\hat{\theta})]} \tag{7.12}$$

Estos intervalos de confianza consideran que toda la variabilidad de la predicción proviene de la estimación de los coeficientes del modelo. Además, dichos intervalos de confianza son una retranscripción directa del hecho de que los coeficientes del modelo siguen una distribución multinormal (también llamada distribución normal multivariante) de media igual a su valor verdadero y de matriz de varianza-covarianza $\hat{\mathbf{\Sigma}}$. En efecto en el caso lineal, si $\hat{\mathbf{a}} = {}^{\mathrm{t}}[\hat{a}_1, \ldots, \hat{a}_p]$ sigue una distribución multinormal de media ${}^{\mathrm{t}}[a_1, \ldots, a_p]$ y de matriz de varianza-covarianza $\hat{\mathbf{\Sigma}}$, entonces la combinación lineal ${}^{\mathrm{t}}\mathbf{x}\,\hat{\mathbf{a}}$ sigue una distribución normal de media ${}^{\mathrm{t}}\mathbf{x}\,\mathbf{a}$ y de varianza ${}^{\mathrm{t}}\mathbf{x}\hat{\mathbf{\Sigma}}\mathbf{x}$ (Saporta, 1990, p.85).

En el caso del modelo lineal, se puede demostrar que la matriz de varianza-covarianza del estimador de coeficientes del modelo es (Saporta, 1990, p.380): $\mathbf{\Sigma} = \sigma^2({}^{\mathrm{t}}\mathbf{X}\mathbf{X})^{-1}$. Así pues, una estimación de esta matriz de varianza-covarianza es: $\hat{\mathbf{\Sigma}} = \hat{\sigma}^2({}^{\mathrm{t}}\mathbf{X}\mathbf{X})^{-1}$. Al colocar esta expresión en (7.11), creamos una expresión parecida en (7.5). Asimismo, se verifica, en el caso no lineal, que el intervalo de confianza (7.12) es una aproximación de (7.7).

En el caso no lineal (7.12), si queremos evitar tener que calcular las derivadas parciales de f, podremos usar un método de Montecarlo. Es un método basado en la simulación que consiste en hacer Q simulaciones de coeficientes θ según una distribución multinormal de media $\hat{\theta}$ y de matriz de varianza-covarianza $\hat{\mathbf{\Sigma}}$, en calcular la predicción para cada uno de esos valores simulados, y en calcular a continuación el intervalo de confianza empírico de esas Q predicciones. En la bibliografía se dice que este método brinda "intervalos de predicción de la población" ("population prediction intervals", en inglés (Bolker, 2008; Paine *et al.*, 2012). El seudoalgorimo es el siguiente:

1. Para k que va de 1 a Q:
   a) escoger un vector $\hat{\theta}^{(k)}$ que sigue una distribución multinormal de media $\hat{\theta}$ y de matriz de varianza-covarianza $\hat{\mathbf{\Sigma}}$;
   b) calcular la predicción $\hat{Y}^{(k)} = f(X_1, \ldots, X_p; \hat{\theta}^{(k)})$.
2. El intervalo de confianza de la predicción es el intervalo de confianza empírico de los Q valores $\hat{Y}^{(1)}, \ldots, \hat{Y}^{(Q)}$.

Con mucha frecuencia no se conoce la matriz de varianza-covarianza $\hat{\mathbf{\Sigma}}$, pero se dispone, al menos, de una estimación de las desviaciones estándar de los coeficientes. Siendo $\mathrm{Var}(\hat{a}_i) = \Sigma_i$ (caso lineal) o $\mathrm{Var}(\hat{\theta}_i) = \Sigma_i$ (caso no lineal) la varianza del i-ésimo coeficiente del modelo. En este caso dejaremos de lado la correlación entre los coeficientes y nos aproximaremos a la matriz de varianza-covarianza de los coeficientes mediante una matriz diagonal:

$$\hat{\mathbf{\Sigma}} \approx \begin{bmatrix} \hat{\Sigma}_1 & & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & \hat{\Sigma}_p \end{bmatrix}$$

Intervalo de confianza para un árbol escogido al azar

El error resultante de la estimación de los coeficientes del modelo tal como se describió en el párrafo anterior puede acumularse con el error residual descrito en el penúltimo párrafo, para construir un intervalo de confianza de la predicción para un árbol tomado al azar. Ésas son las varianzas de las predicciones que se suman unas con otras α será entonces:

- en el caso del modelo lineal:

$$(\hat{a}_0 + \hat{a}_1 X_1 + \ldots + \hat{a}_p X_p) \pm q_{1-\alpha/2} \sqrt{\hat{\sigma}^2 + {}^{\mathrm{t}}\mathbf{x}\hat{\mathbf{\Sigma}}\mathbf{x}}$$

  que es una aproximación de (7.6),

- en el caso no lineal:

$$f(X_1, \ldots, X_p; \hat{\theta}) \pm q_{1-\alpha/2} \sqrt{\hat{k}^2 X_1^{2\hat{c}} + {}^{\mathrm{t}}[\mathrm{d}_\theta f(\hat{\theta})]\ \hat{\mathbf{\Sigma}}\ [\mathrm{d}_\theta f(\hat{\theta})]}$$

  que es una aproximación de (7.8).

Al igual que antes, si queremos evitar hacer demasiados cálculos, podríamos usar un método de Montecarlo según el seudoalgoritmo siguiente:

1. Para k que va de 1 a Q:
   a) escoger un vector $\hat{\theta}^{(k)}$ según una distribución multinormal de media $\hat{\theta}$ y de matriz de varianza-covarianza $\hat{\mathbf{\Sigma}}$;
   b) escoger un residuo $\hat{\varepsilon}^{(k)}$ según una distribución normal centrada de desviación estándar $\hat{\sigma}$ (caso lineal) o $\hat{k} X_1^{\hat{c}}$ (caso no lineal);
   c) calcular la predicción $\hat{Y}^{(k)} = f(X_1, \ldots, X_p; \hat{\theta}^{(k)}) + \hat{\varepsilon}^{(k)}$.
2. El intervalo de confianza de la predicción es el intervalo de confianza empírico de los Q valores $\hat{Y}^{(1)}, \ldots, \hat{Y}^{(Q)}$.

Intervalo de confianza con incertidumbres de medición

El ajuste de los modelos de volumen y de biomasa supone que las variables explicativas $X_1, \ldots, X_p$ se conocen exactamente. En realidad esta hipótesis no es más que una aproximación ya que esas magnitudes se miden y, en consecuencia, están sujetas a un error de medición. No hay que confundir el error de medición con el error residual de la variable de respuesta. El primero está asociado al instrumento de medición y, en principio, puede volverse tan pequeño como queramos usando instrumentos de medición cada vez más precisos. El segundo refleja una variabilidad biológica intrínseca entre los individuos. Podemos incluir el impacto del error de medición en la predicción al incorporarlo en el intervalo de confianza de la predicción. En consecuencia, las variables explicativas $X_1, \ldots, X_p$ ya no se consideran fijas sino como parte de una cierta distribución. Típicamente, para predecir el volumen o la biomasa de un árbol de características $X_1, \ldots, X_p$, se considera que la i-ésima característica está distribuida en función de una distribución normal de media X_i y de desviación estándar τ_i. Típicamente, si X_i es un diámetro, tomaremos τ_i del orden de 3–5 mm; si X_i es una altura, τ_i es del orden del 3 % de X_i para $X_i \leq 15$ m y del orden de 1 m para $X_i > 15$ m.

Es difícil calcular una expresión explícita del intervalo de confianza de la predicción cuando las variables explicativas son consideradas como aleatorias, ya que eso implica calcular

las varianzas de productos de variables aleatorias algunas de las cuáles están correlacionadas entre sí. El δ-método ofrece una solución analítica aproximada (Serfling, 1980). O bien, más sencillamente, se puede usar nuevamente el método de Montecarlo. El seudoalgoritmo se convierte en:

1. Para k que va de 1 a Q:
   a) para i que va de 1 a p, escoger $\hat{X}_i^{(k)}$ que sigua una distribución normal de media X_i y de desviación estándar τ_i;
   b) escoger un vector $\hat{\theta}^{(k)}$ que sigua una distribución multinormal de media $\hat{\theta}$ y de matriz de varianza-covarianza $\hat{\mathbf{\Sigma}}$;
   c) escoger un residuo $\hat{\varepsilon}^{(k)}$ que sigua una distribución normal centrada de desviación estándar $\hat{\sigma}$ (caso lineal) o $\hat{k}X_1^{\hat{c}}$ (caso no lineal);
   d) calcular la predicción $\hat{Y}^{(k)} = f(\hat{X}_1^{(k)}, \ldots, \hat{X}_p^{(k)}; \hat{\theta}^{(k)}) + \hat{\varepsilon}^{(k)}$.
2. El intervalo de confianza de la predicción es el intervalo de confianza empírico de los Q valores $\hat{Y}^{(1)}, \ldots, \hat{Y}^{(Q)}$.

Este intervalo de confianza corresponde en este caso a la predicción de un árbol tomado al azar. Para obtener el intervalo de confianza para el árbol promedio, basta aplicar el mismo seudoalgoritmo remplazando la etapa (c) por:

(...)
   c) plantear $\hat{\varepsilon}^{(k)} = 0$;
(...)

7.2.4. Transformación inversa de variables

En la Sección 6.1.5 vimos cómo una transformación de variable podía linealizar un modelo que inicialmente no correspondía a las hipótesis del modelo lineal. La transformación de variable actúa al mismo tiempo sobre la media y sobre el error residual. Lo mismo ocurrirá con la transformación inversa, con las consecuencias para el cálculo de la esperanza de las predicciones. La transformación logarítmica es la más frecuente. Sin embargo, existen también otros tipos de transformaciones.

Transformación logarítmica

Consideremos primero el caso de la transformación logarítmica sobre el volumen o la biomasa, que es, por mucho, el caso más frecuente para los modelos de volumen y de biomasa. Supongamos que una transformación logarítmica fue aplicada a la biomasa B para ajustar un modelo lineal con respecto a las variables explicativas $X_1, \ldots, X_p$:

$$Y = \ln(B) = a_0 + a_1 X_1 + \ldots + a_p X_p + \varepsilon \tag{7.13}$$

con

$$\varepsilon \underset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma)$$

Esto equivale a decir que $\ln(B)$ sigue una distribución normal de media $a_0 + a_1 X_1 + \ldots + a_p X_p$ y de desviación estándar σ o incluso, por definición, que B sigue una distribución lognormal de parámetros $a_0 + a_1 X_1 + \ldots + a_p X_p$ y σ. La esperanza de esta distribución lognormal es:

$$\mathrm{E}(B) = \exp\left(a_0 + a_1 X_1 + \ldots + a_p X_p + \frac{\sigma^2}{2}\right)$$

Comparado al modelo inverso de (7.13) que es $B = \exp(a_0 + a_1X_1 + \ldots + a_pX_p)$, la transformación inversa del error residual induce un sesgo de predicción que puede corregirse multiplicando la predicción $\exp(a_0 + a_1X_1 + \ldots + a_pX_p)$ por un factor de corrección (Parresol, 1999):

$$\mathrm{CF} = \exp\left(\frac{\hat{\sigma}^2}{2}\right) \tag{7.14}$$

Los modelos de biomasa en la bibliografía que fueron ajustadas después de la transformación logarítmica de la biomasa no siempre incluyen el factor de corrección, por lo que hay que ser precavidos.

Si se usó el logaritmo decimal $\log_{10}$ para la transformación de la variable en vez del logaritmo neperiano, el coeficiente corrector es:

$$\mathrm{CF} = \exp\left[\frac{(\hat{\sigma}\ln 10)^2}{2}\right] \approx \exp\left(\frac{\hat{\sigma}^2}{0{,}3772}\right)$$

35

Factor de corrección de la biomasa predicha

Retomemos el ejemplo del modelo de biomasa ajustado en la Línea roja 31 mediante regresión múltiple a partir de los datos transformados logarítmicamente:

$$\ln(B) = -8{,}38900 + 0{,}85715\ln(D^2H) + 0{,}72864\ln(\rho)$$

Si volvemos a los datos de partida usando la función exponencial (sin tener en cuenta el factor de corrección), obtenemos una predicción subestimada: $B = \exp(-8{,}38900) \times (D^H)^{0{,}85715}\rho^{0{,}72864} = 2{,}274 \times 10^{-4}(D^2H)^{0{,}85715}\rho^{0{,}72864}$. Consideremos `m` como el objeto que contiene el modelo ajustado (cf. Línea roja 31). El factor de corrección $\mathrm{CF} = \exp(\hat{\sigma}^2/2)$ se obtiene mediante el comando:

```
exp(summary(m)$sigma^2/2)
```

y resulta ser 1,061035. El modelo correcto es entonces: $B = 2{,}412 \times 10^{-4}(D^2H)^{0{,}85715}\rho^{0{,}72864}$.

Cualquier otra transformación

En el caso general, consideremos ψ como una transformación de variable de la biomasa (o del volumen) tal que la variable de respuesta $Y = \psi(B)$ pueda predecirse mediante una regresión lineal con respecto a las variables explicativas $X_1, \ldots, X_p$. Supongamos que la función ψ derivable e invertible. Como $\psi(B)$ sigue una distribución normal de media $a_0 + a_1X_1 + \ldots + a_pX_p$ y de desviación estándar σ, $B = \psi^{-1}[\psi(B)]$ tiene por esperanza (Saporta, 1990, p.26):

$$\mathrm{E}(B) = \int \psi^{-1}(x)\ \phi(x)\ \mathrm{d}x \tag{7.15}$$

donde ϕ es la densidad de probabilidad de la distribución normal de media $a_0 + a_1X_1 + \ldots + a_pX_p$ y de desviación estándar σ. Esta esperanza es generalmente diferente de $\psi^{-1}(a_0 + a_1X_1 + \ldots + a_pX_p)$: la transformación de variable induce un sesgo de predicción cuando se vuelve a la variable de partida mediante la transformación inversa. El inconveniente de la fórmula (7.15) es que necesita el cálculo de una integral.

