Modelos Lineales Aplicados en R

Juan Aparicio, Mª Asunción Martínez Mayoral y Javier Morales
Depto. Estadística, Matemáticas e Informática
Centro de Investigación Operativa
Universidad Miguel Hernández

Índice general

Índice de figuras

Índice de tablas

Tema 1

Introducción

1.1. Introducción

Las siguientes secciones están basadas fundamentalmente en los textos de Trívez(2004), Álvarez(1999), Peña(1995) y Stanton(2001).

1.1.1. Qué es la Econometría

Desde un punto de vista etimológico, *Econometría* procede de la unión de la palabras griegas *oikonomia* (economía) y *metron* (medida). Es decir, Econometría significaría algo así como la medición de la economía. Cuando revisamos bibliografía, encontramos diversas acepciones que han dado los estudiosos sobre el concepto de Econometría, en las que han señalado los diversos objetivos que persigue esta materia. Destacamos las siguientes definiciones de Econometría:

Spanos : La Econometría está dedicada al estudio sistemático de los fenómenos económicos usando los datos observados.

Cowles Commission : La Econometría tiene por objeto la explicación de la economía y el pronóstico económico mediante el conocimiento de las estructuras o relaciones que describen conductas humanas.

Malinvaud : La Econometría tiene por objeto la determinación empírica de leyes económicas.

Intriligator : La Econometría tiene por objeto la estimación empírica de las relaciones económicas.

Maddala : La Econometría tiene por objeto la aplicación de los métodos estadísticos a datos económicos.

Otero : la Econometría se ocupa de la cuantificación de los fenómenos y de la verificación de las teorías económicas, haciendo uso para ello de las estadísticas económicas y aplicando a tal fin, métodos especiales de inferencia estadística.

Tinbergen : Econometría es un nombre para un campo de la ciencia, en el que la investigación económico-matemática y estadístico-matemática se aplican en combinación.

Tintner : La Econometría es la aplicación de un método específico en el campo general de la ciencia económica, en un esfuerzo de buscar resultados numéricos y de verificar teoremas económicos. Consiste en la aplicación de la teoría económica matemática y los procedimientos estadísticos a los datos económicos en orden a establecer resultados numéricos en el campo de la economía y verificar teoremas económicos.

Gollnick : El análisis econométrico combina la Teoría Económica, las Matemáticas y la Estadística para cuantificar y verificar las relaciones entre las variables económicas.

Judge, Hill, Griffiths, Lütkepohl y Lee : la Econometría, haciendo uso de la Teoría Económica, las Matemáticas y la Inferencia Estadística como fundamentos analíticos, y de los datos económicos como base informativa, proporciona una base para: 1) modificar, refinar o posiblemente refutar las conclusiones contenidas en el cuerpo de conocimientos conocidos como Teoría Económica, y 2) conseguir signos, magnitudes y afirmaciones de calidad para los coeficientes de las variables en las relaciones económicas, de modo que esta información puede usarse como base para la elección y toma de decisiones.

Las teorías o leyes económicas, entendiendo por éstas a las hipótesis aceptadas con validez general, suelen ser formuladas como proposiciones cualitativas de carácter necesario. La inmensa mayoría de estas leyes se expresan de forma sencilla a través de un lenguaje matemático dependiente de una serie de constantes (parámetros). Es entonces cuando la Econometría, a través de un conjunto de datos observados procedentes del fenómeno económico de interés, puede proporcionar las herramientas necesarias para la cuantificación empírica de los valores

de los parámetros. Por ejemplo, la ley de la demanda siempre dice que ésta presenta pendiente negativa en su relación con el precio. Las técnicas estadísticas pertinentes permitirán al econometra obtener el valor concreto de tal pendiente (estimación).

Sobre las anteriores mediciones se fundamentan las predicciones y análisis de los efectos de la aplicación de distintas políticas económicas. Además, han sido ampliamente usadas como base para la toma de decisiones por parte de entidades tanto públicas como privadas.

1.1.2. El modelo económico y el modelo econométrico: un ejemplo

Para llegar a comprender la diferencia entre un modelo "matemático económico" y uno "econométrico", describiremos a continuación un ejemplo que puede ser hallado en Trívez(2004).

Se entiende por modelo teórico a una representación simplificada y en símbolos matemáticos de cierto conjunto de relaciones. Así, por ejemplo, un modelo teórico sobre el consumo puede enunciarse, siguiendo a Keynes, estableciendo que el consumo para un individuo C, viene dado en función de su renta disponible R, esto es,

$$C = f(R).$$

Para pasar de este modelo teórico al econométrico, será necesario especificar una forma funcional concreta para f, e introducir un término de perturbación aleatoria que nos permita razonar en términos probabilísticos y no exactos. Esta perturbación aleatoria servirá para aceptar cierta variabilidad en la relación entre renta y consumo para individuos distintos.

Así, si se intuye una relación lineal entre renta y consumo, un modelo econométrico podría ser:

$$C = \alpha + \beta R + u,$$

donde u contendría toda la información sobre la variabilidad en la relación entre la renta y el consumo, esto es, básicamente, información sobre la heterogeneidad de los individuos respecto a la relación renta-consumo y la influencia de otros posibles factores (variables) relacionados también con el consumo, útiles para predecirlo, pero no considerados en el modelo. Asimismo, el término aleatorio contiene el efecto de los posibles errores de medida en que se incurre a la hora de la toma de los datos para las distintas variables involucradas en el modelo.

1.1.3. Historia de la Econometría

Hacer notar en principio, que estudiosos distintos de la Historia de la Econometría citan como precursores de la misma a científicos diferentes, eso sí, economistas en su mayoría.

Schumpeter, por ejemplo, marca a los economistas de los siglos XVII y XVIII, como Petty, King, Cantillon y Quesnay como los principales precursores de esta disciplina. Otros incluso incluyen en esta lista a Davenant.

Cowles sostiene sin embargo, que "la Econometría estrictamente hablando, tuvo sus orígenes en la Europa del siglo XIX, principalmente con los trabajos de Von Thünen, Cournot, Walras, Jevons, Edgeworth, Pareto y Wicksell". En este sentido, cabe destacar que hacia el año 1870 se registran antecedentes de importancia en los primeros pasos de la Econometría. Sobresalen algunos trabajos que hacen uso de datos atemporales, que habrían de servir más tarde para enunciar las leyes de Engel.

Arrow, junto a Trívez, defiende que la Econometría, como disciplina autónoma, es mucho más reciente, y sitúa su punto de arranque en la segunda década del siglo XX. La investigación econométrica, iniciada en esas fechas, tenía como objetivo principal el análisis de la demanda. Este hecho parece ser debido a la existencia de una teoría de la demanda bien estructurada, la posibilidad de obtener series estadísticas sobre cantidades y precios y, finalmente, como consecuencia de la teoría estadística introducida por Pearson a través del estudio de las correlaciones entre variables.

Se le achaca a Moore ser el máximo fundador del análisis econométrico de la demanda. Moore recurrió en sus investigaciones al empleo de transformaciones de variables, inclusión de variables con retardo, correlaciones múltiples y análisis armónico. Junto a Moore destacan otros científicos: Working, Ezequiel y Shultz. Sobresale el trabajo de Shultz sobre el análisis de la demanda de azúcar.

El empuje de éstos y otros economistas llevó a Fisher en 1912 a intentar crear, apoyado en la Asociación Americana Para El Avance de la Ciencia, una organización compuesta por economistas, estadísticos y matemáticos. Su idea no tuvo éxito.

En 1926 Frisch, con las mismas pretensiones que Fisher, promueve una organización de objetivos similares entre los economistas europeos. Es en una carta de Frisch a Divisa en septiembre de 1926, donde se sugiere, además de la creación de la susodicha asociación, la edición de una revista científica a la que denomina

"*Econometrica*" en analogía con la revista ya existente "*Biometrika*". Parece ser que ésta representa la primera referencia histórica que se posee sobre el término "Econometría".

En cuanto a *Econometrica*, es una revista de carácter científico que publica artículos originales en todas las ramas de la economía, teórica y empírica, abstracta y aplicada. A su vez, promueve los estudios que tienen como objetivo la unificación de lo teórico-cuantitativo y el acercamiento empírico-cuantitativo a los problemas económicos. Para hacernos una idea de la relevancia actual de la revista dentro del panorama internacional, bastará indicar que en el ISI *Journal Citation Reports*®Ranking de 2003, *Econometrica* aparece en primera posición dentro del área de Ciencias Sociales y Métodos Matemáticos, y la cuarta en el área de Estadística y Probabilidad.

Por otro lado, cierta polémica suscitada a raíz de un artículo de Roos en Estados Unidos, generó en 1928 que se discutiera en una reunión de la Asociación Americana para el Avance de la Ciencia, sobre la creación de una nueva sección dedicada plenamente al desarrollo de la Economía Política y la Sociología como ciencias.

Frisch y Roos, de intereses comunes, con el fin de dar origen a una institución completamente independiente, solicitan ayuda a Fisher en el año 1928. Del fruto de ésta y otras reuniones, se funda la *Sociedad de Econometría* (Econometric Society) el 29 de diciembre de 1930, al auspicio de la *Asociación Americana para el Avance de la Ciencia*. El propio Fisher fue nombrado presidente de la sociedad.

Una vez establecida la Sociedad de Econometría, era importante la existencia de una institución donde se localizaran y centraran las investigaciones econométricas. Nacerá así la *Cowles Commission*, financiada íntegramente por el multimillonario Alfred Cowles, quien ofrecerá los fondos necesarios para sacar a la luz el primer número de *Econometrica*, en enero de 1933.

La *Cowles Commission for Research in Economics* se creó en el año 1932 en Colorado Springs, con el objetivo fundamental de "fomentar la investigación científica en Economía, Finanzas, Comercio e Industria".

Dentro de las famosas monografías de la *Cowles Commission* destacan los trabajos de Haavelmo, Hood, Koopmans, Marschak, Rubin, Chernoff, Divinsk y Hurwicz, entre otros. En éstos aparecen cuestiones relacionadas con la inferencia estadística en modelos econométricos multiecuacionales, teoría económica, números índice, etc.

Nos es posible citar, además, las monografías número 10, *Statistical Inference in Dynamic Economic Models*, y la 14, *Studies in Econometric Methods*, como piezas clave de la obra clásica de la Econometría. No obstante, la mayoría de los anteriores trabajos forman parte de un periodo inicial de lo que se entiende hoy en día por estudio econométrico. Más recientemente destacan los siguientes trabajos:

- Chow(1960): test sobre permanencia estructural.

- Zellner(1962, 1963): sistemas de ecuaciones aparentemente no relacionadas.

- Shapiro y Wilk(1963): test de normalidad.

- Box y Cox(1964): transformación de variables.

- Goldfeld y Quandt(1965): contrastes sobre heterocedasticidad.

- Durbin y Watson(1971): autocorrelación.

Todo lo anterior hace referencia a los inicios y desarrollo posterior de la Econometría, no obstante, en cuanto a las pretensiones científicas de relacionar linealmente una variable Y con otra X, cabe señalar que fueron Laplace y Gauss los primeros en aplicar modelos matemáticos lineales en disciplinas como la Astronomía y la Física. Sin embargo, el nombre por el que son conocidos estos modelos matemáticos, *Modelos de Regresión*, proviene de los trabajos de Galton a finales del siglo XIX. Galton, primo de Darwin, encaminó sus investigaciones hacia el estudio de la herencia biológica, concibiendo las nociones modernas de correlación y regresión.

Es habitual encontrar en los libros de texto la definición de coeficiente de correlación lineal de Pearson antes de la definición de recta de regresión. Sin embargo, en Stanton (2001) hallamos la verdadera secuencia temporal en la introducción de estas herramientas en la literatura científica. La primera recta de regresión fue presentada en 1877 en un estudio desarrollado por Galton, mientras que Pearson presentó formalmente su coeficiente de correlación en el año 1896, en el *Philosophical Transactions of the Royal Society of London.*

El nombre de *Modelo de Regresión*, por otro lado, procede de la naturaleza de algunos de los resultados obtenidos por el propio Galton. Éste estudió la dependencia de la estatura de los hijos (variable que hacía las veces de variable respuesta Y) respecto a la de sus padres (variable explicativa X). Halló así lo que denominó una "regresión" a la media de los datos. Los padres altos tienen, en términos generales, hijos altos, pero, en promedio, no tan altos como sus padres.

Por otro lado, los padres bajos tienen hijos bajos, pero, en media, más altos que sus padres. Desde entonces viene utilizándose dicha denominación en el mundo estadístico y econométrico.

Tema 2

El proyecto R

2.1. ¿Qué es R?

Los ejemplos numéricos que aparecen en el libro han sido desarrollados enteramente con R. R es un proyecto GNU (software libre) que posee implementadas una gran variedad de librerías, entre las que destacan las de gráficas y técnicas estadísticas. Entre ellas se encuentran: el modelo de regresión lineal y no lineal, pruebas estadísticas de todo tipo, análisis de series temporales, análisis multivariante, etc.

Para acceder a este software debemos conectarnos a CRAN. CRAN es una red ftp de servidores de internet alrededor del mundo que almacenan versiones del código y documentación de R idénticas y actualizadas. En las siguientes secciones veremos cómo podemos descargarnos dicho programa desde la red CRAN a través de dos vías distintas.