Cuando la desviación estándar residual σ es pequeña, el δ-método (Serfling, 1980) aporta una expresión aproximada de ese sesgo de predicción:

$$\begin{aligned} \mathrm{E(B)} &\simeq \psi^{-1}[\mathrm{E}(Y)] + \frac{1}{2}\mathrm{Var}(Y)\ (\psi^{-1})''[\mathrm{E}(Y)] \\ &\simeq \psi^{-1}(a_0 + a_1X_1 + \ldots + a_pX_p) + \frac{\sigma^2}{2}\ (\psi^{-1})''(a_0 + a_1X_1 + \ldots + a_pX_p) \end{aligned}$$

Estimación "smearing"

El método de estimación "smearing" (que podríamos traducir como "de dispersión" o " dispersante") es un método no paramétrico de corrección del sesgo de predicción cuando se aplica una transformación inversa a la variable de respuesta de un modelo lineal (Duan, 1983; Taylor, 1986; Manning & Mullahy, 2001). Dado que se puede reescribir la ecuación (7.15) de la esperanza de la biomasa (o del volumen) de la siguiente forma:

$$\begin{aligned} \mathrm{E}(B) &= \int \psi^{-1}(x)\ \phi_0(x - a_0 - a_1X1 - \ldots - a_pX_p)\ \mathrm{d}x \\ &= \int \psi^{-1}(x + a_0 + a_1X1 + \ldots + a_pX_p)\ \mathrm{d}\Phi_0(x) \end{aligned}$$

donde ϕ_0 (respectivamente Φ_0) es la densidad de probabilidad (respectivamente la función de distribución) de la distribución normal centrada de la desviación estándar σ, el método smearing consiste en remplazar Φ_0 por la función de repartición empírica de los residuos del ajuste del modelo, o sea:

$$\begin{aligned} B_{\text{smearing}} &= \int \psi^{-1}(x + a_0 + a_1X_1 + \ldots + a_pX_p) \times \frac{1}{n}\sum_{i=1}^{n}\delta(x - \hat{\varepsilon}_i)\,\mathrm{d}x \\ &= \frac{1}{n}\sum_{i=1}^{n}\psi^{-1}(a_0 + a_1X_1 + \ldots + a_pX_p + \hat{\varepsilon}_i) \end{aligned}$$

donde δ es la distribución de Dirac en cero y $\hat{\varepsilon}_i$ es el residuo del modelo ajustado para la i-ésima observación. Este método de corrección del sesgo de predicción tiene la ventaja de ser, al mismo tiempo, muy general y fácil de calcular. Tiene el inconveniente de que hay que conocer los residuos $\hat{\varepsilon}_i$ del ajuste del modelo. Esto no representa un problema cuando uno mismo ajusta un modelo a los datos, pero sí lo es cuando se usa un modelo publicado para el cual no se dan los residuos.

En el caso particular de la transformación logarítmica, ψ^{-1} la función exponencial y, en consecuencia, la estimación smearing de la biomasa es: $\exp(a_0 + a_1X_1 + \ldots + a_pX_p) \times \mathrm{CF}_{\text{smearing}}$, donde el factore de corrección smearing es:

$$\mathrm{CF}_{\text{smearing}} = \frac{1}{n}\sum_{i=1}^{n}\exp(\hat{\varepsilon}_i)$$

Dado $\hat{\sigma}^2 = (\sum_{i=1}^{n}\hat{\varepsilon}_i^2)/(n-p-1)$, el factor de corrección smearing es diferente del del factor de corrección (7.14). Sin embargo, dentro del límite en el cual $\hat{\sigma} \to 0$, ambos factores son equivalentes.

36

Estimación "smearing" de la biomasa

Retomemos una vez más el ejemplo del modelo de biomasa ajustado en la Línea roja 31 mediante regresión múltiple a partir de los datos transformados logarítmicamente:

$$\ln(B) = -8{,}38900 + 0{,}85715\ln(D^2H) + 0{,}72864\ln(\rho)$$

El factor de corrección smearing se obtiene mediante el comando:

```
mean(exp(residuals(m)))
```

donde `m` es el objeto que contiene el modelo ajustado y vale, en este ejemplo 1,059859. En comparación, el factor de corrección calculado anteriormente (Línea roja 35) fue 1,061035.

7.3. Predicción del volumen o de la biomasa de un rodal

Para predecir el volumen o la biomasa de un rodal con la ayuda de un modelo de biomasa, no es posible medir las entradas de ésta para todos los árboles del rodal. Las entradas sólo se medirán para una muestra de árboles del rodal. El volumen o la biomasa de los árboles de esta muestra se calculará con la ayuda del modelo, luego se extrapolará a todo el rodal. La predicción del volumen o de la biomasa de un rodal conlleva dos fuentes de variabilidad: una asociada a la predicción individual mediante el modelo, y la otra asociada al muestreo de los árboles dentro del rodal. Tener en cuenta rigurosamente ambas fuentes de variabilidad en la predicción a escala del rodal plantea problemas complejos de doble muestreo, que ya evocamos en los párrafos 2.1.2 y 2.3 (Parresol, 1999).

El problema es menos complejo cuando la muestra de los árboles usados para construir el modelo es independiente de la muestra de árboles medidos. En ese caso, se puede considerar que el error de predicción asociado a ese modelo es independiente del error de muestreo. Supongamos que n parcelas de ensayo de superficie unitaria A se hubieran colocado en el rodal, cuya superficie total es $\mathcal{A}$. Consideremos N_i el número de árboles encontrados en la i-ésima parcela ($i = 1, \ldots, n$) y consideremos también $X_{ij1}, \ldots, X_{ijp}$ las p variables explicativas medidas en el j-ésimo árbol de la i-ésima parcela ($j = 1, \ldots, N_i$). Cunia (1965, 1987b) consideró el caso particular en que la biomasa se predice mediante la regresión múltiple a partir de las p variables explicativas. La estimación de la biomasa del rodal es entonces:

$$\begin{aligned}\hat{B} &= \frac{\mathcal{A}}{n}\sum_{i=1}^{n}\frac{1}{A}\sum_{j=1}^{N_i}(\hat{a}_0 + \hat{a}_1 X_{ij1} + \ldots + \hat{a}_p X_{ijp}) \\ &= \hat{a}_0\Big(\frac{\mathcal{A}}{nA}\sum_{i=1}^{n} N_i\Big) + \hat{a}_1\Big(\frac{\mathcal{A}}{nA}\sum_{i=1}^{n}\sum_{j=1}^{N_i} X_{ij1}\Big) + \ldots + \hat{a}_p\Big(\frac{\mathcal{A}}{nA}\sum_{i=1}^{n}\sum_{j=1}^{N_i} X_{ijp}\Big)\end{aligned}$$

donde $\hat{a}_0, \ldots, \hat{a}_p$ son los coeficientes estimados de la regresión. Dado $X_0^\star = (\mathcal{A}/nA)\sum_{i=1}^{n} N_i$ y para todo $k = 1, \ldots, p$,

$$X_k^\star = \frac{\mathcal{A}}{nA}\sum_{i=1}^{n}\sum_{j=1}^{N_i} X_{ijk}$$

Entonces la biomasa estimada del rodal se escribe como:

$$\hat{B} = \hat{a}_0 X_0^\star + \hat{a}_1 X_1^\star + \ldots + \hat{a}_p X_p^\star$$

Lo que resulta interesante es que la variabilidad de $\hat{\mathbf{a}} = {}^{\mathrm{t}}[\hat{a}_0, \ldots, \hat{a}_p]$ depende totalmente del ajuste al modelo y no del muestreo del rodal, mientras que la variabilidad de $\mathbf{x} =$

${}^{\mathrm{t}}[X_0^\star, \ldots, X_p^\star]$ depende, por el contrario, completamente del muestreo y no del modelo. Dado que esos dos errores son independientes,

$$\mathrm{E}(\hat{B}) = \mathrm{E}({}^{\mathrm{t}}\hat{\mathbf{a}}\mathbf{x}) = {}^{\mathrm{t}}\mathrm{E}(\hat{\mathbf{a}})\,\mathrm{E}(\mathbf{x})$$

y

$$\mathrm{Var}(\hat{B}) = {}^{\mathrm{t}}\mathbf{a}\boldsymbol{\Sigma}_{\mathbf{x}}\mathbf{a} + {}^{\mathrm{t}}\mathbf{x}\boldsymbol{\Sigma}_{\mathbf{a}}\mathbf{x}$$

donde $\boldsymbol{\Sigma}_{\mathbf{a}}$ es la matriz $(p+1)\times(p+1)$ de varianza-covarianza de los coeficientes del modelo mientras que $\boldsymbol{\Sigma}_{\mathbf{x}}$ es la matriz $(p+1)\times(p+1)$ de varianza-covarianza de la muestra de $\mathbf{x}$. La primera matriz se deduce del ajuste del modelo mientras que la segunda se deriva del muestreo del rodal. De este modo, el error para la predicción de la biomasa del rodal se descompone en la suma de dos términos, de los cuáles uno está asociado al error de predicción del modelo y el otro al error de muestreo del rodal.

En líneas más generales, el principio es exactamente el mismo que cuando consideramos en la página 184 una incertidumbre asociada a la medición de las variables explicativas X_1, $\ldots$, X_p. Un error de medición no tiene el mismo carácter que un error de muestreo. Pero, desde un punto de vista matemático, los cálculos son los mismos: eso equivale a decir que en ambos casos hay que considerar que las variables explicativas X_1, $\ldots$, X_p son aleatorias en vez de fijas. Así pues, en el caso general, podremos usar un método de Montecarlo para estimar la biomasa del rodal. El seudoalgoritmo de este método de Montecarlo es igual al anterior (cf. p.185):

1. Para k que va de 1 a Q, donde Q es el número de iteraciones de Montecarlo:
   a) para i que va de 1 a p, escoger $\hat{X}_i^{(k)}$ que sigua una distribución que corresponde a la variabilidad de muestreo del rodal (esta distribución depende del tipo de muestreo realizado, del tamaño y del número de parcelas de ensayo inventariadas, etc.);
   b) escoger un vector $\hat{\theta}^{(k)}$ que sigua una distribución multinormal de media $\hat{\theta}$ y de matriz de varianza-covarianza $\hat{\boldsymbol{\Sigma}}$;
   c) calcular la predicción $\hat{Y}^{(k)} = f(\hat{X}_1^{(k)}, \ldots, \hat{X}_p^{(k)}; \hat{\theta}^{(k)})$.
2. El intervalo de confianza de la predicción es el intervalo de confianza empírico de los Q valores $\hat{Y}^{(1)}, \ldots, \hat{Y}^{(Q)}$.

7.4. Expansión y conversión de los modelos de volumen y biomasa

Puede que tengamos un modelo para predecir una magnitud que no es exactamente aquella que necesitamos aunque esté muy estrechamente vinculada con ella. Por ejemplo, disponemos de un modelo que predice la biomasa seca del tronco aunque lo que deseamos conocer es la biomasa total sobre el suelo del árbol. O bien, tenemos un modelo que predice el volumen del tronco cuando queremos conocer su biomasa seca. En vez de renunciar a usar un modelo que no predice exactamente lo que queremos, es preferible usarlo corrigiéndolo mediante un factor. Podemos utilizar factores de *conversión* para convertir un volumen en biomasa (y *vice versa*), factores de *expansión* para extrapolar una parte al todo o combinaciones de ambos. Con esto en mente es que Henry *et al.* (2011) proponen tres métodos para obtener la biomasa total:

- la biomasa del tronco es el producto del volumen del tronco y de la densidad específica de la madera ρ;

- la biomasa epigea es el producto de la biomasa del tronco y de un factor de expansión de la biomasa (FEB);
- la biomasa epigea es el producto del volumen del tronco y de un factor de conversión y de expansión de la biomasa (FCEB = FEB $\times\rho$).

Existen valores tabulados de estos diferentes factores de conversión y de expansión. Dichos valores suelen ser muy variables puesto que integran implícitamente diferentes fuentes de variabilidad. Por muy preciso que sea el modelo predeterminado, suele perderse el beneficio de esta precisión cuando se usa un factor de expansión o de conversión, ya que el error de la predicción acumula todas las fuentes de error que intervienen en su cálculo.

Para los modelos que usan la altura como entrada cuando no se dispone de esa información, se puede utilizar un modelo secundario que predice la altura en función de las entradas disponibles (típicamente un modelo de la relación altura-diámetro). Al igual que para los factores de conversión y de expansión, esto introduce una fuente de error adicional.

7.5. Seleccionar entre diferentes modelos

Cuando se quiere predecir el volumen o la biomasa de árboles dados, suele ocurrir que tengamos varios modelos a nuestra disposición. Por ejemplo, para una especie dada, se ajustaron diferentes modelos en distintos lugares. O bien, disponemos de un modelo local y de otro pantropical. Seleccionar entre los diferentes modelos disponibles no siempre es algo fácil (Henry *et al.*, 2011). ¿Es mejor, por ejemplo, elegir un modelo específico, local, ajustado a pocos datos (en consecuencia, *a priori* sin sesgo pero con una fuerte variabilidad de predicción) o bien un modelo multiespecífico pantropical ajustado a numerosos datos (en consecuencia, potencialmente con sesgo pero con poca variabilidad de predicción)? Esto demuestra que es posible tener en cuenta numerosos criterios de elección: la calidad del modelo (el tamaño de su ámbito de validez, su capacidad de extrapolar predicciones, etc.), su especificidad (con modelos monoespecíficos locales, en un extremo, y modelos pluriespecíficos pantropicales, en el otro), el tamaño del conjunto de datos usado para ajustar el modelo (entonces, implícitamente, la variabilidad de sus predicciones). La selección entre distintos modelos existentes no debe confundirse con la selección de modelos evocada en la Sección 6.3.2 donde a la hora de seleccionar los modelos, no se conocen aún los coeficientes de dichos modelos y se busca el modelo que se ajusta mejor a los datos cuando se estiman sus coeficientes. En este proceso de selección modelos, se trabaja con modelos ya ajustados cuyos coeficientes son conocidos.

Con frecuencia la selección entre diferentes modelos debe hacerse sin datos de biomasa o de volumen. Sin embargo, el caso que nos ocupa ahora es aquel en que se dispone de un conjunto de datos de referencia $\mathcal{S}_n$, con n observaciones de la variable de respuesta (volumen o biomasa) y de las variables explicativas.

7.5.1. Comparación de criterios de validación

Cuando se dispone de un conjunto de datos de referencia $\mathcal{S}_n$, se pueden comparar los distintos modelos disponibles basándose en criterios de validación definidos en el párrafo 7.1.1, usando $\mathcal{S}_n$ como conjunto de datos de validación. En la medida en que los modelos no tienen obligatoriamente el mismo número p de parámetros, y según el principio de parsimonia, favoreceremos los criterios de validación que dependen de p de forma tal que penalicemos los modelos que tengan muchos parámetros.

Cuando se trata de comparar un modelo candidato bien preciso, que se supone que es el "mejor", a diferentes modelos que compiten con ella, se podrá comparar las predicciones del modelo candidato a las predicciones de sus competidoras. Para ello, nos fijaremos en si las predicciones de los modelos competidores entran o no en el intervalo de confianza con nivel α de las predicciones del modelo candidato.

7.5.2. Elección de un modelo

La elección de un modelo puede hacerse con respecto a uno "verdadero" modelo f que no conocemos pero que suponemos que existe. Supongamos que M es el número de modelos de que disponemos. Escribiremos en fórmula abreviada $\hat{f}_m$ la función de las p variables explicativas que predicen el volumen o la biomasa, según el m-ésimo modelo. Esta función es aleatoria puesto que depende de los coeficientes estimados, es decir, los que tienen su propia distribución. La ley de distribución de $\hat{f}_m$ describe así la variabilidad de las predicciones en función de la m-ésimo modelo, tal como se la describe en el párrafo 7.2. Los M modelos pueden tener formas muy diferentes: puede ser que el modelo $\hat{f}_1$ corresponda a una función de potencia, el modelo $\hat{f}_2$ a una función polinomial, etc. Supongamos además que existe una función f de las p variables explicativas que describe la "verdadera" relación entre la variable de respuesta (volumen o biomasa) y esas variables explicativas. Desconocemos esta "verdadera" relación. No sabemos qué forma tiene pero cada una de los M modelos puede verse como una aproximación de la "verdadera" relación f.

En la teoría de la selección de modelos (Massart, 2007), la diferencia entre la verdadera relación f y un modelo $\hat{f}_m$ es cuantifica por una función γ que llamamos la función de *pérdida*. Por ejemplo, la función de pérdida podrá ser la norma L^2 de la diferencia entre f y $\hat{f}_m$:

$$\gamma(f,\, \hat{f}_m) = \int_{x_1} \dots \int_{x_p} [f(x_1,\, \dots,\, x_p) - \hat{f}_m(x_1,\, \dots,\, x_p)]^2 \, \mathrm{d}x_1 \dots \mathrm{d}x_p$$

Se llama *riesgo* (escrito R) la expectativa de pérdida con respecto a la ley de distribución de $\hat{f}_m$ cuando se integra sobre la variabilidad de las predicciones del modelo:

$$R = \mathrm{E}[\gamma(f,\, \hat{f}_m)]$$

El mejor modelo entre las M disponibles es la que minimiza el riesgo. El problema es que no se conoce la verdadera relación de la función f, así que también se desconoce ese modelo "mejor". En la teoría de selección de modelos, ese modelo mejor se llama *oracle*. El modelo elegido será finalmente aquel tal que el riesgo del oráculo quede limitado por una amplia familia de funciones f. En forma intuitiva, el modelo elegido es aquel en la que la diferencia entre ese modelo y la "verdadera" relación sigue siendo limitada, independientemente de cuál sea esa "verdadera" relación (dentro de los límites de una gama de posibilidades realistas). No seguiremos explayándonos sobre esta teoría porque excede el marco de nuestro manual.

7.5.3. Media bayesiana de modelos

En vez de escoger un modelo entre los M disponibles, con el riesgo de no elegir el "mejor", hay una alternativa que consiste en combinar los M modelos competidores en uno nuevo. Esto se llama en inglés "Bayesian model averaging". La media bayesiana de modelos se usó mucho para los modelos de predicción climática (Raftery *et al.*, 2005; Furrer *et al.*, 2007; Berliner & Kim, 2008; Smith *et al.*, 2009) pero sigue usándose todavía poco para los modelos forestales (Li *et al.*, 2008; Picard *et al.*, 2012). Consideremos $\mathcal{S}_n = \{(Y_i,\, X_{i1},\, \dots,\, X_{ip}),\, i = 1, \dots, p\}$ como un conjunto de datos de referencia con n observaciones de la variable de

respuesta Y y de las p explicativas. La media bayesiana de los modelos considera que la ley de distribución de la variable de respuesta Y es una mezcla de las distribuciones de M:

$$g(Y|X_1, \ldots, X_p) = \sum_{m=1}^{M} w_m \, g_m(Y|X_1, \ldots, X_p)$$

donde g es la densidad de distribución de Y, g_m es la densidad de distribución condicional de Y a sabiendas de que el modelo m es el "mejor", y w_m es el peso del m-ésimo modelo en la mezcla, que se puede interpretar como la probabilidad *a posteriori* de que el m-ésimo modelo sea el "mejor". Las probabilidades a *a posteriori* w_m reflejan la calidad del ajuste de los modelos a los datos y tienen una suma igual a un: $\sum_{m=1}^{M} w_m = 1$.

Como en la selección de modelos evocada en el párrafo anterior, la media bayesiana de los modelos supone que existe una "verdadera" relación (pero que sigue siendo desconocida) entre la variable de respuesta y las p variables explicativas, y que cada modelo se aleja de esta "verdadera" relación en función de una distribución normal de desviación estándar σ_m. En otras palabras, la densidad g_m es la densidad de la distribución normal de media $f_m(x_1, \ldots, x_p)$ y de desviación estándar σ_m, donde f_m es la función de las p variables correspondientes a la m-ésimo modelo. Así pues,

$$g(Y|X_1, \ldots, X_p) = \sum_{m=1}^{M} w_m \, \phi(Y; f_m(x_1, \ldots, x_p), \, \sigma_m)$$

donde $\phi(\cdot; \mu, \sigma)$ es la densidad de probabilidad de la distribución normal de esperanza σ. El modelo f_{moy} resultante de la combinación de M modelos competidores se define como la esperanza del modelo de mezcla, es decir:

$$f_{\text{moy}}(X_1, \ldots, X_p) = \mathrm{E}(Y|X_1, \ldots, X_p) = \sum_{m=1}^{M} w_m \, f_m(X_1, \ldots, X_p)$$

De este modo, el modelo resultante de la combinación de los M modelos competidoras es la media ponderada de estos M modelos, el peso del modelo m será la probabilidad *a posteriori* de que dicho modelo m sea el mejor. Asimismo podemos calcular la varianza de las predicciones según el modelo f_{moy} resultante de la combinación de los M modelos competidores:

$$\begin{aligned} \mathrm{Var}(Y|X_1, \ldots, X_p) \;=\; & \sum_{m=1}^{M} w_m \Big[f_m(X_1, \ldots, X_p) - \sum_{l=1}^{M} w_l \, f_l(X_1, \ldots, X_p) \Big]^2 \\ & + \sum_{m=1}^{M} w_m \, \sigma_m^2 \end{aligned}$$

El primer término corresponde a la varianza intermodelos y expresa la variabilidad de las predicciones de un modelo con respecto a otro. El segundo término corresponde a la varianza intramodelo y refleja el error condicional de predicción sabiendo que el modelo es el mejor.

Para poder usar el modelo f_{moy} en vez de los M modelos $f_1, \ldots, f_M$, quedan por estimar los pesos $w_1, \ldots, w_M$ y las desviaciones estándar intramodelo $\sigma_1, \ldots, \sigma_M$. Estos $2M$ parámetros se estiman a partir del conjunto de datos de referencia $\mathcal{S}_n$ usando un algoritmo EM (Dempster *et al.*, 1977; McLachlan & Krishnan, 2008). El algoritmo EM introduce las variables latentes z_{im} de forma que z_{im} es la probabilidad *a posteriori* de que el modelo m sea el mejor modelo para la observación i de $\mathcal{S}_n$. Las variables latentes z_{im} toman valores entre 0 y 1. El algoritmo EM es iterativo y alterno entre dos etapas en cada iteración: la etapa E (como "esperanza" o "expectativa") y la etapa M (como "maximización"). El algoritmo EM es el siguiente:

1. Elegir los valores iniciales $w_1^{(0)}, \ldots, w_M^{(0)}, \sigma_1^{(0)}, \ldots, \sigma_M^{(0)}$ de los $2M$ parámetros por estimar.

2. Alternar ambas etapas:

   a) etapa E: calcular el valor de z_{im} en la iteración j usando los valores de los parámetros en la iteración $j-1$:

   $$z_{im}^{(j)} = \frac{w_m^{(j-1)}\, \phi[Y_i; f_m(X_{i1}, \ldots, X_{ip}), \sigma_m^{(j-1)}]}{\sum_{k=1}^{M} w_k^{(j-1)}\, \phi[Y_i; f_k(X_{i1}, \ldots, X_{ip}), \sigma_k^{(j-1)}]}$$

   b) etapa M: estimar los parámetros en la iteración j utilizando como pesos los valores actuales de los z_{im}, es decir:

   $$\begin{aligned} w_m^{(j)} &= \frac{1}{n}\sum_{i=1}^{n} z_{im}^{(j)} \\ {\sigma_m^{(j)}}^2 &= \frac{\sum_{i=1}^{n} z_{im}^{(j)}[Y_i - f_m(X_{i1}, \ldots, X_{ip})]^2}{\sum_{i=1}^{n} z_{im}^{(j)}} \end{aligned}$$

   de forma que $\sum_{m=1}^{M} |w_m^{(j)} - w_m^{(j-1)}| + \sum_{m=1}^{M} |\sigma_m^{(j)} - \sigma_m^{(j-1)}|$ sea mayor que un umbral infinitesimal fijo (por ejemplo 10^{-6}).

3. El valor estimado de w_m es $w_m^{(j)}$ y el valor estimado de σ_m es $\sigma_m^{(j)}$.

Conclusiones y recomendaciones

Los métodos de estimación del volumen y de la biomasa de los árboles están en constante evolución. Cada vez más se quieren obtener estimaciones que sean lo más próximas posibles a la realidad. Los modelos de volumen y biomasa no han seguido la misma evolución en diferentes zonas ecológicas. En las zonas tropicales secas, donde el problema del suministro de leña es muy antiguo, las ecuaciones alométricas han sido elaboradas principalmente para cuantificar la leña. En la zona tropical muy húmeda, donde el aprovechamiento forestal se hace principalmente para obtener madera de construcción, las ecuaciones han sido elaboradas principalmente para calcular volumen. En la actualidad hay una preocupación cada vez mayor por el cambio climático y el interés despertado por los modelos de biomasa es similar en los bosques secos y húmedos.

Las mediciones de biomasa deberían aumentar en los años venideros para satisfacer las necesidades de estimación de las reservas de carbono y de comprensión de la contribución de los ecosistemas terrestres en el ciclo del carbono. La experiencia adquirida en la determinación del volumen ha demostrado que hacen falta entre dos y tres mil observaciones para estimar el volumen del tronco de *una* especie dada con una precisión aceptable para abarcar la variabilidad comprendida en su área geográfica de distribución (CTFT, 1989). En comparación, el modelo de biomasa de Chave *et al.* (2005), que es uno de los más usados actualmente, fue calibrado a partir de 2410 observaciones y se trata de un modelo pantropical que abarca todas las especies y todas las zonas ecológicas, desde las zonas secas a las zonas muy húmedas. La similitud entre estos dos tamaños de muestras, a pesar de que la variabilidad difiere en varias magnitudes, destaca que todavía hay un margen de progresión considerable en el ámbito de la medición de la biomasa, para llegar a explorar la totalidad de la variabilidad natural. A ello se suma el hecho de que la biomasa, que engloba todos los compartimientos del árbol, tiene probablemente una variabilidad intrínseca mucho mayor que el volumen de un solo tronco.

Para aumentar la fiabilidad de los modelos de biomasa hay que aumentar también el número de observaciones disponibles. Pero medir la biomasa epigea de un árbol exige un esfuerzo de medición mucho mayor que medir el volumen de su tronco. El esfuerzo necesario es aún mayor cuando se trata de la biomasa de las raíces. Actualmente es poco probable que puedan financiarse grandes campañas de medición para la biomasa epigea y radicular. Al igual que Chave *et al.* (2005), la construcción de nuevas ecuaciones alométricas tendrá que basarse en compilaciones de conjuntos de datos recopilados en distintos lugares por equipos independientes. Los métodos estandarizados para medir la biomasa y las estadísticas de ajuste de los modelos capaces de integrar la información complementaria *por medio* de covariables explicativas resultan pues cruciales para permitir avanzar en cuanto a la estimación de la biomasa de los árboles en los próximos años. Los experimentos con rodales regulares (efectos de la ontogenia, de la densidad de la plantación, de la fertilidad de los suelos o de la fertilización, de la silvicultura en general) facilitarán la construcción de estos modelos genéricos.

Al contrario de los manuales existentes, quisimos que el presente manual abarcara todo el proceso de construcción de una ecuación alométrica, desde el trabajo de campo a la

predicción, pasando por el ajuste del modelo. No obstante, no pretendemos haber cubierto todas las situaciones posibles. Muchos son los casos en los que es necesario elaborar métodos específicos. Los árboles grandes con aletones o contrafuertes, por ejemplo, plantean un reto para la predicción de su biomasa — comenzando por el hecho de que no se puede medir su diámetro a la altura del pecho, que es la primera variable de entrada de la mayoría de los modelos. Los árboles huecos, amates, bambú y las grandes epifitas, son algunas de las especies y particularidades que no permitirán el seguimiento de los métodos propuestos en este manual sin plantear problemas. Probablemente habrá que elaborar nuevos métodos dendrométricos para tratar esos casos específicos. El uso del modelado tridimensional, la fotogrametría, el radar y el láser, tanto en tierra como aerotransportados, serán instrumentos que facilitarán o revolucionarán los métodos de estimación de la biomasa y, quizás, remplazarán más adelante la motosierra y la báscula.

Asimismo la estadística es una ciencia en constante evolución. Una comparación del informe de Whraton & Cunia (1987) con los métodos de ajuste utilizados en la actualidad muestra el progreso realizado en el ámbito forestal con respecto al uso de métodos estadísticos cada vez más sofisticados, que intentamos presentar didácticamente en este manual. El hecho de tomar en cuenta la variabilidad entre fustes podría convertirse en algo común para el ajuste modelos de biomasa en el futuro.