2.1.1. Motivos para usar R

R es un paquete estadístico tan potente o más que otros actualmente de moda en el panorama mundial dentro del contexto del uso de la Estadística. La ventaja fundamental que presenta frente a otros, y que lo hace del todo atractivo, es su gratuidad total de uso. Esto implica que desde un punto de vista empresarial de minimización de costes resulte más adecuada su adquisición que otro tipo de software existente en el mercado actual. Al mismo tiempo cabe

destacar la continua actualización y mejora que se viene produciendo de R, y de las cuales, nos podemos beneficiar descargando las actualizaciones de manera sencilla a través de internet. Sin embargo, el interface de R difiere del típico usado en otras muchas aplicaciones para Windows (como con el SPSS, por ejemplo). En concreto, para comunicarnos con R lo tendremos que hacer a través de la utilización de una línea de comandos.

2.1.2. Cómo descargarnos R desde internet

Aunque las formas de descargarse R y sus múltiples librerías son variadas, hemos optado por mostrar dos de ellas. Mediante la primera haremos uso de la web oficial del proyecto R: www.r-project.org. La segunda opción necesita la conexión a la redIRIS (Interconexión de los Recursos InformáticoS) patrocinada por el Plan Nacional de I+D. Evidentemente, en ambos casos necesitamos estar conectados a internet.

1. **Con el navegador de internet**. La dirección URL a la que debemos dirigirnos es http://www.r-project.org. En el menú de la izquierda seleccionamos *Download CRAN*. Elegimos como servidor el español (Spain). Para descargarnos R, pinchamos sobre *Windows (95 and later)* → *Base* → *fichero.exe*. Donde *fichero.exe* denota al programa ejecutable que contiene la última versión de R. Éste es el archivo que debéis descargaros a vuestro disco duro. Una vez completado el proceso de descarga, bastará con pinchar dos veces sobre el archivo y aceptar todas las opciones por defecto. Se instalará el programa y aparecerá un icono de acceso en el escritorio de vuestro ordenador. Asimismo, las librerías son accesibles desde este mismo entorno en *Windows (95 and later)/Contributed extension packages*.

2. **Vía ftp**. Haremos uso de cualquier cliente ftp (CuteFTP, WSFTP,...). El nombre del servidor ftp es: *ftp.rediris.es*. En usuario y contraseña debemos introducir, en ambos casos, *anonymous*. Dentro de la carpeta *mirror* buscaremos *CRAN*. Una vez hayamos entrado en *CRAN* pincharemos dos veces sobre la carpeta denominada *bin*. Dentro de bin seleccionamos *windows*. Y dentro de la subcarpeta *base* encontraremos el fichero ejecutable .exe conteniendo la última versión de R. Como con el método anterior, una vez completado el proceso de descarga, bastará con pinchar dos veces sobre el archivo y aceptar todas las opciones por defecto. Indicar que se os instalará en el sistema la versión correspondiente, apareciendo un icono identificativo en el escritorio de vuestro ordenador.

Las librerías de R se encuentran en *bin*, dentro de la versión correspondiente en la carpeta *contrib*.

Existe la posibilidad de actualizar todas las librerías de R automáticamente por Internet. Para ello bastará con ejecutar el icono de R de vuestro escritorio y, posteriormente, ir al menú *Packages*. Bajo este menú se encuentra la opción *Update packages from CRAN*. Al seleccionar ésta, comenzará la descarga de librerías. En la siguiente Sección se cuenta cómo cargar una librería en concreto para poder ser utilizada por R.

Ha llegado el momento de comenzar a dar nuestros primeros pasos con R.

2.2. Principios Básicos de Uso

En esta sección mostraremos los comandos básicos de uso de R que aparecen en los menús del programa, así como diversas funciones de interés para el desarrollo de los ejemplos y ejercicios del libro. No obstante, nos gustaría señalar que para obtener información más pormenorizada sobre las funciones implementadas en R, deberá acudirse a la ayuda que el propio programa contiene.

Operación	**Menú**
Consulta de ayuda sobre una función concreta	?funcion help("funcion") Help/R functions (text)
Consulta de ayuda sobre funciones que contienen una "cadena" de caracteres	apropos("cadena") Help/Apropos
Consulta de ayuda sobre funciones relacionadas con una "cadena" de caracteres	help.search("cadena") Help/Search help
Consulta de ayuda html	Help/Html help/Search Engine...
Especificar directorio de trabajo	File/Change dir
Ejecutar fichero de código .r Abrir Editor y crear fichero ASCII (p.e.sintaxis) Abrir Editor y fichero ASCII (p.e.sintaxis) Ejecutar sintaxis Guardar fichero ASCII de edición Visualizar fichero ASCII (p.e.datos)	File/Source R Code File/New Script File/Open Script Pulsar botón derecho ratón Run Line or selection (Ctrl+R) Cursor en editor File/Save File/Display File(s)
Inspección de objetos creados Salvar objetos creados en un fichero.RData Cargar objetos guardados en un fichero.RData	objects(), ls() File/Save Workspace File/Load Workspace
Guardar historia de comandos a file.RHistory Cargar objeto .RData	File/Save History File/Load Workspace source("fichero.R")

Operación	**Menú**
Volcado de resultados a un fichero	sink("fichero.out") ... sink() cat(...,file="fichero")
Guardar gráfico	Cursor sobre gráfico File/Save as/
Inspección bases de datos disponibles	data()
Cargar base de datos disponible	data(nombre base de datos)
Inspección de librerías cargadas	search()
Instalar librerías	Packages/Install package(s) ...
Cargar librerías instaladas	Packages/Load package library(librería)
Consulta de ayuda sobre librerías	Help/Html help/Packages

2.3. Creación y manipulación de objetos

Para crear un objeto es preciso utilizar el operador de asignación `<-`, subrayado, o directamente el comando *assign()*. En las últimas versiones de R también funciona "=", pero conviene usarlo con cautela, pues no siempre funciona.

```
x<-5
x=5
assign("x",5)
```

2.3.1. Clases de objetos

Los tipos de objetos en R son:

Vectores	vector()
Factores (*cualitativos*)	factor(),gl()
Listas	list()
(admite elementos de distinto tipo)	
Matrices	matrix(ncol,nrow), array(dim)
Bases de datos	data.frame()
(matriz con columnas de distinto tipo)	

En general, salvo los factores, todos los objetos admiten elementos de tipo numérico, carácter, complejo o lógico. Los factores sólo aceptan elementos numéricos o caracteres.

Para inspeccionar y/o modificar los atributos de un objeto se utiliza el comando `attribute`. Todo objeto tiene dos atributos intrínsecos: tipo y longitud. Los

tipos básicos de los objetos son: "logical", "integer", "numeric", "complex", "character" y "list". Para inspeccionar el tipo de un objeto se utilizan los comandos `typeof` y `mode`.

```
attributes(objeto)
typeof(objeto)
mode(objeto)
```

2.3.2. Operadores Lógicos

El símbolo # identifica como comentario todo lo que sigue en la misma línea.

>	mayor que	<	menor que
>=	mayor o igual que	<=	menor o igual que
==	igual que	! =	distinto de
|	uno u otro	&	uno y otro

2.3.3. Vectores, Secuencias y Factores

Un vector es una colección ordenada de números, símbolos lógicos, caracteres o complejos.

```
x <-1:30
c(1, 1.5, 2, 2.5, 3, 3.5, 4, 4.5, 5)
seq(1, 5, by=0.5)
seq(1, 5, length=20)
seq(length=9, from=1, to=5)
z <- scan() # e introducimos los números que queremos
# pulsando al final dos veces la tecla Intro
rep(1, 30)
sequence(c(10,5))
```

Es posible leer datos sin formato de ficheros externos con el comando `scan( file, what, sep, skip, nlines, flush)`. Donde *file* es el nombre del archivo externo, *what* identifica al tipo de dato que va a ser leído, *sep* identifica al separador de los datos, *skip* es el número de líneas del fichero de entrada que debe saltar R antes de empezar a leer los datos, *nlines* es el número máximo de líneas que deben ser leídas y, por último, *flush* es un valor lógico indicando si aparecen en el fichero comentarios al final de cada línea.

```
cat("Probamos a leer cosas","2 3 5 7", "11 13 17",
file="prueba.data", sep="\n")
pp <- scan("prueba.data")
scan("prueba.data", skip = 1)
scan("prueba.data", skip = 1, nlines=1)
str(scan("prueba.data", what = list("","","")))
str(scan("prueba.data", what = list("","",""), flush = TRUE))
unlink("prueba.data") # Borra el fichero.
```

Las funciones `gl()` y `factor()` generan factores, esto es, vectores cuyos elementos son de tipo cualitativo y que sirven para identificar grupos o categorías de clasificación. Es posible identificar los niveles y asignarles etiquetas.

```
# Un factor con 3 niveles, cada uno repetido 5 veces
gl(3, 5)
# un factor con 3 niveles, que salta al nivel siguiente
# cada 5 observaciones y tiene una longitud total de 30
gl(3, 5, length=30)
gl(2, 6, labels=c("Macho", "Hembra"))

factor(rep(1:2,5),levels=c(1,2),labels=c("Blanco","Negro"))
```

Podemos categorizar variables de tipo numérico con el comando `cut()`, y particionar los elementos de un vector en función de los niveles de un factor con `split()`. También podemos seleccionar los elementos de un vector/factor, con corchetes:

```
# datos
edad<- c(7,15,84,45,21,11,35,57,71,55,62,12,35,36,25,44,45,80,18)
edad.f <- cut(edad,breaks=c(0,15,30,50,75,100))

# partición por grupos
x<-seq(1,5,length=6)
grupos<-gl(2,3)
split(x,grupos)

# selección
x[2]
x[(1:15)*2]
x[x>10]
x[x<3 | x>27]
x[x<20 & x>15]
```

```
# prescindimos del primer elemento de x
x[-1]

# selección a través de un objeto de dimensiones similares:
color <- gl(2,15,30,labels=c("Blanco","Rojo"))
x[color=="Rojo"]
```

Los datos faltantes se identifican con `NA`:

```
x  <-  c(1,5,6,2,NA)
is.na(x)
y <- x[!is.na(x)]
x[is.na(x)] <- 999
```

2.3.4. Matrices y Arrays

Un array es una colección de entradas de datos del mismo tipo, ya sea numérico, categórico o lógico.

Las matrices son un caso particular de arrays. Los arrays son matrices multidimensionales. La selección de elementos se hace con corchetes.

```
matrix(1:8,ncol=4)
matrix(ncol=2,nrow=2)
mat<-matrix(1:8,ncol=2,byrow=T)
dimnames(mat)<-list(c("F1","F2","F3","F4"),c("Q1","Q2")); mat

# Se pueden crear pegando vectores y matrices, por filas o columnas
cbind(1:6,7:8)
rbind(1:6,7:9)

# Un array tridimensional
a<-array(c(1:8,11:18,111:118),dim= c(2,4,3)
# una matriz con 2 filas y 3 columnas
b<-array(dim=c(2,3))

# Selección de elementos
a[1,1,2]
a[,,2]
```

2.3.5. Listas

En una lista podemos mezclar elementos de diferentes tipos. La selección de elementos se hace con doble corchete o con el nombre del elemento precedido de $:

```
lista  <- list(array=a,matriz=mat,vector=x)

lista[[1]]
names(lista)
lista$array}

# Concatenar varias listas en una sola:
lista.A <- list(xA=1:5,yA=c("rubio","moreno"))
lista.B<- list(xB=85,yB=gl(2,1,10,labels=c("ojo","oreja")))
lista.AB <-  c(lista.A,lista.B)
```

2.3.6. Operaciones con vectores, matrices y listas

Las operaciones habituales (suma, resta, multiplicación, división) se denotan con sus respectivos símbolos (+,-,*,/).

Otras operaciones de interés son:

v*z	producto interno de los vectores v y z
A+B	suma las matrices A y B
A-B	resta las matrices A y B
A*B	multiplica las matrices, elemento a elemento
A %B	divide las matrices, elemento a elemento
A %* %B	producto matricial
t(A)	matriz transpuesta de A
solve(A)	matriz inversa de A
solve(A,b)	resuelve la ecuación Ax=b
svd()	proporciona la descomposición en valores singulares
qr()	obtiene la descomposición QR
eigen()	calcula valores y vectores propios
sort(v)	ordena el vector v
diag(v)	proporciona una matriz diagonal el vector v
diag(A)	extrae un vector con los elementos de la diagonal de la matriz A

tapply(v,factor,función)	aplica la *función* sobre los valores del vector *v* clasificado en las categorías del *factor*
sapply(split(v,factor),fun))	tiene el mismo efecto que la anterior tapply
lapply(lista,fun)	calcula la función fun en cada uno de los elementos de una lista
apply(A,index,fun)	aplica la función fun sobre cada una de las filas (index=1) o columnas (index=2) de la matriz A

2.3.7. Data Frames

Una hoja de datos es una lista que pertenece a la clase data.frame. Están caracterizadas por lo siguiente:

- las componentes han de ser vectores, factores, matrices numéricas, listas u otras hojas de datos.
- Las matrices, listas y hojas de datos contribuyen a la nueva hoja de datos con tantas variables como columnas, elementos o variables posean, respectivamente.
- Los vectores numéricos y los factores se incluyen sin modificar; los vectores no numéricos se fuerzan a factores cuyos niveles son los únicos valores que aparecen en el vector.
- Los vectores que constituyen la hoja de datos deben tener todos la misma longitud, y las matrices el mismo número de filas.