La mejora de los métodos de medición y de ajuste de los modelos, el aumento de las mediciones de campo, sólo contribuirán a mejorar los procesos de investigación científicos y de estimación de la biomasa de los árboles si los modelos y los métodos producidos se ponen a disposición en forma transparente. Muchos datos permanecen en las bibliotecas y que nunca se publican en revistas científicas o en la Internet. Además, para un país que no dispone de datos de biomasa para algunas de sus regiones ecológicas, no es fácil tener acceso a los datos existentes en los países vecinos o en zonas ecológicas idénticas. Por ello alentamos a los representantes del sector forestal a identificar los datos ya disponibles para las zonas ecológicas o los países de particular interés. Los datos pueden integrarse en una base de datos y servir para identificar las lagunas. Una vez hecho esto, pueden realizarse las mediciones de campo usando los consejos y el hilo conductor propuestos en este manual.

Para poder seguir mejorando las estimaciones, hace falta instaurar un sistema para archivar los datos. Ese es el punto de partida para disponer de mejores estimaciones en el futuro. Un sistema adecuado permitiría reducir los esfuerzos de los futuros equipos para entender y recalcular las estimaciones existentes. Por otra parte, es importante crear métodos que sean coherentes a lo largo del tiempo. El manual propone distintos métodos de medición. Es preferible adoptar uno que pueda reproducirse y que sea menos dependiente de factores financieros, tecnológicos o humanos. En caso de que se elabore un método alternativo por motivos prácticos, habrá que indicarlo y ponerlo a disposición para permitir que el próximo manual pueda tomar más en cuenta la diversidad de las metodologías posibles. Por último, es preferible adoptar métodos simples y fáciles de reproducir.

Bibliografía

AFNOR. 1985. Bois – détermination de la masse volumique. Tech. Rep. NF B51-005, AFNOR. 65

AGO. 2002. Field measurement procedures for carbon accounting. Bush for Greenhouse Report 2, Australian Greenhouse Office, Canberra, Australia. 30

Akaike, H. 1974. A new look at the statistical model identification. *IEEE T. Automat. Contr.*, 19(6): 716–723. 156

Alder, D. 1980. *Estimation des volumes et accroissement des peuplements forestiers – Vol. 2. Étude et prévision de la production.* Études FAO : forêts No. 22/2. Rome, FAO. 194 pp. 26

Andrews, J.A. & Siccama, T.G. 1995. Retranslocation of calcium and magnesium at the heartwood-sapwood boundary of Atlantic white cedar. *Ecology*, 76(2): 659–663. 24

Araújo, T.M., Higuchi, N. & de Carvalho, J.A. 1999. Comparison of formulae for biomass content determination in a tropical rain forest site in the state of Pará, Brazil. *Forest Ecology and Management*, 117(1-3): 43–52. 106

Archibald, S. & Bond, W.J. 2003. Growing tall vs growing wide: tree architecture and allometry of *Acacia karroo* in forest, savanna, and arid environments. *Oikos*, 102(1): 3–14. 23

Assmann, E. 1970. *The Principles of Forest Yield Study.* Oxford, UK, Pergamon Press. 506 pp. 24, 26

Augusto, L., Meredieu, C., Bert, D., Trichet, P., Porté, A., Bosc, A., Lagane, F., Loustau, D., Pellerin, S., Danjon, F., Ranger, J. & Gelpe, J. 2008. Improving models of forest nutrient export with equations that predict the nutrient concentration of tree compartments. *Annals of Forest Science*, 65(8): 808. 24

Basuki, T.M., van Laake, P.E., Skidmore, A.K. & Hussin, Y.A. 2009. Allometric equations for estimating the above-ground biomass in tropical lowland *Dipterocarp* forests. *Forest Ecology and Management*, 257(8): 1684–1694. 106

Batho, A. & García, O. 2006. De Perthuis and the origins of site index: a historical note. *Forest Biometry, Modelling and Information Science*, 1: 1–10. 24

Becking, J.H. 1953. Einige gesichtspunkte für die durchführung von vergleichenden durchforstungsversuchen in gleichälterigen beständen. *In 11^{e} Congrès de l'Union Internationale des Instituts de Recherches Forestiers, Rome, 1953 : comptes rendus.* IUFRO, pp. 580–582. 218

Bellefontaine, R., Petit, S., Pain-Orcet, M., Deleporte, P. & Bertault, J.G. 2001. *Les arbres hors forêt : vers une meilleure prise en compte*. Cahier FAO Conservation No. 35. Rome, FAO. 214 pp. 33

Bergès, L., Nepveu, G. & Franc, A. 2008. Effects of ecological factors on radial growth and wood density components of sessile oak (*Quercus petraea* Liebl.) in Northern France. *Forest Ecology and Management*, 255(3-4): 567–579. 24, 27

Berliner, L.M. & Kim, Y. 2008. Bayesian design and analysis for superensemble-based climate forecasting. *Journal of Climate*, 21(9): 1891–1910. 191

Bloom, A.J., Chapin, F.S. & Mooney, H.A. 1985. Resource fimitation in plants–an economic analogy. *Annual Review of Ecology and Systematics*, 16: 363–392. 28

Bohlman, S. & O'Brien, S. 2006. Allometry, adult stature and regeneration requirement of 65 tree species on Barro Colorado Island, Panama. *Journal of Tropical Ecology*, 22(2): 123–136. 23

Bolker, B. 2008. *Ecological Models and Data in R*. Princeton, NJ, Princeton University Press. 183

Bontemps, J.D., Hervé, J.C. & Dhôte, J.F. 2009. Long-term changes in forest productivity: a consistent assessment in even-aged stands. *Forest Science*, 55(6): 549–564. 26

Bontemps, J.D., Hervé, J.C., Leban, J.M. & Dhôte, J.F. 2011. Nitrogen footprint in a long-term observation of forest growth over the twentieth century. *Trees–Structure and Function*, 25(2): 237–251. 26

Bormann, F.H. 1953. The statistical efficiency of sample plot size and shape in forest ecology. *Ecology*, 34(3): 474–487. 48, 49

Bouchon, J. 1974. Les tarifs de cubage. Tech. rep., ENGREF, Nancy, France. 31

Bouriaud, O., Leban, J.M., Bert, D. & Deleuze, C. 2005. Intra-annual variations in climate influence growth and wood density of Norway spruce. *Tree Physiology*, 25(6): 651–660. 27

Box, G.E.P. & Draper, N.R. 1987. *Empirical Model Building and Response Surfaces*. Wiley series in probability and mathematical statistics. New York, NY, Wiley. 669 pp. 41

Bozdogan, H. 1987. Model selection and Akaike's Information Criterion (AIC): The general theory and its analytical extensions. *Psychometrika*, 52(3): 345–370. 156

Bradley, P.N. 1988. Survey of woody biomass on farms in western Kenya. *Ambio*, 17(1): 40–48. 30

Brown, I.F., Martinelli, L.A., Thomas, W.W., Moreira, M.Z., Victoria, R.A. & Ferreira, C.A.C. 1995. Uncertainty in the biomass of Amazonian forests: An example from Rondônia, Brazil. *Forest Ecology and Management*, 75(1-3): 175–189. 40

Brown, S. 1997. *Estimating Biomass and Biomass Change of Tropical Forests: a Primer*. FAO Forestry Paper No. 134. Rome, FAO. 65 pp. 43, 106

Brown, S., Gillespie, A.J.R. & Lugo, A.E. 1989. Biomass estimation methods for tropical forests with applications to forest inventory data. *Forest Science*, 35(4): 881–902. 106, 125

Burdon, R.D., Kibblewhite, R.P., Walker, J.C.F., Megraw, E.R. & Cown, D.J. 2004. Juvenile versus mature wood: a new concept, orthogonal to corewood versus outerwood, with special reference to *Pinus radiata* and *P. taeda*. *Forest Science*, 50(4): 399–415. 27

Burnham, K.P. & Anderson, D.R. 2004. Multimodel inference: understanding AIC and BIC in model selection. *Sociol. Method. Res.*, 33(2): 261–304. 156

Burnham, K.P. & Anderson, D.R. 2002. *Model Selection and Multimodel Inference. A Practical Information-Theoretic Approach*. New York, NY, Springer Science+Business Media, Inc., 2nd edn. 488 pp. 156

Cailliez, F. 1980. *Forest volume estimation and yield prediction. Volume estimation*, *Études FAO forêts*, vol. 1. Rome, FAO. 98 pp. 31

Cairns, M.A., Brown, S., Helmer, E.H. & Baumgardner, G.A. 1997. Root biomass allocation in the world's upland forests. *Oecologia*, 111(1): 1–11. 28

Calama, R., Barbeito, I., Pardos, M., del Río, M. & Montero, G. 2008. Adapting a model for even-aged *Pinus pinea* L. stands to complex multi-aged structures. *Forest Ecology and Management*, 256(6): 1390–1399. 28

Cavaignac, S., Nguyen Thé, N., Melun, F. & Bouvet, A. 2012. élaboration d'un modèle de croissance pour l'Eucalyptus gundal. *FCBA INFO*, p. 16. 27

Charru, M., Seynave, I., Morneau, F. & Bontemps, J.D. 2010. Recent changes in forest productivity: An analysis of national forest inventory data for common beech (*Fagus sylvatica* L.) in north-eastern France. *Forest Ecology and Management*, 260(5): 864–874. 26

Chave, J., Andalo, C., Brown, S., Cairns, M.A., Chambers, J.Q., Eamus, D., Fölster, H., Fromard, F., Higuchi, N., Kira, T., Lescure, J.P., Nelson, B.W., Ogawa, H., Puig, H., Riéra, B. & Yamakura, T. 2005. Tree allometry and improved estimation of carbon stocks and balance in tropical forests. *Oecologia*, 145(1): 87–99. 106, 195

Chave, J., Coomes, D., Jansen, S., Lewis, S.L., Swenson, N.G. & Zanne, A.E. 2009. Towards a worldwide wood economics spectrum. *Ecology Letters*, 12(4): 351–366. 24

Chave, J., Riéra, B. & Dubois, M.A. 2001. Estimation of biomass in a neotropical forest of French Guiana: spatial and temporal variability. *Journal of Tropical Ecology*, 17(1): 79–96. 106

Chave, J., Condit, R., Aguilar, S., Hernandez, A., Lao, S. & Perez, R. 2004. Error propagation and scaling for tropical forest biomass estimates. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, 359(1443): 409–420. 40, 46, 51

Chave, J., Condit, R., Lao, S., Caspersen, J.P., Foster, R.B. & Hubbell, S.P. 2003. Spatial and temporal variation of biomass in a tropical forest: results from a large

census plot in Panama. *Journal of Ecology*, 91(2): 240–252. 48, 49, 51

Cochran, W.G. 1977. *Sampling Techniques*. Wiley Series in Probability and Mathematical Statistics. New York, NY, John Wiley & Sons, 3rd edn. 428 pp. 33, 38, 39, 42

Colin-Belgrand, M., Ranger, J. & Bouchon, J. 1996. Internal nutrient translocation in chesnut tree stemwood: III. Dynamics across an age series of *Castanea sativa* (Miller). *Annals of Botany*, 78(6): 729–740. 24

Cotta, H. 1804. *Principes fondamentaux de la science forestière*. Paris, Bouchard-Huzard. 495 pp. 31

Courbaud, B., Goreaud, F., Dreyfus, P. & Bonnet, F.R. 2001. Evaluating thinning strategies using a tree distance dependent growth model: some examples based on the CAPSIS software "uneven-aged spruce forests" module. *Forest Ecology and Management*, 145(1): 15–28. 28

Cressie, N. 1993. *Statistics for Spatial Data*. Wiley Series in Probability and Mathematical Statistics. New York, NY, John Wiley & Sons, 2nd edn. 900 pp. 48

CTFT. 1989. *Mémento du forestier*. Paris, France, Ministère de la Coopération et du Développement, 3rd edn. 1266 pp. 33, 40, 42, 43, 48, 195

Cunia, T. 1964. Weighted least squares method and construction of volume tables. *Forest Science*, 10(2): 180–191. 31, 125

Cunia, T. 1965. Some theory on reliability of volume estimates in a forest inventory sample. *Forest Science*, 11(1): 115–128. 188

Cunia, T. 1987a. Construction of tree biomass tables by linear regression techniques. *In* E.H. Whraton & T. Cunia, eds., *Estimating tree biomass regressions and their error. Proceedings of the workshop on tree biomass regression functions and their contribution to the error of forest inventory estimates, May 26–30, 1986, Syracuse, N.Y. – Part E.* Broomall, PA, USA, USDA Forest Service, Northeastern Forest Experiment Station, General Technical Report no. NE-117, pp. 27–36. 125

Cunia, T. 1987b. Error of forest inventory estimates: its main components. *In* E.H. Whraton & T. Cunia, eds., *Estimating tree biomass regressions and their error. Proceedings of the workshop on tree biomass regression functions and their contribution to the error of forest inventory estimates, May 26–30, 1986, Syracuse, N.Y. – Part E.* Broomall, PA, USA, USDA Forest Service, Northeastern Forest Experiment Station, General Technical Report no. NE-117, pp. 1–14. 34, 39, 46, 188

Cunia, T. 1987c. An optimization model for subsampling trees for biomass measurement. *In* E.H. Whraton & T. Cunia, eds., *Estimating tree biomass regressions and their error. Proceedings of the workshop on tree biomass regression functions and their contribution to the error of forest inventory estimates, May 26–30, 1986, Syracuse, N.Y. – Part E.* Broomall, PA, USA, USDA Forest Service, Northeastern Forest Experiment Station, General Technical Report no. NE-117, pp. 109–118. 34, 39, 46, 49

Cunia, T. 1987d. An optimization model to calculate the number of sample trees and plots.

In E.H. Whraton & T. Cunia, eds., *Estimating tree biomass regressions and their error. Proceedings of the workshop on tree biomass regression functions and their contribution to the error of forest inventory estimates, May 26–30, 1986, Syracuse, N.Y. – Part E.* Broomall, PA, USA, USDA Forest Service, Northeastern Forest Experiment Station, General Technical Report no. NE-117, pp. 15–24. 34, 39, 46, 49

Cunia, T. & Briggs, R.D. 1984. Forcing additivity of biomass tables: some empirical results. *Canadian Journal of Forest Research*, 14: 376–384. 171

Cunia, T. & Briggs, R.D. 1985a. Forcing additivity of biomass tables: use of the generalized least squares method. *Canadian Journal of Forest Research*, 15: 23–28. 171

Cunia, T. & Briggs, R.D. 1985b. Harmonizing biomass tables by generalized least squares. *Canadian Journal of Forest Research*, 15: 331–340. 174

de Vries, P.G. 1986. *Sampling Theory for Forest Inventory– A Teach-Yourself Course.* Berlin, Springer-Verlag. 399 pp. 38, 47

Dean, C. 2003. Calculation of wood volume and stem taper using terrestrial single-image close-range photogrammetry and contemporary software tools. *Silva Fennica*, 37(3): 359–380. 174

Dean, C. & Roxburgh, S. 2006. Improving visualisation of mature, high-carbon sequestering forests. *For. Biometry Model. Inform. Sci.*, 1: 48–69. 174

Dean, C., Roxburgh, S. & Mackey, B. 2003. Growth modelling of *Eucalyptus regnans* for carbon accounting at the landscape scale. *In* A. Amaro, D. Reed & P. Soares, eds., *Modelling Forest Systems.* Wallingford, UK, CAB International Publishing, pp. 27–39. 174

Deans, J.D., Moran, J. & Grace, J. 1996. Biomass relationships for tree species in regenerating semi-deciduous tropical moist forest in Cameroon. *Forest Ecology and Management*, 88(3): 215–225. 40

Decourt, N. 1973. Production primaire, production utile : méthodes d'évaluation, indices de productivité. *Ann. Sci. For.*, 30(3): 219–238. 25

Deleuze, C., Blaudez, D. & Hervé, J.C. 1996. Fitting a hyperbolic model for height versus girth relationship in spruce stands. Spacing effects. *Ann. Sci. For.*, 53(1): 93–111. 27

Dempster, A.P., Laird, N.M. & Rubin, D.B. 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, 39(1): 1–38. 192

Dhôte, J.F. 1990. *Modèles de la dynamique des peuplements forestiers : articulation entre les niveaux de l'arbre et du peuplement. Applications à la sylviculture des hêtraies.* Thèse de doctorat, Université Claude Bernard-Lyon I, Lyon, France. 27

Dhôte, J.F. 1991. Modélisation de la croissance des peuplements réguliers de hêtre : dynamique des hiérarchies sociales et facteurs de production. *Ann. Sci. For.*, 48(4): 389–416. 24

Dhôte, J.F. 1996. A model of even-aged beech stands productivity with process-based

interpretations. *Ann. Sci. For.*, 53(1): 1–20. 26

Díaz, S. & Cabido, M. 1997. Plant functional types and ecosystem function in relation to global change. *Journal of Vegetation Science*, 8: 463–474. 169

Dietz, J. & Kuyah, S. 2011. Guidelines for establishing regional allometric equations for biomass estimation through destructive sampling. Report of the carbon benefits project: Modelling, measurement and monitoring, World Agroforestry Centre (ICRAF), Nairobi, Kenya. 30

Dietze, M.C., Wolosin, M.S. & Clark, J.S. 2008. Capturing diversity and interspecific variability in allometries: A hierarchical approach. *Forest Ecology and Management*, 256(11): 1939–1948. 23

Djomo, A.N., Ibrahima, A., Saborowski, J. & Gravenhorst, G. 2010. Allometric equations for biomass estimations in Cameroon and pan moist tropical equations including biomass data from Africa. *Forest Ecology and Management*, 260(10): 1873–1885. 46, 106

Dong, J., Kaufmann, R.K., Myneni, R.B., Tucker, C.J., Kauppi, P.E., Liski, J., Buermann, W., Alexeyev, V. & Hughes, M.K. 2003. Remote sensing estimates of boreal and temperate forest woody biomass: carbon pools, sources, and sinks. *Remote Sens. Environ.*, 84: 393–410. 30

Dreyfus, P. 2012. Joint simulation of stand dynamics and landscape evolution using a tree-level model for mixed uneven-aged forests. *Annals of Forest Science*, 69(2): 283–303. 28

Duan, N. 1983. Smearing estimate: a nonparametric retransformation method. *Journal of the American Statistical Association*, 78(383): 605–610. 31, 187

Durbin, J. & Watson, G.S. 1971. Testing for serial correlation in least squares regression. III. *Biometrika*, 58(1): 1–19. 115

Ebuy Alipade, J., Lokombé Dimandja, J.P., Ponette, Q., Sonwa, D. & Picard, N. 2011. Biomass equation for predicting tree aboveground biomass at Yangambi, DRC. *Journal of Tropical Forest Science*, 23(2): 125–132. 40

Efron, B. & Tibshirani, R.J. 1993. *An Introduction to the Bootstrap*. Monographs on Statistics and Applied Probability No. 57. New York, NY, Chapman & Hall. 436 pp. 176, 177

Eichhorn, F. 1904. Beziehungen zwischen Bestandshöhe und Bestandsmasse. *Allgemeine Forst- und Jagdzeitung*, 80: 45–49. 25, 218

Enquist, B.J., Brown, J.H. & West, G.B. 1998. Allometric scaling of plant energetics and population density. *Nature*, 395(6698): 163–165. 23, 105

Enquist, B.J., West, G.B., Charnov, E.L. & Brown, J.H. 1999. Allometric scaling of production and life-history variation in vascular plants. *Nature*, 401(6756): 907–911. 23, 105

Enquist, B.J. 2002. Universal scaling in tree and vascular plant allometry: toward a

general quantitative theory linking plant form and function from cells to ecosystems. *Tree Physiology*, 22(15-16): 1045–1064. 24

Eyre, F.H. & Zillgitt, W.M. 1950. Size-class distribution in old-growth northern hardwoods twenty years after cutting. Station Paper 21, U.S. Department of Agriculture, Forest Service, Lake States Forest Experiment Station, Saint Paul, Minnessota, USA. 28

Fairfield Smith, H. 1938. An empirical law describing heterogeneity in the yields of agricultural crops. *Journal of Agricultural Science*, 28: 1–23. 48, 50

Fang, Z. & Bailey, R.L. 1999. Compatible volume and taper models with coefficients for tropical species on Hainan island in southern China. *Forest Science*, 45(1): 85–100. 174

FAO. 2006. *Global Forest Resources Assessment 2005. Progress towards sustainable forest management*, *FAO Forestry Paper*, vol. 147. Rome, Food and Agriculture Organization of the United Nations. 29

Favrichon, V. 1998. Modeling the dynamics and species composition of tropical mixed-species uneven-aged natural forest: effects of alternative cutting regimes. *Forest Science*, 44(1): 113–124. 28

Fonweban, J.N. & Houllier, F. 1997. Tarif de peuplement et modèle de production pour *Eucalyptus saligna* au Cameroun. *Bois et Forêts des Tropiques*, 253: 21–36. 96

Fournier-Djimbi, M. 1998. Le matériau bois : structure, propriétés, technologie. Cours, ENGREF, Département de foresterie rurale et tropicale, Montpellier, France. 65

Franc, A., Gourlet-Fleury, S. & Picard, N. 2000. *Introduction à la modélisation des forêts hétérogènes*. Nancy, France, ENGREF. 312 pp. 28, 105

Furnival, G.M. 1961. An index for comparing equations used in constructing volume tables. *Forest Science*, 7(4): 337–341. 128, 161

Furrer, R., Knutti, R., Sain, S.R., Nychka, D.W. & Meehl, G.A. 2007. Spatial patterns of probabilistic temperature change projections from a multivariate Bayesian analysis. *Geophysical Research Letters*, 34(L06711): 1–4. 191

Gambill, C.W., Wiant, H. V., J. & Yandle, D.O. 1985. Optimum plot size and BAF. *Forest Science*, 31(3): 587–594. 49, 50

García, O. 2003. Dimensionality reduction in growth models: an example. *Forest Biometry, Modelling and Information Science*, 1: 1–15. 26

García, O. 2011. Dynamical implications of the variability representation in site-index modelling. *Eur. J. For. Res.*, 130(4): 671–675. 24, 26

Gayon, J. 2000. History of the concept of allometry. *Am. Zool.*, 40(5): 748–758. 23

Gehring, C., Park, S. & Denich, M. 2004. Liana allometric biomass equations for Amazonian primary and secondary forest. *Forest Ecology and Management*, 195: 69–83. 32

Genet, A., Wernsdörfer, H., Jonard, M., Pretzsch, H., Rauch, M., Ponette, Q.,

Nys, C., Legout, A., Ranger, J., Vallet, P. & Saint-André, L. 2011. Ontogeny partly explains the apparent heterogeneity of published biomass equations for *Fagus sylvatica* in central Europe. *Forest Ecology and Management*, 261(7): 1188–1202. 28, 55

Gerwing, J.J., Schnitzer, S.A., Burnham, R.J., Bongers, F., Chave, J., DeWalt, S.J., Ewango, C.E.N., Foster, R., Kenfack, D., Martínez-Ramos, M., Parren, M., Parthasarathy, N., Pérez-Salicrup, D.R., Putz, F.E. & Thomas, D.W. 2006. A standard protocol for liana censuses. *Biotropica*, 38(2): 256–261. 32

Gerwing, J.J. & Farias, D.L. 2000. Integrating liana abundance and forest stature into an estimate of total aboveground biomass for an eastern Amazonian forest. *Journal of Tropical Ecology*, 16(3): 327–335. 32

Gibbs, H.K., Brown, S., Niles, J.O. & Foley, J.A. 2007. Monitoring and estimating tropical forest carbon stocks: making REDD a reality. *Environmental Research Letters*, 2(4): 1–13. Doi:10.1088/1748-9326/2/4/045023. 29

Gomat, H.Y., Deleporte, P., Moukini, R., Mialounguila, G., Ognouabi, N., Saya, R.A., Vigneron, P. & Saint-André, L. 2011. What factors influence the stem taper of *Eucalyptus*: growth, environmental conditions, or genetics? *Annals of Forest Science*, 68(1): 109–120. 27

Gonzalez, P., Asner, G.P., Battles, J.J., Lefsky, M.A., Waring, K.M. & Palace, M. 2010. Forest carbon densities and uncertainties from Lidar, QuickBird, and field measurements in California. *Remote Sens. Environ.*, 114(7): 1561–1575. 30

Gould, S.J. 1979. An allometric interpretation of species-area curves. The meaning of the coefficient. *American Naturalist*, 114(3): 335–343. 105

Gould, S.J. 1966. Allometry and size in ontogeny and phylogeny. *Biological Reviews*, 41(4): 587–638. 23

Gould, S.J. 1971. Geometric similarity in allometric growth: a contribution to the problem of scaling in the evolution of size. *American Naturalist*, 105(942): 113–136. 23

Goupy, J. 1999. *Plans d'expériences pour surfaces de réponse*. Paris, Dunod. 409 pp. 41

Gourlet-Fleury, S. & Houllier, F. 2000. Modelling diameter increment in a lowland evergreen rain forest in French Guiana. *Forest Ecology and Management*, 131(1-3): 269–289. 28

Gourlet-Fleury, S., Rossi, V., Rejou-Mechain, M., Freycon, V., Fayolle, A., Saint-André, L., Cornu, G., Gérard, J., Sarrailh, J.M., Flores, O., Baya, F., Billand, A., Fauvet, N., Gally, M., Henry, M., Hubert, D., Pasquier, A. & Picard, N. 2011. Environmental filtering of dense-wooded species controls above-ground biomass stored in African moist forests. *Journal of Ecology*, 99(4): 981–990. 28, 65

Gregoire, T.G. & Dyer, M.E. 1989. Model fitting under patterned heterogeneity of variance. *Forest Science*, 35(1): 105–125. 31

Guilley, E., Hervé, J.C. & Nepveu, G. 2004. The influence of site quality, silviculture and region on wood density mixed model in *Quercus petraea* Liebl. *Forest Ecology and*

Management, 189(1-3): 111–121. 24, 27

Hairiah, K., Sitompul, S.M., van Noordwijk, M. & Palm, C.A. 2001. *Methods for sampling carbon stocks above and below ground.* ASB Lecture Note No. 4B. Bogor, Indonesia, International Centre for Research in Agroforestry (ICRAF). 32 pp. 30

Härdle, W. & Simar, L. 2003. *Applied Multivariate Statistical Analysis.* Berlin, Springer-Verlag. 496 pp. 101

Hart, H.M.J. 1928. *Stamtal en dunning: een orienteerend onderzoek naar de beste plantwijdte en dunningswijze voor den djati.* Ph.D. thesis, Wageningen University, Wageningen, The Netherlands. 218

Hawthorne, W. 1995. *Ecological Profiles of Ghanaian Forest Trees.* Tropical Forestry Paper No. 29. Oxford, UK, Oxford Forestry Institute, Department of Plant Sciences, University of Oxford. 74

Hebert, J., Rondeux, J. & Laurent, C. 1988. Comparaison par simulation de 3 types d'unités d'échantillonnage en futaies feuillues de hêtre (*Fagus silvatica* l.). *Annales des Sciences Forestières*, 45(3): 209–221. 49

Henry, M., Besnard, A., Asante, W.A., Eshun, J., Adu-Bredu, S., Valentini, R., Bernoux, M. & Saint-André, L. 2010. Wood density, phytomass variations within and among trees, and allometric equations in a tropical rainforest of Africa. *Forest Ecology and Management*, 260(8): 1375–1388. 7, 8, 9, 13, 24, 32, 71, 88, 90, 91, 96, 97, 102, 103, 105, 106, 117, 118, 123, 124, 130, 131, 133, 140, 141, 142, 143, 158, 159, 160, 161, 164, 170, 180

Henry, M., Picard, N., Trotta, C., Manlay, R., Valentini, R., Bernoux, M. & Saint-André, L. 2011. Estimating tree biomass of sub-Saharan African forests: a review of available allometric equations. *Silva Fennica*, 45(3B): 477–569. 40, 105, 189, 190

Hitchcock, H.C.I. & McDonnell, J.P. 1979. Biomass measurement: a synthesis of the literature. *In Proceedings of IUFRO workshop on forest resource inventories, July 23-26, 1979.* Fort Collins, Colorado, USA, SAF-IUFRO, pp. 544–595. 31

Hoeting, J.A., Madigan, D., Raftery, A.E. & Volinsky, C.T. 1999. Bayesian model averaging: a tutorial. *Statistical Science*, 14(4): 382–417. 137

Hofstad, O. 2005. Review of biomass and volume functions for individual trees and shrubs in Southeast Africa. *Journal of Tropical Forest Science*, 17(1): 151–162. 105

Holmgren, P., Masakha, E.J. & Sjöholm, H. 1994. Not all African land is being degraded: a recent survey of trees on farms in Kenya reveals rapidly increasing forest resources. *Ambio*, 23(7): 390–395. 30