Las hojas de datos pueden tratarse generalmente como matrices cuyas columnas pueden tener diferentes modos y atributos. Pueden imprimirse en forma matricial y se pueden extraer sus filas o columnas según la indexación de matrices.

```
clima.cont <- c("no","no","si")
ciudades <-  c("Sevilla","Alicante","Albacete")
temp.max <-  c(41,35,39)
ciudad.datos <- data.frame(ciudades,clima.cont,temp.max)
```

Es posible cargar datos disponibles en R como hojas de datos (data.frame). Una vez cargadas, podemos trabajar con sus variables directamente, utilizando el comando `attach(dataframe)`. Para olvidar estas variables, basta utilizar `detach(dataframe)`.

```
data(infert)
infert$education
attach(infert)
education
detach(infert)
```

Podemos leer datos con formato de base de datos con el comando `read.table(file, header)`, donde *file* es el nombre del fichero y *header* es un valor lógico indicando si el fichero contiene el nombre de las variables en la primera línea.

```
df <-  data.frame(pelo=c("rubio","moreno"),talla=c(1.69,1.85),
CI=c(140,110))
sink(file="prob")
df
sink()
read.table("prob",header=T)
unlink("prob")
```

2.4. Ejercicios

Inicialmente especifica el directorio de trabajo en que quieres trabajar.

1. Crear un vector de datos (var X) en el que los 10 primeros datos identifiquen a un tipo de individuos (A), los 4 siguientes a otro (B) y los restantes a un tercero (C).

2. Sean los datos:

```
ctl <-  c(4.17,5.58,5.18,6.11,4.50,4.61,5.17,4.53,5.33,5.14)
trt <-  c(4.81,4.17,4.41,3.59,5.87,3.83,6.03,4.89,4.32,4.69)
```

   que identifican las mediciones correspondientes a individuos de dos grupos: control (ctl) y tratamiento (trt). Crear un vector con todos los datos y un factor que identifique individuos de uno y otro grupo.

3. Transforma en data.frame los datos de los Ejercicios 1 y 2. Direcciónalo a un fichero externo y comprueba cómo quedó (sin salir de R).

4. Construye una matriz A cuya diagonal esté constituida por la secuencia de enteros del 1 al 6. Construye otra matriz B de idéntico tamaño en la que la primera fila sean todo 1, la segunda todo 2, y así hasta la última, con todos los elementos igual a 6. Construye un vector b de dimensión 6 con la diagonal de la matriz B. Realiza las siguientes operaciones:

   a) b'(A+B)

   b) b'(AB)

   c) $A^{-1}B$

   d) Resuelve (A+B)x=b

   e) A los elementos iguales a cero en la matriz A asígnales un valor muestreado al azar de entre los números del 0 al 9. (Nota: utiliza el comando *sample()*).

   f) Selecciona la segunda fila de B, llámala b2 y calcula b'b2 y b2'A.

5. Guarda la sintaxis y los objetos generados en la sesión. Sal de R y vuelve a acceder a R cargando todo lo que ejecutaste anteriormente. Comprueba que todo funciona.

Tema 3

Ejemplos y Modelización Estadística

3.1. Introducción

Cuando estamos estudiando una población respecto de ciertas características de interés, es habitual el buscar interrelaciones e incluso intentar predecir alguna o algunas de ellas en función de otras. Es entonces cuando la modelización estadística interviene y el analista busca el mejor modelo que ajusta los datos disponibles y proporciona predicciones fiables. Nosotros trabajaremos aquí con modelos de tipo lineal; en concreto profundizaremos en el modelo de regresión lineal. Dentro de este modelo se engloban los modelos de regresión lineal simple, múltiple, los modelos de Anova y de Ancova. Con el fin de ilustrar mejor los objetivos que se persiguen con la modelización de cada uno de ellos, preferimos presentar primeramente un conjunto de problemas diversos a resolver, que constituirán luego la justificación de dichos modelos. Este conjunto de problemas son resueltos a lo largo del curso conforme avanzamos presentando los pasos a dar para conseguir nuestros propósitos.

3.2. Ejemplos

Ejemplo 3.1 (Tractores)**.** *Parece ser que el dinero gastado en la manutención de tractores es mayor a medida que aumenta la edad del tractor. Se pretende ratificar esta hipótesis utilizando los datos del Apéndice B, representados en la*

Figura 3.1 (de Draper y Smith, 1998, pag.100). Se aprecia cierta tendencia lineal para explicar el gasto en función de la edad del tractor.

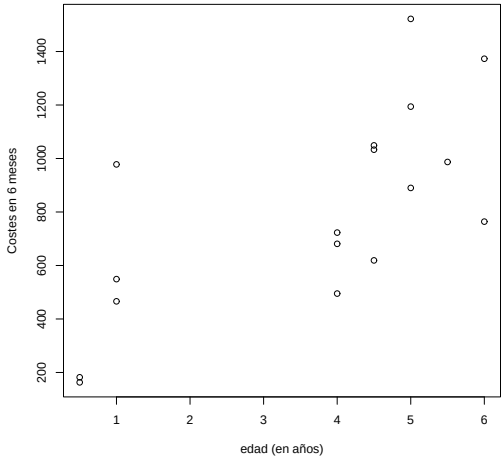


Figura 3.1: Tractores. Relación entre la edad del tractor y el gasto que ocasiona su manutención.

Ejemplo 3.2 (Bosque)**.** *Para estimar la producción en madera de un bosque se suele realizar un muestreo previo en el que se toman una serie de mediciones no destructivas. Disponemos de mediciones para 20 árboles, así como el volumen de madera que producen una vez cortados (ver datos en el Apéndice B). Las variables observadas son:*

HT *= altura en pies*
DBH *= diámetro del tronco a 4 pies de altura (en pulgadas)*
D16 *= diámetro del tronco a 16 pies de altura (en pulgadas)*
VOL *= volumen de madera obtenida (en pies cúbicos).*

El objetivo del análisis es determinar cuál es la relación entre dichas medidas y el volumen de madera, con el fin de poder predecir este último en función de las primeras. Los datos están representados en la Figura 3.2 (de Freund y Wilson, 1998, pag.39).

Ejemplo 3.3 (Bosque2)**.** *Planteada la misma problemática que en el Ejemplo 3.2, disponemos de otro banco de datos en* $\mathcal{R}$ *que contiene las siguientes variables:*

Girth *= circunferencia del árbol a 137 cm (4 piés y 6 pulgadas)*
Height *= altura del árbol*
Volume *= volumen de madera obtenida (en pies cúbicos).*

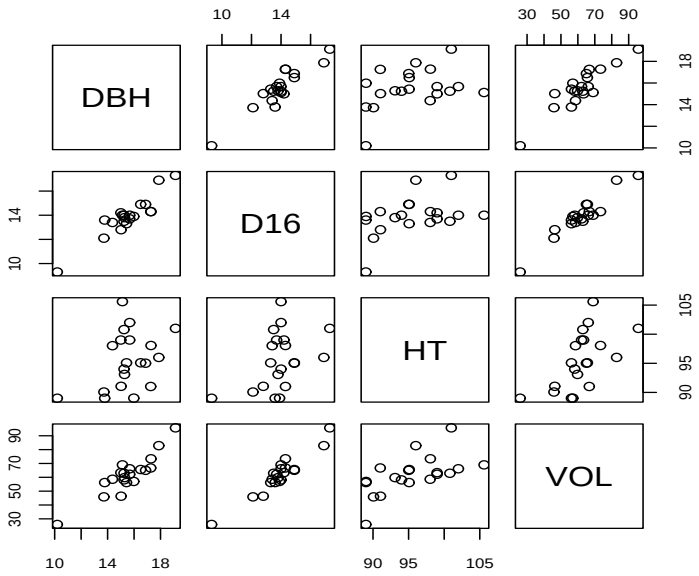


Figura 3.2: Bosque. Relación entre las variables observadas para los 20 árboles de la muestra.

El objetivo es pues, predecir el volumen de madera en función de las otras dos variables. Los datos (de Ryan et al, 1976), disponibles en $\mathcal{R}$ e identificados como 'tree' están representados en la Figura 3.3.

Ejemplo 3.4 (Pizza)**.** *Para un periodo de 12 meses, el gerente de Pizza Shack ha invertido en la aparición de anuncios en el periódico local. Los anuncios se programan y pagan un mes antes de que aparezcan publicados. Cada uno de estos anuncios contiene un cupón de 2×1. Al gerente le gustaría ser capaz de predecir las ventas de pizza (sales) en relación al número de anuncios publicitados (ads) y al coste de éstos (cost). Es un punto a considerar la estrecha relación existente entre el número de anuncios programado y el coste agregado de los mismos, la cual puede ocasionar problemas en la fiabilidad del ajuste obtenido. Los datos provienen de Levin y Rubin (1998), Cap.13 y se presentan en el Apéndice B. Asimismo están representados en la Figura 3.4, donde se aprecia la relación lineal entre las variables que han de actuar como predictoras (ads y cost).*

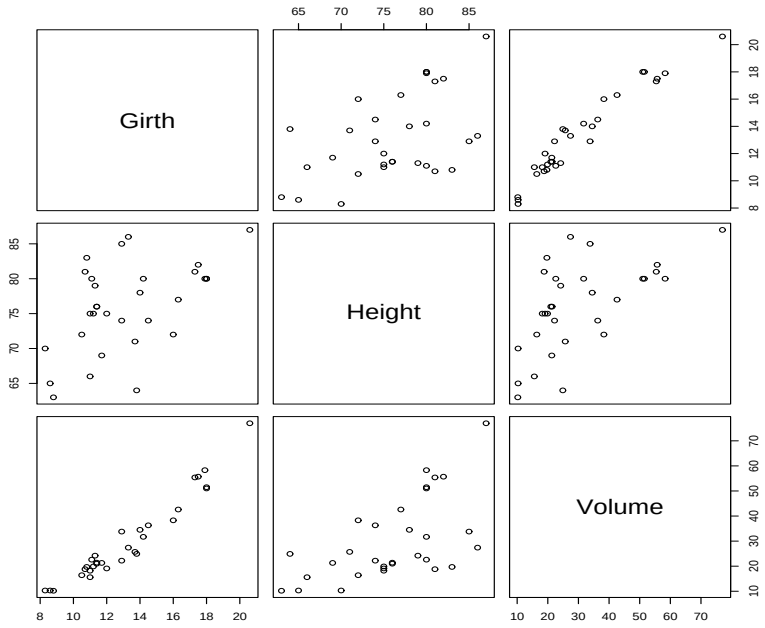


Figura 3.3: Bosque2 (tree). Relación entre las variables observadas para los árboles de la muestra.

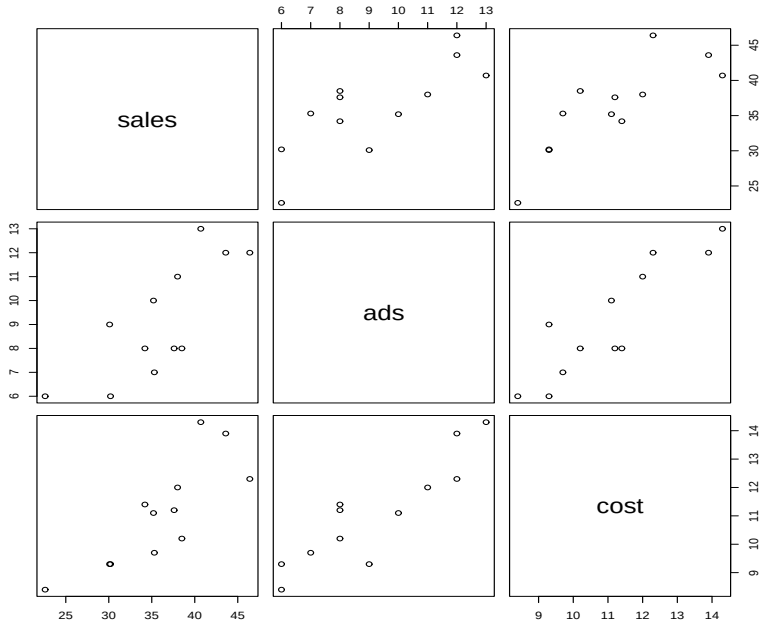


Figura 3.4: Pizza. Relación entre las variables consideradas para la predicción de ventas en función de la inversión publicitaria.