Huxley, J.S. 1924. Constant differential growth-ratios and their significance. *Nature*, 114: 895–896. 23

Ikonen, V.P., Kellomäki, S., Väisänenet, H. & Peltola, H. 2006. Modelling the distribution of diameter growth along the stem in Scots pine. *Trees–Structure and Function*, 20(3): 391–402. 27

Jackson, R.B., Canadell, J., Ehleringer, J.R., Mooney, H.A., Sala, O.E. & Schulze, E.D. 1996. A global analysis of root distributions for terrestrial biomes. *Oecologia*, 108(3): 389–411. 28

Jacobs, M.W. & Cunia, T. 1980. Use of dummy variables to harmonize tree biomass tables. *Canadian Journal of Forest Research*, 10: 483–490. 174

Johnson, F.A. & Hixon, H.J. 1952. The most efficient size and shape of plot to use for cruising in old growth Douglas-fir timber. *Journal of Forestry*, 50: 17–20. 48

Keller, M., Palace, M. & Hurtt, G. 2001. Biomass estimation in the Tapajos National Forest, Brazil. Examination of sampling and allometric uncertainties. *Forest Ecology and Management*, 154(3): 371–382. 48

Kelly, J.F. & Beltz, R.C. 1987. A comparison of tree volume estimation models for forest inventory. Research Paper SO-233, U.S. Department of Agriculture, Forest Service, Southern Forest Experiment Station, New Orleans, LA, USA. 31

Ketterings, Q.M., Coe, R., van Noordwijk, M., Ambagau, Y. & Palm, C.A. 2001. Reducing uncertainty in the use of allometric biomass equations for predicting above-ground tree biomass in mixed secondary forests. *Forest Ecology and Management*, 146(1-3): 199–209. 45, 106

King, D.A. 1996. Allometry and life history of tropical trees. *Journal of Tropical Ecology*, 12: 25–44. 23

Knapic, S., Louzada, J.L. & Pereira, H. 2011. Variation in wood density components within and between *Quercus faginea* trees. *Canadian Journal of Forest Research*, 41(6): 1212–1219. 24

Kozak, A. 1970. Methods for ensuring additivity of biomass components by regression analysis. *Forestry Chronicle*, 46(5): 402–405. 32

Lahti, T. & Ranta, E. 1985. The SLOSS principle and conservation practice: an example. *Oikos*, 44(2): 369–370. 49

Lanly, J.P. 1981. *Manuel d'inventaire forestier, avec références particulières aux forêts tropicales hétérogènes.* Études FAO : forêts No. 27. Rome, Italie, FAO. 208 pp. 47

Larson, P.R. 1963. Stem form development of forest trees. *For. Sci. Monog.*, 5: 1–42. 27

Lavorel, S. & Garnier, E. 2002. Predicting changes in community composition and ecosystem functioning from plant traits: revisiting the Holy Grail. *Functional Ecology*, 16: 545–556. 169

Lefsky, M.A., Cohen, W.B., Harding, D.J., Parker, G.G., Acker, S.A. & Gower, S.T. 2002. Lidar remote sensing of above-ground biomass in three biomes. *Global Ecol. Biogeogr.*, 11(5): 393–399. 30

Levillain, J., Thongo M'Bou, A., Deleporte, P., Saint-André, L. & Jourdan, C. 2011. Is the simple auger coring method reliable for below-ground standing biomass estimation in *Eucalyptus* forest plantations? *Annals of Botany*, 108(1): 221–230. 75, 77

Li, Y., Andersen, H.E. & McGaughey, R. 2008. A comparison of statistical methods for estimating forest biomass from light detection and ranging data. *Western Journal of Applied Forestry*, 23(4): 223–231. 191

Loetsch, F. & Haller, K.E. 1973. *Forest Inventory. Statistics of Forest Inventory and Information from Aerial Photographs*, vol. 1. Munchen, BLV Verlagsgesellschaft mbH. 436 pp. 47

Louppe, D., Koua, M. & Coulibaly, A. 1994. Tarifs de cubage pour *Afzelia africana* Smith en forêt de Badénou (nord Côte d'Ivoire). Tech. rep., Institut des Forêts (IDEFOR), département foresterie, Côte d'Ivoire. 96

MacDicken, K.G. 1997. A guide to monitoring carbon storage in forestry and agroforestry projects. Report of the forest carbon monitoring program, Winrock Internationl Institute for Agricultural Development, Arlington, VA, USA. 30

Magnus, J.R. & Neudecker, H. 2007. *Matrix Differential Calculus with Applications in Statistics and Econometrics*. Wiley series in probability and statistics. Chichester, UK, John Wiley and Sons, 3rd edn. 450 pp. 120, 126

Magnussen, S., Kleinn, C. & Picard, N. 2008a. Two new density estimators for distance sampling. *Eur. J. For. Res.*, 127(3): 213–224. 51

Magnussen, S., Picard, N. & Kleinn, C. 2008b. A gamma-poisson distribution of the point to the k nearest event distance. *Forest Science*, 54(4): 429–441. 51

Maguire, D.A. & Batista, J.L.F. 1996. Sapwood taper models and implied sapwood volume and foliage profiles for coastal Douglas-fir. *Canadian Journal of Forest Research*, 26: 849–863. 174

Maniatis, D., Saint-André, L., Temmerman, M., Malhi, Y. & Beeckman, H. 2011. The potential of using xylarium wood samples for wood density calculations: a comparison of approaches for volume measurements. *iForest – Biogeosci. For.*, 4: 150–159. 68

Manning, W.G. & Mullahy, J. 2001. Estimating log models: to transform or not to transform? *J. Health Econ.*, 20: 461–494. 187

Martinez-Yrizar, A., Sarukhan, J., Perez-Jimenez, A., Rincon, E., Maass, J.M., Solis-Magallanes, A. & Cervantes, L. 1992. Above-ground phytomass of a tropical deciduous forest on the coast of Jalisco, México. *Journal of Tropical Ecology*, 8: 87–96. 106

Massart, P. 2007. *Concentration Inequalities and Model Selection. École d'Été de Probabilités de Saint-Flour XXXIII – 2003*. Lecture Notes in Mathematics No. 1896. Berlin Heidelberg, Springer-Verlag. 335 pp. 191

McCarthy, M.C. & Enquist, B.J. 2007. Consistency between an allometric approach and optimal partitioning theory in global patterns of plant biomass allocation. *Functional Ecology*, 21(4): 713–720. 28

McLachlan, G.J. & Krishnan, T. 2008. *The EM Algorithm and Extensions*. Wiley Series in Probability and Statistics. Hoboken, NJ, John Wiley & Sons, 2nd edn. 360 pp. 192

Meredieu, C., Perret, S. & Dreyfus, P. 2003. Modelling dominant height growth: effect of stand density. *In* A. Amaro, D. Reed & P. Soares, eds., *Modelling Forest Systems. Proceedings of the IUFRO 4.01 and 4.11 Conference, Instituto Superior de Gestão and Instituto Superior de Agronomia, Sesimbra, Portugal, 2-5 June 2002*. Wallingford, UK, CAB International Publishing, pp. 111–121. 26

Metcalf, C.J.E., Clark, J.S. & Clark, D.A. 2009. Tree growth inference and prediction when the point of measurement changes: modelling around buttresses in tropical forests. *Journal of Tropical Ecology*, 25(1): 1–12. 174

Mokany, K., Raison, R.J. & Prokushkin, A.S. 2006. Critical analysis of root : shoot ratios in terrestrial biomes. *Global Change Biology*, 12(1): 84–96. 28

Monreal, C.M., Etchevers, J.D., Acosta, M., Hidalgo, C., Padilla, J., López, R.M., Jiménez, L. & Velázquez, A. 2005. A method for measuring above- and below-ground C stocks in hillside landscapes. *Can. J. Soil Sci.*, 85(Special Issue): 523–530. 30

Muller, K.E. & Stewart, P.W. 2006. *Linear Model Theory. Univariate, Multivariate and Mixed Models*. Wiley series in probability and statistics. Hoboken, NJ, John Wiley & Sons. 410 pp. 173

Muller-Landau, H.C., Condit, R.S., Chave, J., Thomas, S.C., Bohlman, S.A., Bunyavejchewin, S., Davies, S., Foster, R., Gunatilleke, S., Gunatilleke, N., Harms, K.E., Hart, T., Hubbell, S.P., Itoh, A., Kassim, A.R., Lafrankie, J.V., Lee, H.S., Losos, E., Makana, J.R., Ohkubo, T., Sukumar, R., Sun, I.f., Nur Supardi, M.N., Tan, S., Thompson, J., Valencia, R., Villa Muñoz, G., Wills, C., Yamakura, T., Chuyong, G., Dattaraja, H.S., Esufali, S., Hall, P., Hernandez, C., Kenfack, D., Kiratiprayoon, S., Suresh, H.S., Thomas, D., Vallejo, M.I. & Ashton, P. 2006. Testing metabolic ecology theory for allometric scaling of tree size, growth and mortality in tropical forests. *Ecology Letters*, 9(5): 575–588. 24, 106

Myers, R.H. & Montgomery, D.C. 2002. *Response Surface Methodology: Process and Product Optimization Using Designed Experiments*. Wiley series in probability and statistics. New York, NY, Wiley. 824 pp. 41

Namaalwa, J., Eid, T. & Sankhayan, P. 2005. A multi-species density-dependent matrix growth model for the dry woodlands of Uganda. *Forest Ecology and Management*, 213(1-3): 312–327. 28

Návar, J. 2009. Allometric equations for tree species and carbon stocks for forests of northwestern Mexico. *Forest Ecology and Management*, 257(2): 427–434. 106

Návar, J., Méndez, E. & Dale, V. 2002. Estimating stand biomass in the Tamaulipan thornscrub of northeastern Mexico. *Annals of Forest Science*, 59(8): 813–821. 31, 32

Navarro, M.N.V., Jourdan, C., Sileye, T., Braconnier, S., Mialet-Serra, I., Saint-André, L., Dauzat, J., Nouvellon, Y., Epron, D., Bonnefond, J.M., Berbigier, P., Rouzière, A., Bouillet, J.P. & Roupsard, O. 2008. Fruit development, not GPP, drives seasonal variation in NPP in a tropical palm plantation. *Tree Physiology*, 28(11): 1661–1674. 75

Nelson, B.W., Mesquita, R., Pereira, L.G., Garcia Aquino de Souza, J.S., Teixeira Batista, G. & Bovino Couto, L. 1999. Allometric regressions for improved estimate of secondary forest biomass in the central Amazon. *Forest Ecology and Management*, 117(1-3): 149 –167. 106

Ngomanda, A., Moundounga Mavouroulou, Q., Engone Obiang, N.L., Midoko Iponga, D., Mavoungou, J.F., Lépengué, N., Picard, N. & Mbatchi, B. 2012. Derivation of diameter measurements for buttressed trees, an example from Gabon. *Journal of Tropical Ecology*, 28(3): 299–302. 137

Nicolini, É., Chanson, B. & Bonne, F. 2001. Stem growth and epicormic branch formation in understorey beech trees (*Fagus sylvatica* L.). *Annals of Botany*, 87(6): 737–750. 27

Nogueira, E.M., Fearnside, P.M., Nelson, B.W., Barbosa, R.I. & Keizer, E.W.H. 2008. Estimates of forest biomass in the Brazilian Amazon: New allometric equations and adjustments to biomass from wood-volume inventories. *Forest Ecology and Management*, 256(11): 1853–1867. 106

Nogueira, E.M., Nelson, B.W. & Fearnside, P.M. 2006. Volume and biomass of trees in central Amazonia: influence of irregularly shaped and hollow trunks. *Forest Ecology and Management*, 227(1-2): 14–21. 32

Paine, C.E.T., Marthews, T.R., Vogt, D.R., Purves, D., Rees, M., Hector, A. & Turnbull, L.A. 2012. How to fit nonlinear plant growth models and calculate growth rates: an update for ecologists. *Method. Ecol. Evol.*, 3(2): 245–256. 183

Pardé, J. 1980. Forest biomass. *Forestry Abstracts*, 41(8): 343–362. 31

Pardé, J. & Bouchon, J. 1988. *Dendrométrie*. Nancy, France, ENGREF, 2nd edn. 328 pp. 26, 31, 35, 40, 41, 42, 43

Parresol, B.R. 1993. Modeling multiplicative error variance: an example predicting tree diameter from stump dimensions in baldcypress. *Forest Science*, 39(4): 670–679. 31

Parresol, B.R. 1999. Assessing tree and stand biomass: a review with examples and critical comparisons. *Forest Science*, 45(4): 573–593. 31, 32, 46, 125, 161, 171, 174, 176, 186, 188

Parresol, B.R. 2001. Additivity of nonlinear biomass equations. *Canadian Journal of Forest Research*, 31(5): 865–878. 31

Parresol, B.R. & Thomas, C.E. 1989. A density-integral approach to estimating stem biomass. *Forest Ecology and Management*, 26: 285–297. 174

Patenaude, G., Hill, R.A., Milne, R., Gaveau, D.L.A., Briggs, B.B.J. & Dawson, T.P. 2004. Quantifying forest above ground carbon content using LiDAR remote sensing. *Remote Sens. Environ.*, 93(3): 368–380. 30

Pearson, T. & Brown, S. 2005. Guide de mesure et de suivi du carbone dans les forêts et prairies herbeuses. Report, Winrock International, Arlington, VA, USA. 30

Peng, C. 2000. Growth and yield models for uneven-aged stands: past, present and future.