Ejemplo 3.5 (Papel)**.** *En la Figura 3.5 están representados unos datos (ver Apéndice B) medidos con los que se pretende concluir sobre la relación existente entre la concentración de madera contenida en la pulpa, a partir de la que se elabora papel, y la resistencia (en términos de tensión que soporta) del papel resultante. El objetivo del análisis es describir la tendencia observada, curva según el gráfico, a través de un modelo aceptable de predicción de la resistencia en función de la concentración de madera. Los datos provienen de Montgomery y Peck (1992), pag.205.*

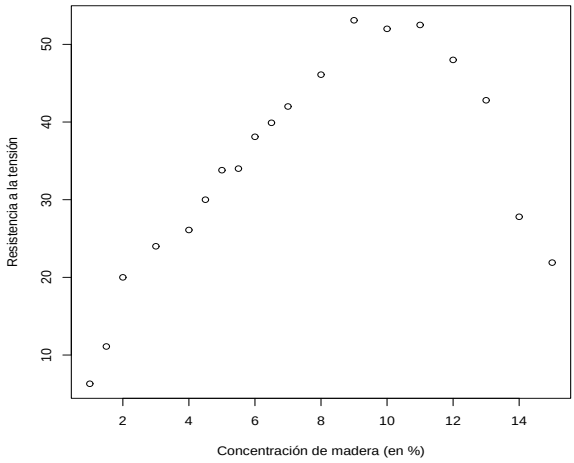


Figura 3.5: Papel. Relación entre la concentración de madera en la pulpa para la elaboración del papel y la resistencia del papel resultante.

Ejemplo 3.6 (Hipertensión)**.** *Se lleva a cabo un experimento clínico en el que un grupo aleatorizado de pacientes, identificado como 'grupo T', prueba un nuevo fármaco contra la hipertensión, y otro grupo también aleatorizado y llamado 'grupo P' toma un placebo. A todos los pacientes se les mide la tensión arterial antes de administrar el tratamiento (fármaco o placebo) y después de finalizar el período de observación. El objetivo del análisis es investigar si el fármaco tiene un efecto sobre la reducción de la tensión arterial y cuantificar dicho efecto si existe. Las diferencias, para cada grupo de pacientes, entre la tensión arterial antes y después del tratamiento están representadas en la Figura 3.6. Se aprecian diferencias entre los dos grupos, que habrán de ser ratificadas estadísticamente.*

Ejemplo 3.7 (Insecticidas)**.** *Se trata de un banco de datos de $\mathcal{R}$ (data(InsectSprays)). Los datos provienen de un experimento agronómico para testar varios insecticidas (InsectSprays). Las variables observadas son 'count' el número de insectos que quedan tras la vaporización con el insecticida, y 'spray',*

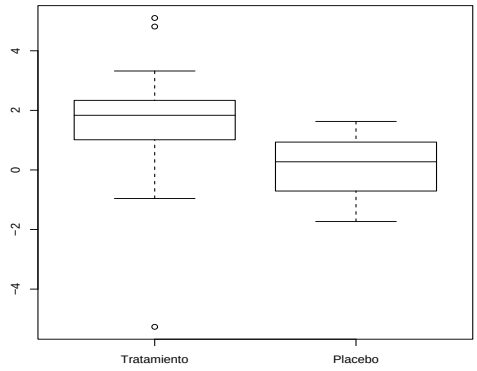


Figura 3.6: Hipertensión. Diferencias en la tensión arterial antes y después del tratamiento, para los dos grupos de pacientes: tratamiento y placebo.

el tipo de insecticida utilizado (un factor de clasificación). El objetivo es concluir sobre la efectividad de los diferentes insecticidas, evaluada en términos del número de insectos que mueren con ellos. Los datos están representados en la Figura 3.7.

Ejemplo 3.8 (Torno)**.** *En la Figura 3.8 están representados los datos sobre tiempo de vida de una pieza cortadora de dos tipos, A y B, en función de la velocidad del torno en el que está integrada (en revoluciones por segundo). Este ejemplo proviene de Montgomery y Peck (1992), pag.239. El objetivo del análisis es describir la relación entre el tiempo de vida de la pieza y la velocidad del torno, teniendo en cuenta de qué tipo es la pieza; detectar diferencias, si existen, entre los dos tipos de piezas; y predecir el tiempo de vida de la pieza en función de su tipo y de la velocidad del torno. El gráfico de la Figura 3.8 sugiere una relación lineal entre velocidad y tiempo de vida, pero diferente para cada tipo de pieza.*

3.3. Modelización Estadística

Una vez establecidos los objetivos del problema que queremos estudiar, conviene definir con precisión cuáles son las características a medir sobre un conjunto de individuos seleccionados de una población de interés sobre la que queremos extraer conclusiones. Comenzaremos exponiendo unos cuantos conceptos básicos de cara al tratamiento estadístico de un problema.

Una variable estadística es una característica observable en una población de

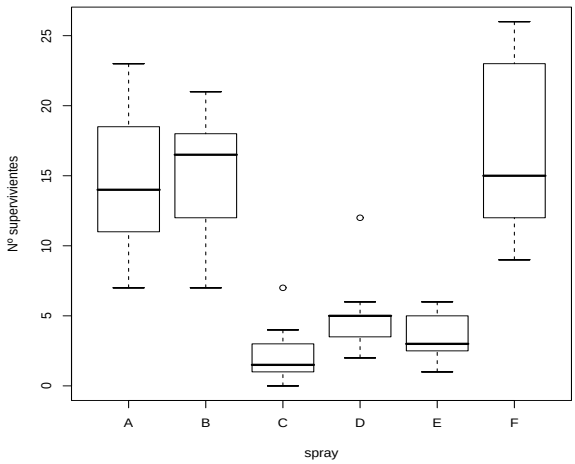


Figura 3.7: Insectos. Número de insectos supervivientes tras la vaporización con los diferentes insecticidas.

interés; por ejemplo, los ingresos mensuales de un individuo. Los conceptos de población (conjunto de individuos/unidades experimentales de interés) y población estadística difieren. La población estadística es el conjunto de todas las observaciones que podemos registrar del valor de la variable estadística sobre todas las unidades experimentales de interés; en nuestro caso, la población estadística estaría constituída por los ingresos mensuales de todos los individuos sobre los que queremos concluir, pertenecientes a un colectivo de interés (población). Los parámetros son características numéricas que sintetizan la información sobre todos los elementos de la población, y que, como las poblaciones mismas, pueden ser conocidos o no; p.e., el ingreso mensual medio de la población, o el porcentaje de individuos en la población con ingresos mensuales superiores a 1000 euros. Generalmente, las poblaciones de interés, por su tamaño, son imposibles de observar completamente, por lo que para su estudio se utiliza sólamente un subconjunto de la misma, esto es, una muestra; cuando es escogida al azar, se denomina muestra aleatoria. Las muestras aleatorias son la herramienta primaria de la inferencia estadística y el motivo por el cual se puede utilizar la probabilidad para extraer conclusiones sobre la población de interés; pretenderíamos en nuestro ejemplo, concluir sobre los ingresos en el colectivo de interés a través de una muestra aleatoria de ingresos recogidos sobre unos cuantos individuos seleccionados al azar de dicho colectivo. Siempre es conveniente en un estudio estadístico sobre una población, garantizar lo máximo posible la representatividad de la muestra elegida en la población de interés, esto es, garantizar que hemos recogido, y en su justa medida, la heterogeneidad existente en la población total;

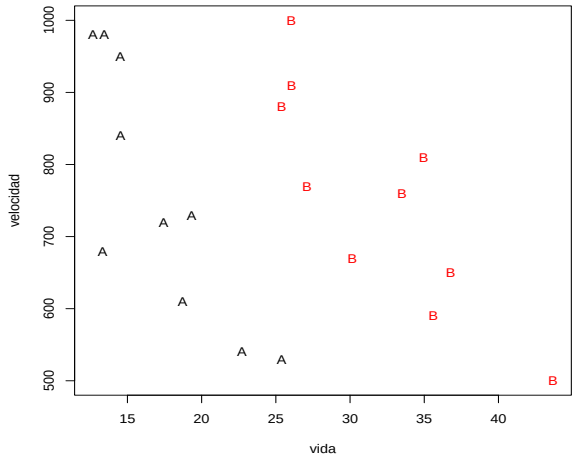


Figura 3.8: Torno. Relación entre el tiempo de vida de una pieza cortadora y la velocidad del torno en el que se integra, en función del tipo de pieza (A o B).

para ello hay diversos métodos de muestreo (selección) aleatorio, que conviene saber utilizar adecuadamente. El método de muestreo aleatorio más común es el muestreo aleatorio simple, en el que todos los elementos de la población tienen la misma probabilidad de ser seleccionados y además, son seleccionados independientemente unos de otros. Los estadísticos son características de una muestra y juegan el mismo papel sobre una muestra que los parámetros describiendo a una población; de hecho, algunos estadísticos se utilizarán para aproximar parámetros, pasando a denominarse entonces, estimadores. El ingreso mensual medio de la muestra, o el porcentaje de individuos en la muestra con ingresos mensuales superiores a 1000 euros son estadísticos y además estimadores de sus análogos en la población.

Según Ferrándiz (1994), "la probabilidad está inmersa en el razonamiento científico general, y en particular en el estudio de fenómenos sociales, económicos, biológicos, ya que i) el azar se encuentra presente en muchos procesos, ii) los resultados experimentales u observacionales presentan siempre una cierta variabilidad atribuible a factores no controlados y que varían aleatoriamente y iii) los investigadores utilizan mecanismos de selección aleatoria de unidades experimentales a la hora de diseñar experimentos. Las conclusiones de un análisis estadístico y la fiabilidad de las mismas se formulan en términos probabilísticos. Eso es debido, por un lado a las razones anteriores, y por otro al hecho de que los modelos probabilísticos fundamentan la justificación teórica de la inferencia estadística. Es decir, no hay inferencia estadística sin probabilidad".

Los "modelos estadísticos" son representaciones matemáticas de sucesos reales manifestables en la observación de fenómenos. Cuando podemos observar un fenómeno que es variable (varía de unos sujetos a otros y en función de ciertas condiciones) y recoger información cualitativa o numérica sobre lo que sucede, podemos emplear la Estadística. La modelización estadística de los datos recogidos nos permitirá representar la realidad de un modo fácil que nos facilite entenderla y predecirla.

El proceso de modelización y análisis estadístico de un banco de datos se puede sintetizar en las siguientes pautas de actuación:

1. Contextualización del problema. Definición de objetivos y variables.
2. Inspección gráfica e identificación de tendencias.
3. Consideración de hipótesis distribucionales y relacionales. Propuesta de modelos.
4. Ajuste y revisión (diagnóstico) de modelos. Comparación y selección del mejor modelo.
5. Valoración de la capacidad predictiva del modelo. Validación.
6. Interpretación y conclusiones.

Si la revisión del modelo nos lleva a descartarlo, será preciso una nueva propuesta, de modo que entraríamos en un bucle entre los puntos (4) y (5), que culminará cuando quedemos satisfechos con el diagnóstico y la validación del modelo.

3.4. Notación

En general, en esta asignatura, dispondremos de una variable *respuesta* de tipo continuo, que trataremos de explicar o predecir, y que denotaremos por y. Las variables *explicativas* o *predictoras* son aquellas que se utilizan para explicar y predecir el comportamiento de la respuesta y las notaremos con $\mathrm{x}_i, i = 1, \ldots, p-1$. Estas variables pueden ser de tipo cuantitativo (comúnmente denominadas covariables) o cualitativo (también llamadas factores de clasificación). El objetivo de la modelización será el de establecer qué relaciones existen entre las

variables explicativas y la respuesta, y cómo las primeras pueden predecir el comportamiento de la respuesta. La modelización se abordará sobre las observaciones conseguidas en una muestra aleatoria de individuos de la población de interés. Así pues, una muestra de tamaño n está constituida por las observaciones:

$$\{y_1, y_2, \ldots, y_n\}, \qquad \{x_{i1}, x_{i2}, \ldots, x_{in}\}_{i=1}^{p-1}.$$

Con esta notación, un modelo estadístico se compone de:

- Una componente sistemática, que representa la tendencia general manifestada entre la variable respuesta y sus predictores, esto es, entre lo que se espera observar para y, $E(\mathsf{y})$, y las variables explicativas involucradas $\mathrm{x}_1, \mathrm{x}_2, \ldots, \mathrm{x}_{p-1}$. Dicha tendencia general se explicita a través de un funcional $f()$ que puede tener formas diversas, dando lugar a modelos diferentes:

$$E(\mathsf{y}) = f(\mathrm{x}_1, \mathrm{x}_2, \ldots, \mathrm{x}_{p-1})$$

- Una componente aleatoria, que añade a la tendencia general la variabilidad que proviene de las diferencias entre las unidades de la población. Esta variabilidad extra puede ser debida a factores y variables no considerados en el modelo, así como al error de medida. Dichas diferencias, a las que denominamos errores, ϵ, vendrán modelizadas por una distribución de probabilidad, $\epsilon \sim F$ con media cero $E(\epsilon) = 0$ y $Var(\epsilon) = Var(\mathsf{y})$, que determinará las hipótesis distribucionales asumidas sobre el modelo. Así, el modelo completo quedará especificado según:

$$\mathsf{y} = f(\mathrm{x}_1, \mathrm{x}_2, \ldots, \mathrm{x}_{p-1}) + \epsilon.$$

En lo que sigue, trabajaremos con modelos lineales, esto es, modelos donde el funcional $f()$ será del tipo:

$$f(\mathrm{x}_1, \mathrm{x}_2, \ldots, \mathrm{x}_{p-1}) = \beta_0 + \beta_1 \mathrm{x}_1 + \beta_2 \mathrm{x}_2 + \ldots + \beta_{p-1} \mathrm{x}_{p-1}.$$

Tema 4

Análisis de Correlación

4.1. Introducción

Cuando trabajamos en la modelización de una variable respuesta continua (con predictores continuos o categóricos), lo habitual será acudir a modelos de regresión que permitan predecir la primera en función de las restantes variables observadas que estén relacionadas con ella. Sin embargo, previamente a dicha modelización, convendrá llevar a cabo, aparte de la inspección gráfica, un análisis de asociación. Cuando las variables explicativas son de tipo continuo y se estudia si las relaciones son de tipo lineal, dicho análisis se denomina de correlación. Cuando disponemos de una única variable predictora, proponemos como medida de asociación lineal el coeficiente de correlación simple; cuando tenemos varios predictores, hemos de trabajar con los coeficientes de correlación parcial, pues el de correlación simple puede dar lugar a conclusiones engañosas.

4.2. Análisis gráfico de asociación

El primer paso a dar cuando pretendemos modelizar un banco de datos es llevar a cabo una inspección gráfica de los mismos, con el fin de descubrir de qué tipo son las relaciones entre las variables disponibles, si las hay. Básicamente, los gráficos de asociación son de dos tipos:

1. Gráficos de dispersión: sirven para visualizar relaciones entre un par de variables continuas.

2. Gráficos de cajas: sirven para visualizar relaciones entre una variable continua y un factor.

Así, por ejemplo, en la Figura 4.1 hemos representado varios pares de variables observadas. En los gráficos se aprecian distintos tipos de relación, más y menos lineal, entre pares de variables continuas.

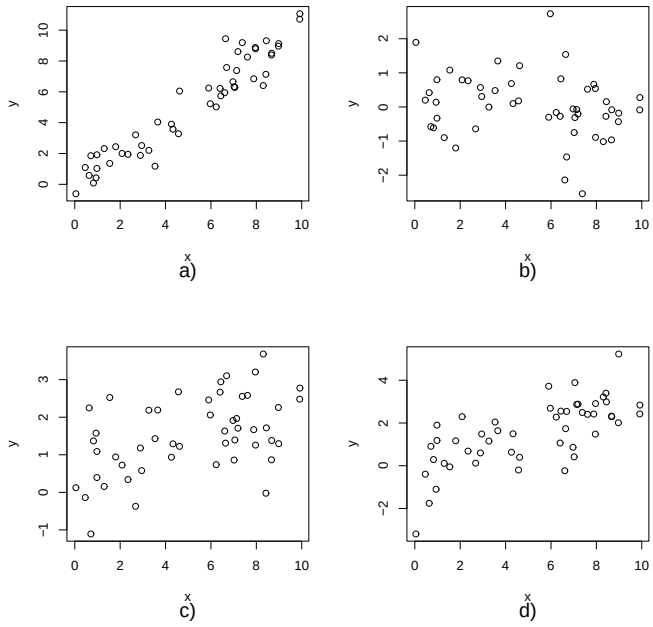


Figura 4.1: Diversos gráficos de dispersión.

Cuando la variable explicativa es de tipo factor (está categorizada), su asociación con la respuesta se detecta al comprobar si provoca diferentes respuestas en los diferentes niveles (categorías) observados. Tales diferencias se aprecian generalmente bien con los gráficos de cajas (boxplot). Estos gráficos nos dan información sobre la dispersión de los datos respuesta (amplitud de la caja con bigotes, de la caja sola, de los bigotes solos,...), sobre la simetría (en función de que la mediana divida a la caja con bigotes en dos mitades iguales o no, de la desigualdad de las partes de la caja separadas por la mediana, de la desigualdad de los bigotes, ...). Cuando representamos una caja para cada uno de los grupos identificados por una variable factor, tendremos información sobre la diferencia entre dichos grupos en función de que las cajas queden encajadas en un mismo rango de valores, o aparezcan desencajadas a distintas alturas.

En la Figura 4.2 tenemos representadas diversas variables respuesta clasificadas en dos grupos (factor explicativo con dos niveles de clasificación), y encontramos situaciones de clara diferencia entre los grupos (d), hasta semejanzas manifiestas (b) entre ambos. Los gráficos a) y c) no muestran grupos explícitamente diferentes; de hecho, el solapamiento de las cajas-bigotes en a) y la inclusión de la caja 1 en los bigotes de la caja 0, cuyos datos son mucho más dispersos, hacen intuir diferencias NO significativas.

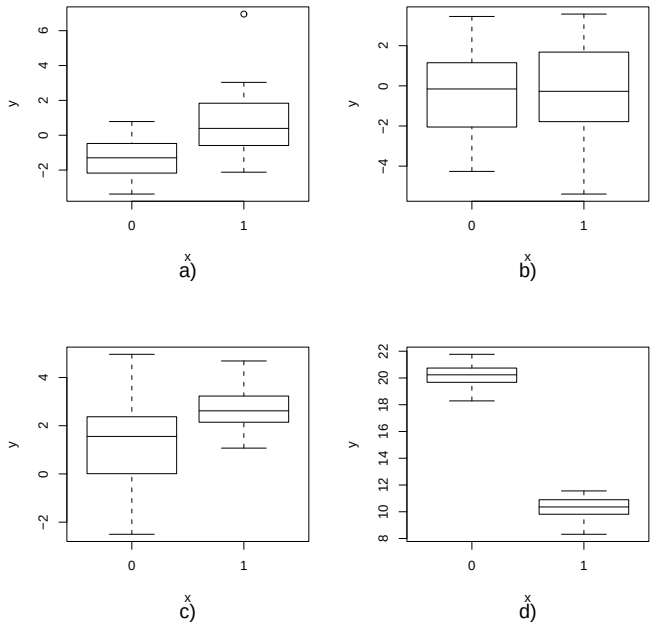


Figura 4.2: Diversos gráficos de cajas.

4.3. Análisis de Correlación

Nos concentramos ahora en el caso de que todas las variables involucradas son de naturaleza continua. Una vez hecha la inspección gráfica de los datos con los correspondientes diagramas de dispersión, hemos de valorar algo más objetivamente si el tipo de asociación se puede catalogar como lineal o no. En ocasiones, aun no apreciándose a priori una relación lineal, es posible transformar los datos con alguna función que linealice la relación. El conseguir una buena linealización del problema será resultado muchas veces de la experiencia acumulada.

Un análisis de correlación nos permitirá cuantificar el grado de asociación lineal entre variables, y en particular, entre las variables explicativas continuas disponibles y la respuesta de interés. Utilizaremos para ello el coeficiente de correlación simple cuando sólo estamos interesados en relacionar una variable explicativa con la respuesta. Cuando el número de variables explicativas sea superior, la medida a usar para medir el grado de asociación lineal será el coeficiente de correlación parcial.

4.4. Correlación Lineal Simple

Comenzamos con el caso más sencillo, en el que consideramos una variable respuesta continua y observada y una variable explicativa continua x. Un análisis de correlación servirá para poner de manifiesto si existe una relación de tipo lineal entre ambas variables. Otro tipo de relaciones no serán manifiestas a través de la correlación lineal. Si tal relación lineal existe, tenemos la justificación para ajustar un modelo de regresión lineal simple.

Se define el coeficiente de correlación (ρ) de Pearson como:

$$\rho = \frac{Cov(\mathrm{x}, \mathsf{y})}{\sqrt{Var(\mathrm{x})Var(\mathsf{y})}},$$

donde Var identifica la varianza y Cov la covarianza:

$$Cov(\mathrm{x}, \mathsf{y}) = E[(\mathrm{x} - E(\mathrm{x}))(\mathsf{y} - E(\mathsf{y}))].$$

Dada una muestra $(x_1, y_1), \ldots, (x_n, y_n)$ de dos variables x e y, se calcula el coeficiente de correlación lineal simple (r) de Pearson como:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}, \tag{4.1}$$

Cuando existe correlación lineal entre dos variables, este coeficiente es útil para cuantificar el grado de asociación entre ellas.

Así tenemos que el coeficiente de correlación de Pearson es una medida de asociación lineal libre de escala, con valores comprendidos entre -1 y 1, invariante a transformaciones lineales de las variables. En particular,

- $r = 0 \rightsquigarrow$ nada de asociación (lineal)
- $r = 1$ ó $-1 \rightsquigarrow$ asociación lineal perfecta
- $r < 0 \rightsquigarrow$ correlación negativa: $\uparrow x \Rightarrow \downarrow y$, cuando una aumenta de valor, la otra disminuye
- $r > 0 \rightsquigarrow$ correlación positiva: $\uparrow x \Rightarrow \uparrow y$, cuando una aumenta de valor, la otra también.

Ejemplo 4.1. *[Asociación lineal en 'Tractores'] Investigamos la relación lineal entre el dinero gastado en la manutención de tractores y la edad de los mismos,*

```
# Calculamos la correlación lineal de Pearson:
cor(edad,costes)   #= 0.6906927
```

y obtenemos una correlación lineal de Pearson en principio alta, si bien la linealidad era apreciable sólo por trozos en la Figura 3.1.

Ejercicio 4.1. *Simular cuatro bancos de datos, compuestos cada uno de observaciones de un par de variables* x *e* y*, más y menos relacionadas linealmente entre sí. En concreto, generar observaciones $x_1, x_2, \ldots, x_n$ e $y_1, y_2, \ldots, y_n$ a través de una distribución normal bidimensional, ligadas entre sí con una correlación lineal de: $\rho_1 = 0{,}2$, $\rho_2 = 0{,}8$, $\rho_3 = -0{,}5$ y $\rho_4 = -0{,}99$.*

4.4.1. Contraste de Hipótesis

Cuando queremos contrastar si dos variables x e y están relacionadas linealmente, planteamos el contraste:

$$\begin{cases} H_0 & : \rho = 0 \\ H_1 & : \rho \neq 0 \end{cases}$$

Para resolver dicho contraste podemos utilizar el estadístico de la z-transformada de Fisher,

$$Z = arctanh\ r = \frac{1}{2}\, log\left(\frac{1+r}{1-r}\right), \tag{4.2}$$

que se distribuye para muestras moderadas ($n \geq 25$), aproximadamente según una normal $N(arctanh(\rho), 1/(n-3))$.

Sin embargo, generalmente se usa un test t denominado de *correlación del producto de los momentos, de Pearson*:

$$r_t = r\sqrt{\frac{n-2}{1-r^2}} \sim t_{n-2}, \tag{4.3}$$

de modo que se rechazará H_0 a favor de que ambas variables estén relacionadas linealmente cuando:

$$|r_t| > t_{n-2;1-\alpha/2}.$$

Podríamos pretender también corroborar si la correlación es positiva o negativa, en cuyo caso se plantearían, respectivamente, los contrastes:

$$(CORR+)\begin{cases} H_0 & : \rho = 0 \\ H_1 & : \rho > 0 \end{cases} \qquad (CORR-)\begin{cases} H_0 & : \rho = 0 \\ H_1 & : \rho < 0 \end{cases}$$

cuyas regiones críticas vienen dadas respectivamente por:

$$\begin{array}{ll} (CORR+) & r_t > t_{n-2;1-\alpha} \\ (CORR-) & r_t < t_{n-2;\alpha} = 1 - t_{n-2;1-\alpha}. \end{array}$$

Ejemplo 4.2. *[Significatividad de la correlación en 'Tractores'] Queremos saber si es significativa la correlación entre edad del tractor y costes de manutención.*

```
cor.test(edad,costes)
#        Pearson's product-moment correlation
#
#data:  edad and costes
#t = 3.6992, df = 15, p-value = 0.002143
#alternative hypothesis: true correlation is not equal to 0
#95 percent confidence interval:
# 0.3144325 0.8793971
#sample estimates:
#      cor
#0.6906927
```

Según el test 4.3, aceptamos la alternativa ($H_1 : \rho \neq 0$), lo cual implica que hemos de rechazar correlación nula a favor de reconocer una relación lineal entre ambas variables.

Dado que la correlación es positiva, ¿podemos afirmar estadísticamente (a un nivel de confianza del 99 %) que existe una relación directa entre edad y costes, esto es, a más edad del tractor, mayores costes de manutención?

```
cor.test(edad,costes,alternative="greater",conf.level=0.99)
#        Pearson's product-moment correlation
#
#data:  edad and costes
#t = 3.6992, df = 15, p-value = 0.001071
#alternative hypothesis: true correlation is greater than 0
#99 percent confidence interval:
# 0.2236894 1.0000000
#sample estimates:
#      cor
#0.6906927
```

La respuesta es claramente que sí.

4.5. Transformaciones 'pro-linealidad'

Para comprobar linealidad entre las variables, tan importante es la inspección gráfica de los datos como el análisis de correlación. De hecho, este último puede resultar engañoso en ocasiones. Como ejemplo, observemos la Figura 4.3, donde aparecen cuatro bancos de datos (accesibles en el Apéndice B), todos ellos con correlación alta y significativamente distinta de cero, pero no todos con una relación lineal manifiesta, como en los bancos 'datos1', 'datos2' y algo en 'datos4'.