Forest Ecology and Management, 132(2-3): 259–279. 28

Perot, T., Goreaud, F., Ginisty, C. & Dhôte, J.F. 2010. A model bridging distance-dependent and distance-independent tree models to simulate the growth of mixed forests. *Annals of Forest Science*, 67(5): 502. 29

Philippeau, G. 1986. Comment interpréter les résultats d'une analyse en composantes principales ? Manuel de Stat-ITCF, Institut Technique des Céréales et des Fourrages (ITCF), Paris. 101

Picard, N. & Bar-Hen, A. 2007. Estimation of the density of a clustered point pattern using a distance method. *Environmental and Ecological Statistics*, 14(4): 341–353. 48, 51

Picard, N. & Favier, C. 2011. A point-process model for variance-occupancy-abundance relationships. *American Naturalist*, 178(3): 383–396. 48

Picard, N. & Franc, A. 2001. Aggregation of an individual-based space-dependent model of forest dynamics into distribution-based and space-independent models. *Ecological Modelling*, 145(1): 69–84. 28

Picard, N., Kouyaté, A.M. & Dessard, H. 2005. Tree density estimations using a distance method in mali savanna. *Forest Science*, 51(1): 7–18. 51

Picard, N., Sylla, M.L. & Nouvellet, Y. 2004. Relationship between plot size and the variance of the density estimator in West African savannas. *Canadian Journal of Forest Research*, 34(10): 2018–2026. 48

Picard, N., Henry, M., Mortier, F., Trotta, C. & Saint-André, L. 2012. Using Bayesian model averaging to predict tree aboveground biomass. *Forest Science*, 58(1): 15–23. 191

Picard, N., Yalibanda, Y., Namkosserena, S. & Baya, F. 2008. Estimating the stock recovery rate using matrix models. *Forest Ecology and Management*, 255(10): 3597–3605. 28

Ponce-Hernandez, R., Koohafkan, P. & Antoine, J. 2004. *Assessing carbon stocks and modelling win-win scenarios of carbon sequestration through land-use changes.* Rome, Food and Agriculture Organization of the United Nations (FAO). 156 pp. 30

Porté, A. & Bartelink, H.H. 2002. Modelling mixed forest growth: a review of models for forest management. *Ecological Modelling*, 150(1-2): 141–188. 28, 29

Press, W.H., Teukolsky, S.A., Vetterling, W.T. & Flannery, B.P. 2007. *Numerical Recipes: The Art of Scientific Computing.* Cambridge, UK, Cambridge University Press, 3rd edn. 1235 pp. 149

Preßler, M.R. 1864. *Das Gesetz der Stammbildung.* Leipzig, Germany, Arnoldische Buchhandlung. 153 pp. 218

Pretzsch, H. 2009. *Forest Dynamics, Growth and Yield: From Measurement to Model.* Berlin, Springer-Verlag. 664 pp. 24

Pukkala, T., Lähde, E. & Laiho, O. 2009. Growth and yieldmodels for unevensizedforest stands in Finland. *Forest Ecology and Management*, 258(3): 207–216. 28

Putz, F.E. 1983. Liana biomass and leaf area of a “tierra firme" forest in the Rio Negro Basin, Venezuela. *Biotropica*, 15(3): 185–189. 32

R Development Core Team. 2005. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. 21

Raftery, A.E., Gneiting, T., Balabdaoui, F. & Polakowski, M. 2005. Using Bayesian model averaging to calibrate forecast ensembles. *Monthly Weather Review*, 133(5): 1155–1174. 191

Reed, D.D. & Green, E.J. 1985. A method of forcing additivity of biomass tables when using nonlinear models. *Canadian Journal of Forest Research*, 15(6): 1184–1187. 32

Reinecke, L.H. 1933. Perfecting a stand-density index for even-aged forests. *J. Agr. Res.*, 46(7): 627–638. 218

Reyes, G., Brown, S., Chapman, J. & Lugo, A.E. 1992. Wood densities of tropical tree species. General Technical Report SO-88, USDA Forest Service, Southern Forest Experiment Station, New Orleans, Louisiana, USA. 65

Rivoire, M., Genet, A., Didier, S., Nys, C., Legout, A., Longuetaud, F., Cornu, E., Freyburger, C., Motz, A., Bouxiero, N. & Saint-André, L. 2009. Protocole d’acquisition de données volume-biomasse-minéralomasse, Bure. Rapport technique, INRA, Nancy, France. 55

Rösch, H., Van Rooyen, M.W. & Theron, G.K. 1997. Predicting competitive interactions between pioneer plant species by using plant traits. *Journal of Vegetation Science*, 8: 489–494. 169

Russell, C. 1983. *Nutrient cycling and productivity of native and plantation forests at Jari Florestal, Para, Brazil*. Ph.D. thesis, University of Georgia, Athens, GA, USA. 40

Rutishauser, E., Wagner, F., Herault, B., Nicolini, E.A. & Blanc, L. 2010. Contrasting above-ground biomass balance in a Neotropical rain forest. *Journal of Vegetation Science*, 21: 672–682. 51

Rykiel, E.J.J. 1996. Testing ecological models: the meaning of validation. *Ecological Modelling*, 90: 229–244. 175, 176

Saatchi, S.S., Houghton, R.A., Dos Santos Alvalá, R.C., Soares, J.V. & Yu, Y. 2007. Distribution of aboveground live biomass in the Amazon basin. *Global Change Biology*, 13(4): 816–837. 30

Saint-André, L., Laclau, J.P., Bouillet, J.P., Deleporte, P., Miabala, A., Ognouabi, N., Baillères, H., Nouvellon, Y. & Moukini, R. 2002a. Integrative modelling approach to assess the sustainability of the *Eucalyptus* plantations in Congo. *In* G. Nepveu, ed., *Connection between Forest Resources and Wood Quality: Modelling Approaches and Simulation Software. Proceedings of the Fourth workshop IUFRO S5.01.04, Harrison Hot Springs, British Columbia, Canada, September 8-15, 2002*. IUFRO, pp. 611–621. 26, 27

Saint-André, L., Laclau, J.P., Deleporte, P., Ranger, J., Gouma, R., Saya, A. & Joffre, R. 2002b. A generic model to describe the dynamics of nutrient concentrations within stemwood across an age series of a eucalyptus hybrid. *Annals of Botany*, 90(1): 65–76. 24, 60

Saint-André, L., Laclau, J.P., P., D., Gava, J.L., Gonçalves, J.L.M., Mendham, D., Nzila, J.D., Smith, C., du Toit, B., Xu, D.P., Sankaran, K.V., Marien, J.N., Nouvellon, Y., Bouillet, J.P. & R. 2008. Slash and litter management effects on *Eucalyptus* productivity: a synthesis using a growth and yield modelling approach. *In* E.K.S. Nambiar, ed., *Site Management and Productivity in Tropical Plantation Forests. Proceedings of Workshops in Piracicaba (Brazil) 22-26 November 2004 and Bogor (Indonesia) 6-9 November 2006*. Bogor, Indonesia, CIFOR, pp. 173–189. 26

Saint-André, L., Leban, J.M., Houllier, F. & Daquitaine, R. 1999. Comparaison de deux modèles de profil de tige et validation sur un échantillon indépendant. Application à l'épicéa commun dans le nord-est de la France. *Annals of Forest Science*, 56(2): 121–132. 27

Saint-André, L., Thongo M'Bou, A., Mabiala, A., Mouvondy, W., Jourdan, C., Roupsard, O., Deleporte, P., Hamel, O. & Nouvellon, Y. 2005. Age-related equations for above- and below-ground biomass of a *Eucalyptus* hybrid in Congo. *Forest Ecology and Management*, 205(1-3): 199–214. 31, 43, 55, 152, 168

Saporta, G. 1990. *Probabilités, analyse des données et statistique*. Paris, Technip. 493 pp. 36, 38, 41, 47, 177, 178, 179, 181, 183, 186

Savage, V.M., Deeds, E.J. & Fontana, W. 2008. Sizing up allometric scaling theory. *PLoS Computational Biology*, 4(9): e1000171. 28

Schlaegel, B.E. 1982. Testing, reporting, and using biomass estimation models. *In* C.A. Gresham, ed., *Proceedings of the 3rd Annual Southern Forest Biomass Workshop*. Clemson, SC, Belle W. Baruch Forest Science Institute, Clemson University, pp. 95–112. 176

Schnitzer, S.A., DeWalt, S.J. & Chave, J. 2006. Censusing and measuring lianas: a quantitative comparison of the common methods. *Biotropica*, 38(5): 581–591. 32

Schnitzer, S.A., Rutishauser, S. & Aguilar, S. 2008. Supplemental protocol for liana censuses. *Forest Ecology and Management*, 255: 1044–1049. 32

Schreuder, H.T., Banyard, S.G. & Brink, G.E. 1987. Comparison of three sampling methods in estimating stand parameters for a tropical forest. *Forest Ecology and Management*, 21(1-2): 119–127. 49

Schreuder, H.T., Gregoire, T.G. & Wood, G.B. 1993. *Sampling methods for multiressource forest inventory*. New York, NY, Wiley & Sons. 446 pp. 47, 51

Serfling, R.J. 1980. *Approximation Theorems of Mathematical Statistics*. Wiley Series in Probability and Statistics. New York, NY, John Wiley & Sons. 371 pp. 181, 185, 187

Shaw, J.D. 2006. Reineke's stand density index: Where are we and where do we go from here? *In Driving Changes in Forestry. Proceedings of the Society of American Foresters*

2005 National Convention, October 19-23, 2005, Fort Worth, Texas, USA. Bethesda, MD, USA, Society of American Foresters, pp. 1–13. 26

Shinozaki, K., Yoda, K., Hozumi, K. & Kira, T. 1964a. A quantitative analysis of plant form - the pipe model theory. I. Basic analyses. *Japanese Journal of Ecology*, 14: 97–104. 23, 27

Shinozaki, K., Yoda, K., Hozumi, K. & Kira, T. 1964b. A quantitative analysis of plant form - the pipe model theory. II. Further evidence of the theory and its application on forest ecology. *Japanese Journal of Ecology*, 14: 133–139. 23, 27

Shiver, B.D. & Borders, B.E. 1996. *Sampling techniques for forest ressource inventory*. New York, NY, Wiley & Sons. 356 pp. 38, 39, 47

Sillett, S.C., Van Pelt, R., Koch, G.W., Ambrose, A.R., Carroll, A.L., Antoine, M.E. & Mifsud, B.M. 2010. Increasing wood production through old age in tall trees. *Forest Ecology and Management*, 259(5): 976–994. 174

Skovsgaard, J.P. & Vanclay, J.K. 2008. Forest site productivity: a review of the evolution of dendrometric concepts for even-aged stands. *Forestry*, 81(1): 13–31. 24, 26

Smith, R.L., Tebaldi, C., Nychka, D. & Mearns, L.O. 2009. Bayesian modeling of uncertainty in ensembles of climate models. *Journal of the American Statistical Association*, 104(485): 97–116. 191

Soares, P. & Tomé, M.. 2002. Height-diameter equation for first rotation eucalypt plantations in Portugal. *Forest Ecology and Management*, 166(1-3): 99–109. 27

St.-Onge, B., Hu, Y. & Vega, C. 2008. Mapping the height and above-ground biomass of a mixed forest using lidar and stereo Ikonos images. *Int. J. Remote Sens.*, 29(5): 1277–1294. 30

Stoyan, D. & Stoyan, H. 1994. *Fractals, Random Shapes and Point Fields*. Chichester, UK, John Wiley & Sons. 390 pp. 48

Tateno, R., Hishi, T. & Takeda, H. 2004. Above- and belowground biomass and net primary production in a cool-temperate deciduous forest in relation to topographical changes in soil nitrogen. *Forest Ecology and Management*, 193(3): 297–306. 28

Taylor, J.M.G. 1986. The retransformed mean after a fitted power transformation. *Journal of the American Statistical Association*, 81: 114–118. 31, 187

Tedeschi, L.O. 2006. Assessment of the adequacy of mathematical models. *Agricultural Systems*, 89(2-3): 225–247. 176

Thompson, S.K. 1992. *Sampling*. Wiley Series in Probability and Mathematical Statistics. New York, NY, John Wiley & Sons. 343 pp. 38, 39

Thornley, J.H. 1972. A balanced quantitative model for root: shoot ratios in vegetative plants. *Annals of Botany*, 36(2): 431–441. 28

Tomé, M., Barreiro, S., Paulo, J.A. & Tomé, J. 2006. Age-independent difference

equations for modelling tree and stand growth. *Canadian Journal of Forest Research*, 36(7): 1621–1630. 28

Valinger, E. 1992. Effects of thinning and nitrogen fertilization on stem growth and stem form of *Pinus sylvestris* trees. *Scandinavian Journal of Forest Research*, 7(1-4): 219–228. 27

Vallet, P., Dhôte, J.F., Le Moguédec, G., Ravart, M. & Pignard, G. 2006. Development of total aboveground volume equations for seven important forest tree species in France. *Forest Ecology and Management*, 229(1-3): 98–110. 27

Vallet, P. & Pérot, T. 2011. Silver fir stand productivity is enhanced when mixed with Norway spruce: evidence based on large-scale inventory data and a generic modelling approach. *Journal of Vegetation Science*, 22(5): 932–942. 28

van Breugel, M., Ransijn, J., Craven, D., Bongers, F. & Hall, J.S. 2011. Estimating carbon stock in secondary forests: Decisions and uncertainties associated with allometric biomass models. *Forest Ecology and Management*, 262(8): 1648–1657. 40, 43, 45, 46, 51

Van Pelt, R. 2001. *Forest Giants of the Pacific Coast*. Vancouver, Canada, Global Forest Society. 174

Vanclay, J.K. 1994. *Modelling Forest Growth and Yield – Applications to Mixed Tropical Forests*. Wallingford, UK, CAB International Publishing. 312 pp. 28

Vanclay, J.K. 2009. Tree diameter, height and stocking in even-aged forests. *Annals of Forest Science*, 66(7): 702. 26

Verzelen, N., Picard, N. & Gourlet-Fleury, S. 2006. Approximating spatial interactions in a model of forest dynamics as a means of understanding spatial patterns. *Ecological Complexity*, 3(3): 209–218. 28

Violle, C., Navas, M.L., Vile, D., Kazakou, E., Fortunel, C., Hummel, I. & Garnier, E. 2007. Let the concept of trait be functional! *Oikos*, 116: 882–892. 169

Wagner, F., Rutishauser, E., Blanc, L. & Herault, B. 2010. Effects of plot size and census interval on descriptors of forest structure and dynamics. *Biotropica*, 42(6): 664–671. 48, 49, 51

Weiskittel, A.R., Hann, D.W., Hibbs, D.E., Lam, T.Y. & Bluhm, A.A. 2009. Modeling top height growth of red alder plantations. *Forest Ecology and Management*, 258(3): 323–331. 26

West, G.B., Brown, J.H. & Enquist, B.J. 1997. A general model for the origin of allometric scaling laws in biology. *Science*, 276: 122–126. 23, 105

West, G.B., Brown, J.H. & Enquist, B.J. 1999. A general model for the structure and allometry of plant vascular systems. *Nature*, 400(6745): 664–667. 23, 28, 105

West, P.W. 2009. *Tree and Forest Measurement*. Berlin, Springer-Verlag, 2nd edn. 191 pp. 47, 51

White, J.F. & Gould, S.J. 1965. Interpretation of the coefficient in the allometric

equation. *American Naturalist*, 99(904): 5–18. 23

Whraton, E.H. & Cunia, T., eds. 1987. *Estimating tree biomass regressions and their error. Proceedings of the workshop on tree biomass regression functions and their contribution to the error of forest inventory estimates, May 26–30, 1986, Syracuse, N.Y. – Part E*, General Technical Report no. NE-117. Broomall, PA, USA, USDA Forest Service, Northeastern Forest Experiment Station. 51, 125, 196

Yamakura, T., Hagihara, A., Sukardjo, S. & Ogawa, H. 1986. Tree size in a mature dipterocarp forest stand in Sebulu, East Kalimantan, Indonesia. *Southeast Asian Studies*, 23(4): 452–478. 106

Zeide, B. 1980. Plot size optimization. *Forest Science*, 26(2): 251–257. 49, 50

Zianis, D. & Mencuccini, M. 2004. On simplifying allometric analyses of forest biomass. *Forest Ecology and Management*, 187(2-3): 311–332. 24, 175

Zianis, D., Muukkonen, P., Mäkipää, R. & Mencuccini, M. 2005. *Biomass and Stem Volume Equations for Tree Species in Europe.* Silva Fennica Monographs No. 4. Vantaa, Finland, The Finnish Society of Forest Science and The Finnish Forest Research Institute. 63 pp. 24, 40, 105

Glosario

En el presente glosario damos la definición de algunos términos técnicos inusuales o que se utilizan en este manual con una acepción diferente de la habitual.