Una vez investigada la relación lineal entre la variable respuesta y la explicativa, si ésta falla, y antes de abordar la modelización, es recomendable considerar la posibilidad de realizar alguna transformación sobre las variables con el fin de linealizar la relación. Elegir una transformación apropiada es fruto de la experiencia que se va adquiriendo en el análisis de datos. Transformaciones habituales son los logaritmos (especialmente útiles cuando la escala de medida tiene un rango muy grande), exponenciales, raíces, ... Con todo, en la Figura 4.4 hemos ilustrado algunas funciones que nos pueden ayudar a decidir qué transformaciones abordar con el fin de obtener un par de variables relacionadas linealmente y sobre las que formular un modelo lineal.

Ejercicio 4.2. *Para las situaciones ilustradas en la Figura 4.4 sobre la relación entre un par de variables* `x` *e* `y`*, expresada en términos de la relación analítica correspondiente, formula transformaciones apropiadas para conseguir linealidad (de Montgomery y Peck, 1992, pag.90).*

Ejemplo 4.3. *Para las situaciones ilustradas en la Figura 4.3, cuyos datos están disponibles en el Apéndice B, vamos a contrastar correlación lineal y proponer transformaciones que linealicen la relación entre las variables cuando resulte necesario.*

```
# Leemos primeramente los datos, en el Apéndice.
#
# Para el banco 'datos1'
cor.test(x,y1)  # cor = 0.44126, p-valor = 0.01465
# La correlación lineal es significativa al 5%, pero no se aprecia
# lineal en el gráfico. Proponemos la transformación logarítmica:
cor.test(log(x),log(y1)) # cor=0.697845, p-valor =1.813e-05
# incrementando la correlación y la asociación lineal

# Para el banco 'datos2'
cor.test(x,y2)  # cor=0.8765235, p-valor=2.171e-10
# Probamos con el logaritmo de y2
cor.test(x,log(y2))  # cor=0.9962347, p-valor<2.2e-16
# y ganamos en linealidad.

# Para el banco 'datos3'
cor.test(x,y3)  # cor=0.979048, p-valor < 2.2e-16
# La linealidad es bastante patente.

# Para el banco 'datos4'
cor.test(x,y4)  # cor=0.979048, p-valor<2.2e-16
# aun claramente significativa, se aprecia cierta curva.
# Utilizamos la transformación inversa y
cor.test(1/x,y4)    # cor=0.9815995, p-valor<2.2e-16
# la ganancia en correlación revierte en ganancia en linealidad.
```

En todas las situaciones, salvo en 'datos3', a pesar de obtener correlaciones lineales significativas, la relación explícita en la Figura 4.3 arrojaba ciertas sospechas, más y menos notorias de falta de linealidad. Para esos datos, abordamos transformaciones que mejoran la correlación lineal, además de la linealidad, como queda patente en la Figura 4.5.

4.6. Correlación Múltiple y Parcial

Se define el coeficiente de correlación múltiple muestral $r_{1\ldots k}$ para k variables explicativas como aquel que cuantifica el grado de relación lineal existente entre una variable respuesta y el conjunto de esas k variables explicativas. Si S representa la matriz de covarianzas entre todas las variables involucradas, el coeficiente

de correlación múltiple se obtiene como:

$$r_{1\ldots k} = \frac{\left\{S_{12}S_{22}^{-1}S_{21}\right\}^{1/2}}{S_{11}^{1/2}}, \tag{4.4}$$

con S_{11} la parte de la matriz S correspondiente a la variable respuesta, S_{22} la parte de la matriz S correspondiente a las variables explicativas, S_{12} la parte de la matriz S que corresponde a las covarianzas entre la respuesta y las explicativas, y S_{21} la traspuesta de la anterior, $S_{21} = S_{12}'$.

Podemos además, realizar el contraste para verificar si el conjunto de variables explicativas está relacionado lineal y significativamente con la respuesta,

$$H_0 : \rho_{1\ldots k} = 0, \quad H_1 : \rho_{1\ldots k} \neq 0,$$

para lo que se define como estadístico de contraste:

$$\frac{(n-k-1)r_{1\ldots k}^2}{k(1-r_{1\ldots k}^2)} \sim F_{k,n-k-1}, \tag{4.5}$$

Ejemplo 4.4. *[Correlación múltiple en 'Bosque'] En el Ejemplo 3.2, pretendíamos relacionar el volumen de madera con diversas medidas sencillas del árbol. Indaguemos sobre la asociación de todas ellas a través del coeficiente de correlación múltiple:*

```
# Utilizamos la función rho.mult() del Apéndice Sintaxis en R.
rho.mult(bosque)   #=0.9793152
```

La correlación múltiple es pues significativamente distinta de 0, lo que, en principio, habla a favor de una asociación fuerte entre las variables medidas y el volumen de madera producido.

La siguiente pregunta a hacer es cuál de las variables explicativas guarda más relación (lineal) con la respuesta. Para contestarla hemos de acudir al coeficiente de correlación parcial.

Se define el coeficiente de correlación parcial muestral entre una variable respuesta y una explicativa, dado un bloque adicional de variables explicativas, como aquel que cuantifica el grado de relación lineal existente entre las dos primeras considerando (descontando) la relación (lineal) entre ambas y el bloque de variables adicionales. Su utilidad es decisiva para el ajuste de modelos de regresión múltiple, en los que contamos con un grupo de variables explicativas a

incorporar en el modelo de predicción. La correlación simple entre una variable explicativa y la respuesta sólo da información sobre la relación entre ellas, obviando toda relación con otras variables. La correlación parcial aísla la parte de las variables respuesta y explicativa que no está relacionada linealmente con las otras variables explicativas y da información sobre si dichos restos están relacionados linealmente, esto es, si la covariable en cuestión aporta algo de información (lineal) adicional sobre la respuesta, que no hayan aportado las otras variables.

El coeficiente de correlación parcial lo denotamos por:

$$r_{y.k|(-k)},$$

donde k identifica la variable explicativa cuya relación con la respuesta queremos conocer, x_k, y $(-k)$ identifica el bloque de variables restantes cuya relación con la respuesta descontamos, x_{-k}.

El coeficiente de correlación parcial se calcula a partir de la matriz de covarianzas S según:

$$r_{y.k|(-k)} = \frac{[S_{11,2}]_{yl}}{[S_{11,2}]_{yy}^{1/2}\,[S_{11,2}]_{ll}^{1/2}},$$

con $S_{11,2} = S_{11} - S_{12}S_{22}^{-1}S_{21}$, donde S_{11} es la parte de la matriz S correspondiente a la explicativa de interés, x_k, y la variable respuesta y, S_{22} es la parte de la matriz S correspondiente a las variables explicativas restantes (x_{-k}), y S_{12} es la parte de S que corresponde a las covarianzas entre los bloques $(\mathsf{y}, \mathrm{x}_k)$ y x_{-k}.

A continuación podemos realizar el contraste sobre la correlación parcial,

$$H_0 : \rho_{y.l|(q+1,\ldots,k)} = 0, \quad H_1 : \rho_{y.l|(q+1,\ldots,k)} \neq 0,$$

con el estadístico de contraste:

$$\frac{(n-(k-q)-1)^{1/2} r_{y.k|(-k)}}{1 - r_{y.k|(-k)}^2} \sim t_{n-(k-q)-1}, \tag{4.6}$$

Ejemplo 4.5. *[Correlación parcial en 'Bosque']*

Queremos identificar qué covariables están más relacionadas con la respuesta (VOL) en presencia de las restantes. Para ello podemos calcular todas las correlaciones parciales entre la respuesta y las covariables, en la primera fila (columna) de la siguiente matriz:

```
library(ggm)
parcor(cov(bosque))
#          VOL       DBH       D16       HT
#   VOL 1.0000000 0.3683119 0.7627127 0.7285511
```

Esto es,

$$r_{y.DBH|-DBH} = 0{,}3683119, \quad r_{y.D16|-D16} = 0{,}7627127, \quad r_{y.HT|-HT} = 0{,}7285511,$$

de modo que D16 parece ser la covariable más relacionada (linealmente) con VOL en presencia de las restantes, y DBH la que menos.

Por contra, si simplemente observábamos las correlaciones simples:

```
cor(bosque)
#            VOL       DBH       D16        HT
#   VOL 1.0000000 0.9078088 0.9530963 0.6010862
```

$$r_{y,DBH} = 0{,}9078088, \quad R_{y,DB16} = 0{,}9530963, \quad r_{y,HT} = 0{,}6010862$$

obteníamos que la respuesta VOL parecía muy relacionada linealmente con DBH. Esto se produce a consecuencia de que aquí no tenemos en cuenta a las restantes variables. De hecho, VOL y DBH están muy relacionadas linealmente con el bloque (DB16,HT), como indican sus correlaciones múltiples, ambas muy altas y significativas:

```
#  La correlación múltiple entre VOL y (DB16,HT) es
rho.mult(bosque[,-2])   #=0.976029
# y entre DBH y (DB16,HT)
rho.mult(bosque[,-1])   #=0.9267678,
```

Si queremos calcular exclusivamente la correlación parcial entre un par de variables, considerando su relación con un conjunto de variables fijas, podemos utilizar el comando `pcor()`*, también en la librería* `ggm` *de* $\mathcal{R}$ *.*

```
pcor(1:4,cov(bosque))  # VOL versus DBH, dado el resto
pcor(c(1,3,2,4),cov(bosque))  # VOL versus D16, dado el resto
```

Resultado 4.1. *Si* $\mathtt{y}$ *y* x_k *son ambas independientes del bloque de variables* x_{-k}*, entonces el coeficiente de correlación parcial* $r_{y.k|(-k)}$ *es igual al coeficiente de correlación simple* $r_{y,k}$*.*

Ejemplo 4.6. *Pretendemos corroborar la afirmación del Resultado 4.1. Para ello vamos a simular un banco de datos con tres variables* $\mathtt{y}, \mathrm{x}_1, \mathrm{x}_2$*, tales que* x_2 *sea independiente de* $\mathtt{y}$ *y de* x_1*. Calcularemos* $r_{y,1|-1}$ *y* $r_{y,1}$*, para verificar que son muy similares, esto es, que en ese caso coinciden los coeficientes de correlación parcial y simple.*

```
# Construimos la matriz de covarianzas para (Y,X1,X2),
#con X2 independiente de Y y de  X1:
cov.mat<-matrix(c(1,0.8,0,0.8,1,0,0,0,1),ncol=3)

# simulamos 50 datos:
datos<-rmvnorm(50,mean=c(0,0,0),sigma=cov.mat)
# calculamos la correlación parcial entre Y y X1
pcor(1:3,cov(datos))    #=0.7540396
# y la correlación simple
cor(datos[,1],datos[,2])    #=  0.7532037
# que salen claramente similares.

# La función correlations(), en la librería ggm,
# proporciona a la vez correlaciones parciales
# (en el triángulo superior) y simples
# (en el triángulo inferior)
correlations(cov(datos))
```

4.7. Resumen

Hemos estudiado básicamente en este tema el modo de descubrir asociaciones lineales entre variables de tipo continuo. Una vez ratificada la linealidad, tendremos la justificación adecuada para formular un modelo de predicción lineal. Cuando algún predictor es categórico, los gráficos de cajas nos ayudan a discernir si la clasificación que proporciona dicha variable afecta a la respuesta.

Un análisis de correlación debería constituir la primera aproximación a los datos de naturaleza continua cuando nos planteamos su modelización. El análisis de correlación consta de dos partes:

- Inspección gráfica de los datos: gráficos de dispersión.
- Análisis de correlación propiamente dicho, que consiste en el cálculo de las correlaciones y la realización de los contrastes pertinentes.

De hecho, ambas partes son imprescindibles, pues sólo utilizar una de ellas para ratificar linealidad y atreverse a formular un modelo lineal, puede ser peligroso en ocasiones. Una vez se aprecia la linealidad en el gráfico correspondiente, el coeficiente de correlación es útil para cuantificar el grado de asociación.

Cuando tenemos una variable respuesta y varias covariables continuas, hemos de inspeccionar la relación de todas ellas respecto de la respuesta, tanto individual como conjuntamente. Especialmente cuando las covariables guardan algo de

relación entre sí, la correlación simple puede proporcionar resultados engañosos respecto de la relación existente entre cada una de ellas con la respuesta. El coeficiente de correlación múltiple permite dar una medida global de la relación lineal que existe entre todo el conjunto de covariables y la respuesta. El coeficiente de correlación parcial proporciona información sobre el grado de asociación lineal entre la respuesta y una covariable en presencia de las restantes variables explicativas, por supuesto, como siempre, cuando efectivamente existe una relación lineal.

Ante problemas de linealidad, y antes de abordar una modelización lineal, conviene intentar linealizar mediante alguna transformación de las variables originales.

4.8. Ejercicios

1. Simula y representa apropiadamente las siguientes propuestas de datos. Concluye sobre el tipo de tendencia/relación apreciada.

   a) $\mathrm{x} \sim Bi(40, 0{,}6)$ e $\mathsf{y} \sim N(1 + 2\mathrm{x}, 2)$

   b) $\mathrm{x} \sim Ber(0{,}6)$ e $\mathsf{y} \sim N(1 + 2\mathrm{x}, 2)$

   c) $\mathrm{x} \sim Bi(40, 0{,}6)$ e $\mathsf{y} \sim N(1 + 2log(\mathrm{x}), 3)$

   d) $\mathrm{x} \sim Un(0, 1)$ e $\mathsf{y} \sim N(1 + 2exp(\mathrm{x}), 2)$

   e) $\mathrm{x} \sim Ber(0{,}6)$, $\mathrm{z} \sim Un(0, 4)$ e $\mathsf{y} \sim N(1 + 2\mathrm{zx}, 2)$

   f) $\mathrm{x} \sim Ber(0{,}6)$, $\mathrm{z} \sim Un(0, 4)$ e $\mathsf{y} \sim N(1 + 2\mathrm{x}\sqrt{\mathrm{z}}, 2)$

2. Para las situaciones presentadas en la Figura 4.6, sabemos que los coeficientes de correlación muestrales son $-0{,}85; -0{,}38; -1; 0{,}06; 0{,}97; 0{,}62$. Identifica cada coeficiente con el gráfico que corresponde.

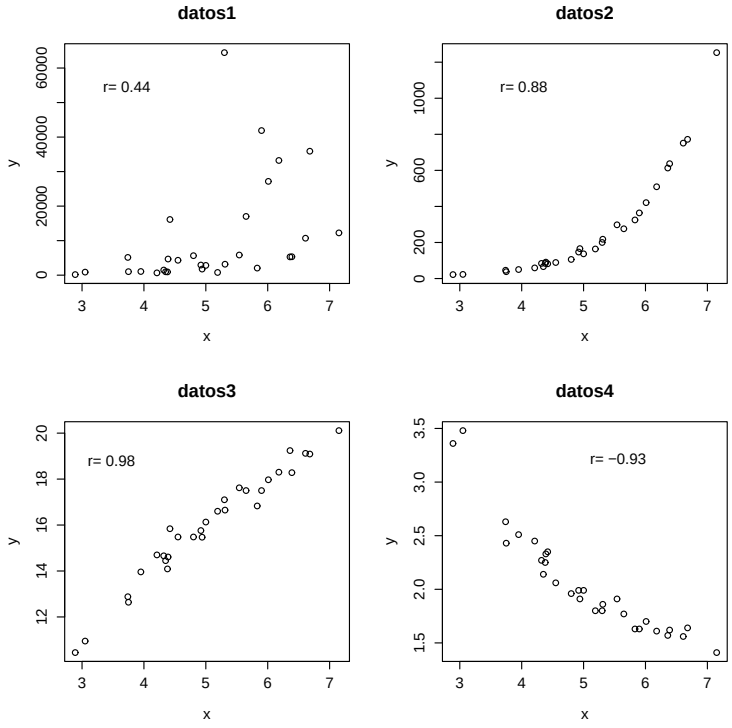


Figura 4.3: ¿Relación lineal con correlación lineal alta?

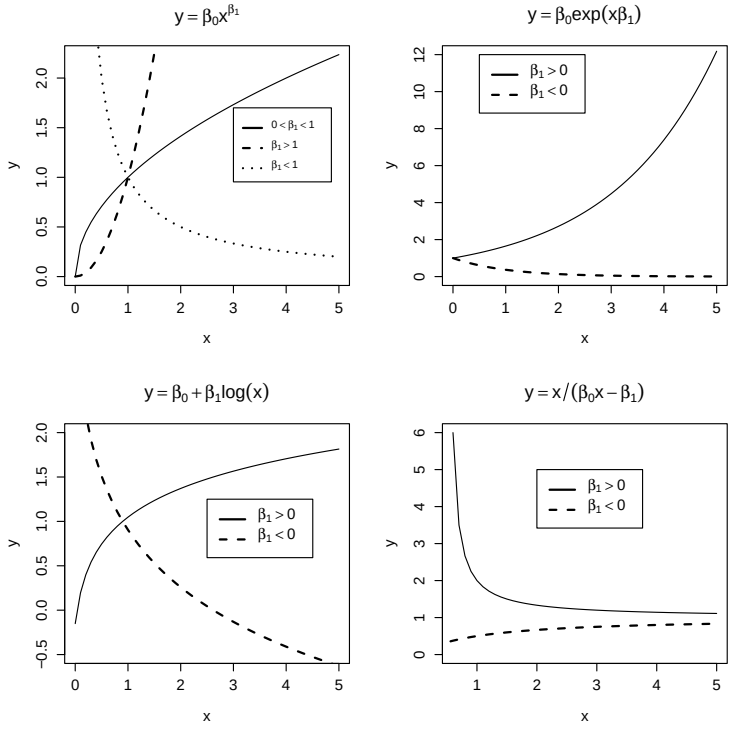


Figura 4.4: Relaciones linealizables.

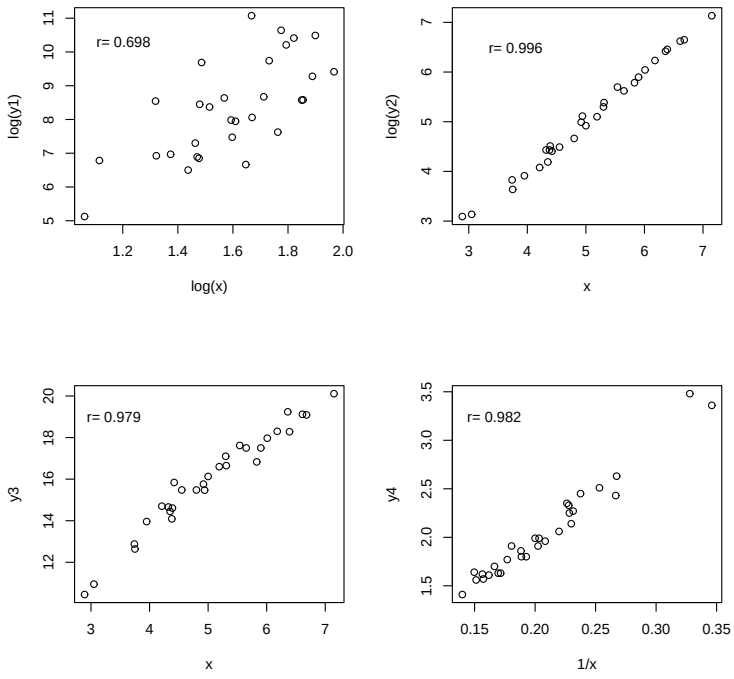


Figura 4.5: Linealización de datos1, datos2 y datos4.

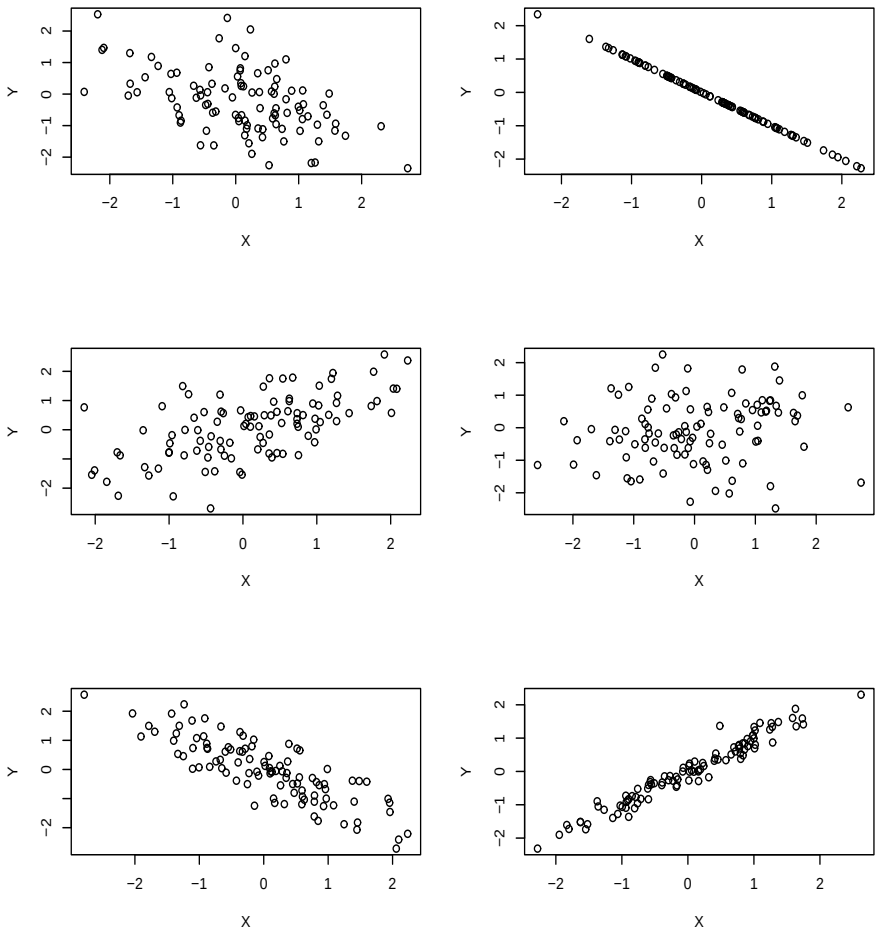


Figura 4.6: Correlación lineal.

Tema 5

El Modelo de Regresión Lineal Simple (RLS)

5.1. Introducción

En el Tema anterior nos referíamos al análisis de correlación como un paso previo a la modelización, útil para verificar relaciones lineales entre las variables. Pues bien, cabe resaltar aquí que, cuando uno lleva a cabo un análisis de correlación, no necesita tener clara la distinción entre variables respuesta y explicativas; simplemente se indaga la relación entre las variables y se concluye sobre si es lineal o no. Cuando uno pretende modelizar, ha de especificar, en base a los objetivos del estudio, cuál es la variable que pretende explicar o predecir y cuáles son las que le van a ayudar a hacerlo, bien porque sean causas o simplemente porque estén relacionadas con la primera y contengan información relevante para poder predecirla.

Nos preocupamos en este tema del Modelo de Regresión Lineal Simple (RLS), que podemos catalogar como el modelo lineal más sencillo, a través del cual pretendemos explicar (predecir) una variable respuesta continua y a partir de una variable *explicativa* también continua x. Por supuesto, asumimos que existe una relación lineal entre ellas, que queremos captar a través de un *modelo de regresión*.

En el experimento o estudio del que obtenemos los datos, los valores de $\mathbf{y}$ se han observado y los de $\mathbf{x}$, bien se han observado, bien se han prefijado por parte del investigador. En cualquier caso, asumimos que la aleatoriedad (incertidum-

bre) está contenida sólo en la variable **y**, mientras que la **x** carece de aleatoriedad y simplemente informa de lo que ocurre en los valores observados.

5.2. Por qué el nombre de REGRESIÓN

A principios del siglo XX, el estudioso de la genética Francis Galton descubrió un fenómeno llamado regresión a la media. Buscando leyes de herencia genética, descubrió que la estatura de los hijos solía ser una regresión a la estatura media poblacional, en comparación con la estatura de sus padres. Los padres altos solían tener hijos algo más bajos, y viceversa. Galton (1894) desarrolló el análisis de regresión para estudiar este fenómeno, al que se refirió de manera optimista como "regresión a la mediocridad".

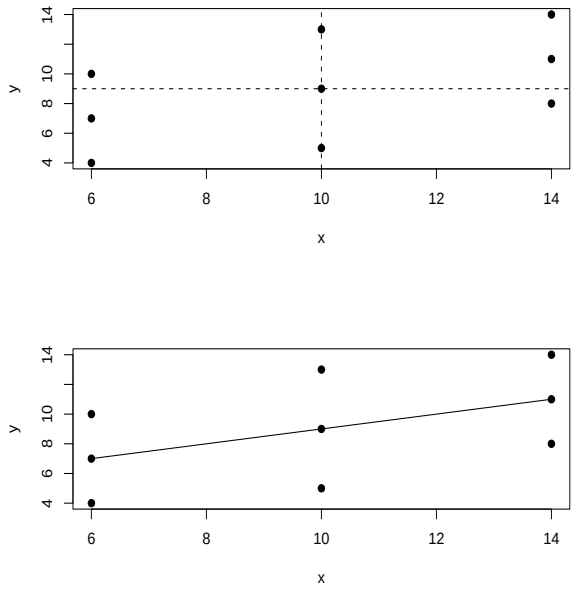


Figura 5.1: Regresión a la media.

Con el fin de mostrar fácilmente en qué consiste la recta de regresión, consideremos los puntos representados en el gráfico superior de la Figura 5.1. Tenemos una serie de 9 observaciones de pares (x_i, y_i); en concreto, tan sólo tenemos tres valores fijos de la variable **x**. Para cada valor de **x**, hemos observado tres valores de **y**, curiosamente equidistantes, de forma que el punto medio representa la media de los valores **y** observados. La tendencia lineal se aprecia claramente en el

gráfico. A la hora de ajustar una recta de regresión para captar dicha tendencia, la recta atravesará los puntos pasando por los puntos medios (regresión a la media), como se ilustra en el gráfico inferior de la Figura 5.1.

5.3. Formulación del modelo RLS

El modelo de Regresión lineal Simple de y sobre x se formula según:

$$\mathsf{y} = \beta_0 + \beta_1 \mathrm{x} + \epsilon, \tag{5.1}$$

de forma que, para un valor dado de $\mathrm{x} = x$, ϵ representa una desviación aleatoria de la respuesta y sobre el valor esperado según la recta de regresión:

$$E(\mathsf{y}|\mathrm{x} = x) = \beta_0 + \beta_1 x. \tag{5.2}$$

Los coeficientes de la regresión, esto es, los parámetros que estimaremos para ajustar el modelo RLS son:

β_0**.-** la *interceptación* de la recta, esto es, la altura de la recta cuando $\mathrm{x} = 0$.