Aditividad. Propiedad de un sistema de ecuaciones alométricas ajustadas a las diferentes partes del árbol y al árbol en su totalidad, tal que la suma de las predicciones para cada compartimiento corresponde realmente a la predicción para todo el árbol.

Alícuota. Parte extraída de un compartimiento del árbol cuya medición sirve para medir la totalidad del compartimiento mediante la regla de tres.

Alometría. Relación estadística, a escala de una población, entre dos características de tamaño de los individuos de dicha población. Esta relación suele ser una forma de potencia. Ejemplo: hay una alometría en los vertebrados entre la masa del cuerpo adulto y el tamaño del cerebro.

Biomasa. Masa de la materia orgánica viva o muerta de un organismo, expresada en masa de materia seca. Para un árbol, la unidad de medida es el kg o sus múltiplos. Por extensión, la biomasa de una zona es la suma de las biomasas de los organismos que se encuentran en esa zona. La unidad de medida es pues el kg (o sus múltiplos) por unidad de superficie.

Ruido blanco, error estadístico. En probabilidades, el proceso aleatorio que genera variables aleatorias que son todas independientes entre sí.

Compartimiento. Una parte de un árbol, generalmente determinada de forma tal que los órganos de un compartimiento tengan densidades (relación biomasa seca sobre volumen fresco) parecidas. El follaje, el tronco, las ramas grandes, etc., son compartimientos.

Covarianza. Cantidad que mide la variación simultánea de dos variables aleatorias. La covarianza se vuelve más positiva para cada par de valores que difieren de su media en el mismo sentido, y más negativa para cada par de valores que difieren de su media en el sentido opuesto. La covarianza de una variable aleatoria y de esta misma variable aleatoria da la varianza.

Doblete. Colección de dos valores numéricos.

Fractal. Objeto cuya estructura es invariante al cambio de escala.

Hart-Becking (índice de). En ciencias forestales, el índice establecido por Hart (1928) y Becking (1953) que mide la densidad de un rodal a partir de la distancia media entre los árboles y la altura dominante del rodal. Este índice se calcula como derazón entre el espaciamiento promedio entre árboles sobre la altura dominante, multiplicado por 100.

Heterocedasticidad. Es lo opuesto de la homocedasticidad, es decir, cuando la varianza del error residual de un modelo no es constante (y típicamente varía con una de las variables explicativas del modelo).

Homocedasticidad. Cuando la varianza del error residual de un modelo es constante. La homocedasticidad es una de las condiciones necesarias para el ajuste de un modelo lineal.

Ley de Eichhorn. En ciencias forestales es la ley empírica enunciada por Eichhorn (1904) que establece que el volumen específico de una masa homogénea, monoespecífica y de dosel cerrado, sólo es función de su altura dominante. Se trata de la segunda ley de Eichhorn; la primera establece que la altura dominante de una masa homogénea, monoespecífica y de dosel cerrado sólo es función de la edad, de la especie y de las condiciones del sitio.

Ley de Pressler. En ciencias forestales es la ley empírica enunciada por Preßler (1864) que estipula que el incremento en área basal es constante desde el tocón del árbol hasta la base de la porción funcional de la copa.

Mineralomasa. Cantidad de elementos minerales en la biomasa.

Distribución de Dirac. Distribución (en el sentido estadístico del término) concentrada en un valor x_0 de una variable aleatoria continua (es decir que la probabilidad de que la variable aleatoria sea $< x$ vale 0 para $x < x_0$ y 1 para $x > x_0$).

Método de Montecarlo. Dícese de un método que tiene por objeto calcular un valor numérico mediante la simulación de un proceso aleatorio.

Posición social. Para un árbol, la posición de su copa en el dosel, que determina su jerarquía con respecto a la competencia por la luz (también se habla de clasificación sociológica). Se suelen distinguir los árboles dominantes, los codominantes y los suprimidos.

Índice de densidad de Reinecke (IDR). En ciencias forestales es el índice establecido por Reinecke (1933) que mide la densidad de un rodal a partir del número de árboles por hectárea (la densidad del rodal) y el área basal media promedio de los árboles (diámetro cuadrático medio). Este índice se calcula como la relación de la densidad del rodal sobre la densidad máxima, determinada a partir del diámetro cuadrático medio por la curva de auto-raleo.

Variable ordinal. Variable que toma valores discretos y permite ordenarlos de acuerdo con sus modalidades. Por ejemplo, el mes del año es una variable ordinal (los meses pueden colocarse en orden cronológico).

Varianza. Cantidad que mide la dispersión de una variable aleatoria con respecto a su valor promedio. Se la calcula como el promedio de las desviaciones con la media, elevadas al cuadrado.

Léxico de símbolos matemáticos

Símbolos latinos

a valor estimado de un coeficiente de un modelo predictivo

A superficie de una parcela de ensayo

$\mathcal{A}$ superficie del rodal

b valor estimado de un coeficiente de un modelo predictivo

B biomasa de una alícuota, de una parte (tronco, ramas, follaje, etc.), de un árbol o de un rodal

CV_X coeficiente de variación de una magnitud X

c exponente de una ley de potencia

C definición 1: circunferencia de un árbol; definición 2: costo del muestreo; definición 3: un criterio de validación de un modelo

D diámetro de un árbol

D_0 diámetro dominante del rodal

E precisión de la estimación de una magnitud estimada

f una función que asocia una variable de respuesta a una o varias variables explicativas

F índice de Furnival

g una función

G área basal de un árbol o de un rodal

h una altura entre cero (el suelo) y la altura H del árbol

H altura de un árbol

H_0 altura dominante del rodal

$\mathbf{I}_n$ matriz de información de Fisher para una muestra del tamaño n

k coeficiente multiplicador de una ley de potencia

K número de partes para una validación cruzada

ℓ verosimilitud de una muestra

$\mathcal{L}$ logaritmo de verosimilitud o log-verosimilitud de una muestra

L longitud de una troza

M definición 1: número de compartimientos de biomasa en un árbol; definición 2: número de modelos concurrentes que predicen una misma variable de respuesta

n tamaño de una muestra

N definición 1: número total de unidades de muestreo (árbol o parcela) en el rodal; definición 2: densidad de un rodal (número de pies por hectárea)

$\mathcal{N}$ la distribución normal (también llamada distribución de Gauss o distribución gaussiana)

p número de variables explicativas de un modelo (intersección no incluida)

P perfil de tronco (curva que da la superficie de la sección del tronco en función de la altura)

q definición 1: número de parámetros estimados de un modelo; definición 2: cuantile de la distribución normal centrada y reducida

Q número de iteraciones de Montecarlo

R definición 1: coeficiente de determinación de un modelo; definición 2 (en la teoría de selección de modelo): un riesgo; definición 3: radio de una troza

S número de estratos de una estratificación

S_X desviación estándar empírica de una variable X

$\mathcal{S}_n$ un conjunto de datos que contiene n observaciones

t_n cuantile de una ley de Student a n grados de libertad

T edad de una plantación

V volumen de una troza, de un árbol o de un rodal

w definición 1: peso de una observación en la regresión ponderada; definición 2: peso de un modelo en una mezcla de modelos

X una variable (en general variable explicativa de un modelo)

$\mathbf{x}$ un vector de variables explicativas

$\mathbf{X}$ matriz del plano para un modelo lineal

Y una variable (en general variable de respuesta de un modelo)

$\mathbf{Y}$ vector de respuesta de un modelo multivariado

z una variable latente para el algoritmo EM

Z una variable (en general une covariable que define una estratificación del conjunto de datos)

Símbolos griegos

α definición 1: valor "verdadero" (desconocido) de un coeficiente de un modelo predictivo; definición 2: umbral de confianza de un intervalo de confianza (generalmente 5 %)

β valor "verdadero" (desconocido) de un coeficiente de un modelo predictivo

γ función de pérdida (en la teoría de la selección de modelo)

δ distribución de Dirac

Δ una diferencia de valor para una magnitud dada

ε error residual de un modelo predictivo

$\boldsymbol{\varepsilon}$ vector de los errores residuales de un modelo multivariado

ζ covarianza residual entre dos compartimientos

η coeficiente de contracción volumétrica

θ un conjunto de parámetros de un modelo

$\boldsymbol{\theta}$ un vector de parámetros de un modelo multivariado

ϑ un conjunto de parámetros

μ esperanza de una variable aleatoria = media "verdadera" (desconocida) de una magnitud por estimar

ξ parámetro de transformación de Box-Cox

ρ densidad de la madera

σ desviación estándar del error residual de un modelo predictivo

$\boldsymbol{\Sigma}$ matriz de varianza-covarianza de una distribución multinormal (también llamada distribución normal multivariante)

τ desviación estándar "verdadera" (desconocida) de una magnitud por estimar

ϕ densidad de probabilidad de la distribución normal

ψ función que define una transformación de variable

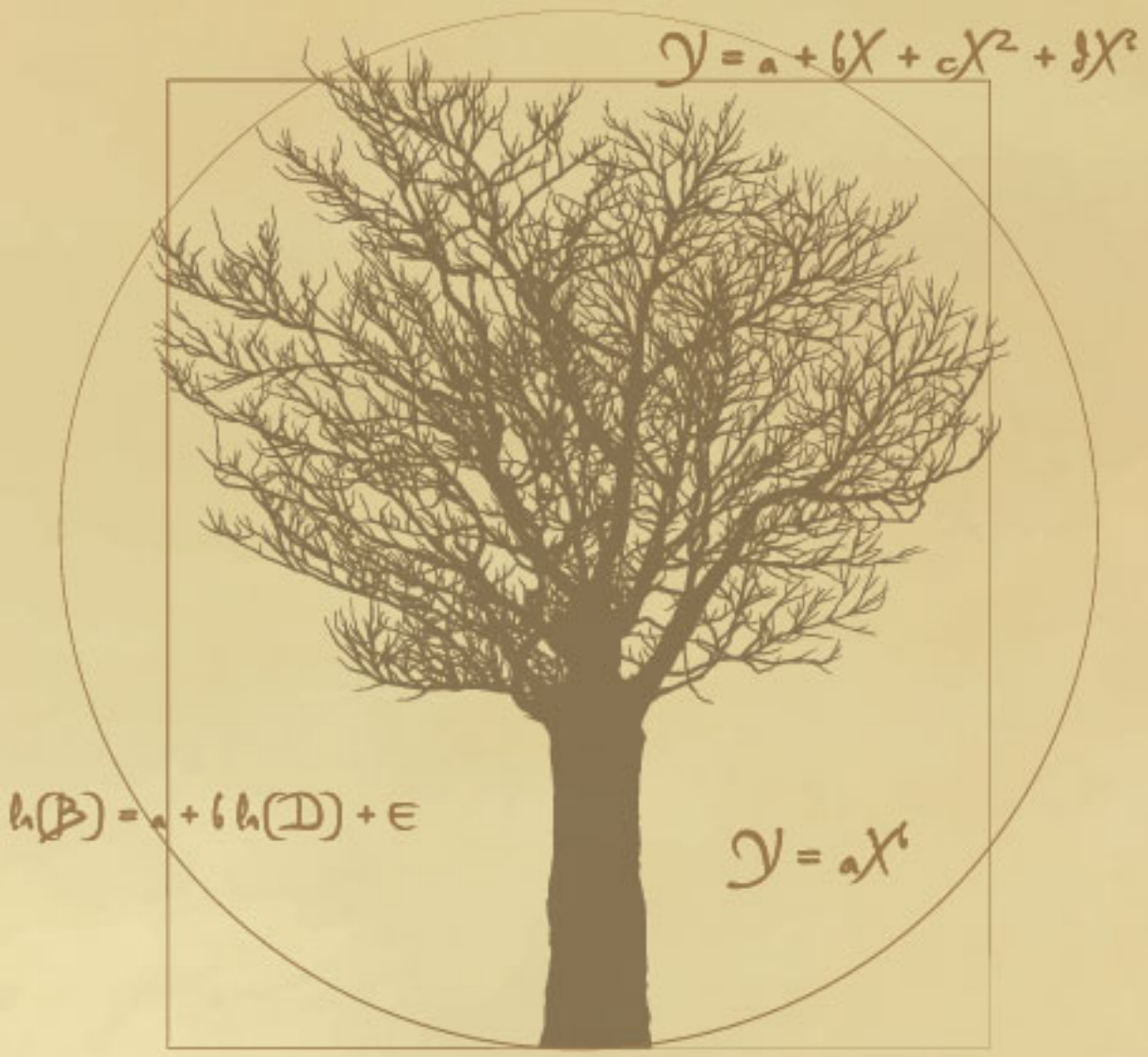
χ contenido de humedad

χ_0 punto de saturación de las fibras

ω proporción (por ejemplo, la proporción en biomasa fresca de la madero de una troza)

Símbolos no alfabéticos

$\varnothing$ diámetro de un árbol, una troza, una rama o una raíz



E-ISBN 978-92-5-307347-4