β_1**.-** la *pendiente* de la recta, que refleja cuánto varía la respuesta media $E(\mathsf{y})$ cuando pasamos de observar $\mathrm{x} = x$ a $\mathrm{x} = x + 1$.

Dada una muestra de valores observados $\{(x_i, y_i)\}_{i=1}^n$, el modelo (5.1) implica que todas las observaciones responden a:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \qquad i = 1, \ldots, n,$$

donde ϵ_i son errores aleatorios e incorrelados, con media cero y varianza constante σ^2, características que identifican las hipótesis básicas del modelo RLS, que formulamos con más detalle a continuación sobre los errores aleatorios ϵ_i:

Incorrelación: $Corr(\epsilon_i, \epsilon_j) = 0$**.** Significa que las observaciones de la respuesta y, $y_1, y_2, \ldots, y_n$ están incorreladas entre sí, esto es, los valores de unas no afectan a los de otras.

Media cero: $E(\epsilon_i) = 0$**.** Lo que implica que la respuesta esperada según el modelo RLS depende linealmente de los coeficientes de regresión β_0 y β_1, al tener como expresión (5.2).

Varianza constante: $Var(\epsilon_i) = \sigma^2$**.** Lo que significa que las observaciones $\{y_i, i = 1, \ldots, n\}$ provienen de una misma población cuya variabilidad respecto de sus medias $\{\beta_0 + \beta_1 x_i, i = 1, \ldots, n\}$ viene dada por σ^2.

5.4. Estimación de la recta de regresión

Estimar la recta de regresión consiste en estimar los coeficientes de la regresión β_0 y β_1 para obtener la recta:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x, \tag{5.3}$$

donde $\hat{y}$ denota el valor de y predicho por la recta para el valor observado de $\mathrm{x} = x$.

Disponemos de dos criterios básicos de estimación, que proporcionan la misma solución. Utilizar uno u otro depende de nuestros intereses estadísticos. Si tan sólo queremos determinar la recta, basta con considerar el criterio de *Mínimos Cuadrados*. Si además pretendemos utilizarla con fines inferenciales o predictivos, hablaremos de que nuestra solución es la *máximo-verosímil*, y las hipótesis del modelo se radicalizarán más al imponer normalidad en los errores, como veremos.

Criterio 1: Mínimos Cuadrados o minimización del error cuadrático medio. Consiste en minimizar las distancias entre los puntos observados y los predichos por la recta de ajuste, como se ilustra en la Figura 5.2. El error cuadrático medio de la recta se define como:

$$S(\beta) = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} [y_i - (\beta_0 + \beta_1 x_i)]^2 \tag{5.4}$$

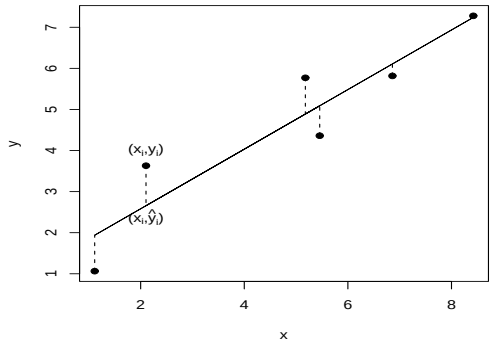


Figura 5.2: Mínimos cuadrados en regresión.

La solución de mínimos cuadrados $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1)$ se obtiene minimizando $S(\beta)$,

esto es, derivando respecto de β_0 y β_1 e igualando a cero:

$$\frac{\partial S(\beta)}{\partial \beta_0} \mid_{\hat{\beta}_0,\hat{\beta}_1} = -2 \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \tag{5.5}$$
$$\frac{\partial S(\beta)}{\partial \beta_1} \mid_{\hat{\beta}_0,\hat{\beta}_1} = -2 \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) x_i = 0.$$

De ahí se obtienen las *ecuaciones normales*:

$$n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i \tag{5.6}$$
$$\hat{\beta}_0 \sum_{i=1}^{n} x_i + \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} y_i x_i$$

de donde se consiguen las estimaciones para β_0 y β_1:

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \tag{5.7}$$
$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}},$$

con:

$$\bar{y} = \sum_{i=1}^{n} y_i/n \qquad \bar{x} = \sum_{i=1}^{n} x_i/n$$
$$S_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2 \qquad S_{xy} = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}). \tag{5.8}$$

Ejemplo 5.1. *Obtener el ajuste de mínimos cuadrados para los datos de Tractores (Ejemplo 3.1).*

```
x<-edad; y<-costes
# La estimación de mínimos cuadrados se obtiene en R con
mc<-lsfit(x,y)$coefficients;mc
# Pintamos los datos y superponemos la recta
plot(x,y)
abline(a=mc[1],b=mc[2])
```

La recta de mínimos cuadrados resulta $\hat{costes} = 323{,}6 + 131{,}7\, edad$*, esto es, un año más de antigüedad del tractor reporta un gasto adicional de 131.72 (\$). En la Figura 5.3 visualizamos el resultado de la recta de mínimos cuadrados ajustada, que captura la tendencia lineal apreciada en los datos.*

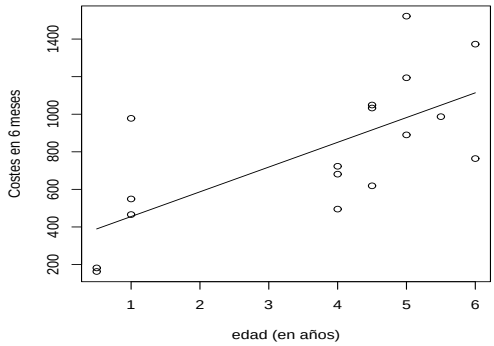


Figura 5.3: Ajuste de mínimos cuadrados en 'Tractores'.

Ejemplo 5.2. *Queremos verificar que, cuando la correlación entre un conjunto de valores* $\mathbf{x}$ *y otro* $\mathbf{y}$ *es fija, tenemos que:*

- *a mayor dispersión de las* $\mathbf{y}$*'s, mayor pendiente de la recta,*
- *a mayor dispersión de los* $\mathbf{x}$*'s, menor pendiente de la recta.*

Para ello, simulamos tres bancos de datos normales bivariantes $\{(x_i, y_i), i = 1, \ldots, 50\}$ *con correlación 0.7, tales que:*

B1	$Var(\mathbf{x}) = 1 = Var(\mathbf{y})$
B2	$Var(\mathbf{x}) = 1, Var(\mathbf{y}) = 5$
B3	$Var(\mathbf{x}) = 5, Var(\mathbf{y}) = 1$

Ajustamos rectas de regresión y las representamos sobre los gráficos de dispersión para apreciar el efecto sobre la pendiente (ver Figura 5.4)

```
library(mvtnorm)
rho<-0.7    # correlación
mu<-c(0,0)  # media

# Función para calcular la matriz de varianzas-covarianzas
# a partir de varianzas y correlación.
varcovar<-function(vx,vy,rho){
covar<-rho*sqrt(vx*vy)  # covarianza (X,Y)
sigma<-matrix(c(vx,covar,covar,vy),nrow=2)
return(sigma) }

# Simulamos los datos y ajustamos una recta de regresión
b1<-rmvnorm(50,mu,varcovar(1,1,rho));fit1<-lsfit(b1[,1],b1[,2])
b2<-rmvnorm(50,mu,varcovar(1,5,rho));fit2<-lsfit(b2[,1],b2[,2])
b3<-rmvnorm(50,mu,varcovar(5,1,rho)); fit3<-lsfit(b3[,1],b3[,2])

#  Dibujamos dicha recta sobre los gráficos de dispersión
# para apreciar cómo cambia la pendiente: par(mfrow=c(2,2))
plot(b1,xlim=c(-5,5),ylim=c(-5,5),xlab="X",ylab="Y",
main="Var(X)=1,Var(Y)=1") abline(a=coef(fit1)[1],b=coef(fit1)[2])
plot(b2,xlim=c(-5,5),ylim=c(-5,5),xlab="X",ylab="Y",
main="Var(X)=5,Var(Y)=5") abline(a=coef(fit2)[1],b=coef(fit2)[2])
plot(b1,xlim=c(-5,5),ylim=c(-5,5),xlab="X",ylab="Y",
main="Var(X)=1,Var(Y)=1") abline(a=coef(fit1)[1],b=coef(fit1)[2])
plot(b3,xlim=c(-5,5),ylim=c(-5,5),xlab="X",ylab="Y",
main="Var(X)=1,Var(Y)=5") abline(a=coef(fit3)[1],b=coef(fit3)[2])
```

Ejercicio 5.1. *Consultar el applet 'Correlation and Regression Demo' en la dirección web* `http://bcs.whfreeman.com/ips4e/`*, en el apartado Statistical Applets.*

Criterio 2: Máxima Verosimilitud**.** Habitualmente el objetivo de un análisis de regresión no consiste únicamente en estimar la recta, sino en *inferir* con ella, esto es, asociar un error a las estimaciones obtenidas, contrastar un determinado valor de los parámetros, o predecir la respuesta para un x dado junto con una banda de confianza. En ese caso, precisamos de distribuciones de probabilidad para controlar la incertidumbre. Añadimos pues, una hipótesis más sobre la distribución de la variable respuesta, o lo que es lo mismo, sobre el error aleatorio ϵ. Dicha hipótesis es la de normalidad de los errores.

Así, el total de hipótesis básicas del modelo de regresión con fines inferenciales, viene resumido en la siguiente expresión:

$$\epsilon_i \overset{iid}{\sim} N(0, \sigma^2), \qquad i = 1, \ldots, n. \tag{5.9}$$

esto es, hablamos de errores aleatorios independientes e idénticamente distribuidos (iid) según una distribución Normal con media cero y varianza σ^2, lo que

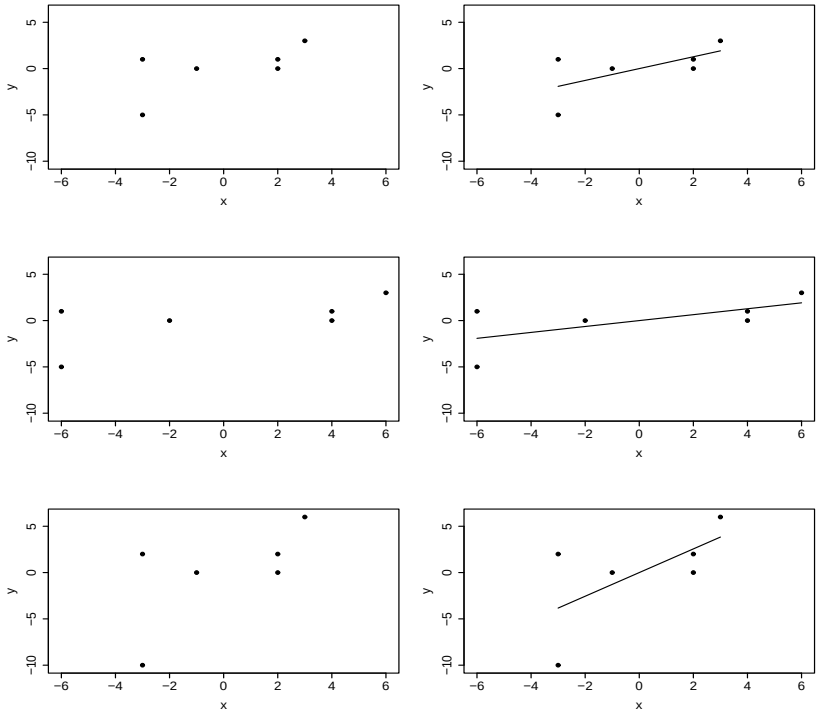


Figura 5.4: Influencia de la dispersión de las variables sobre la pendiente de la recta de regresión

implica directamente que la distribución para la variable respuesta será:

$$\mathsf{y}_i \overset{iid}{\sim} N(\beta_0 + \beta_1 x_i, \sigma^2), \qquad i = 1, \ldots, n.$$

Desde este momento, los datos proporcionan información sobre los parámetros del modelo, $\beta = (\beta_0, \beta_1)$, a través de la verosimilitud conjunta:

$$L(\beta; \mathbf{y}) = exp\left\{ -\frac{\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2}{2\sigma^2} \right\}. \tag{5.10}$$

Obtener la solución más factible a la vista de los datos observados $\{(x_i, y_i), i = 1, \ldots, n\}$ equivale a obtener la solución máximo-verosímil, esto es, la que maximiza la verosimilitud (5.10). Maximizar la verosimilitud es equivalente a maximizar la log-verosimilitud $l(\beta, \mathbf{y})$, que tiene una expresión más sencilla sin exponenciales. La solución máximo-verosímil se obtiene derivando e igualando a cero $l(\beta, \mathbf{y})$, lo que da lugar, de nuevo, a las ecuaciones normales (5.6). Así pues, la solución máximo-verosímil coincide con la de mínimos cuadrados (5.7).

Ejemplo 5.3. *Obtener el ajuste por máxima verosimilitud para los datos del Ejemplo 3.1.*

```
x<-edad; y<-costes
emv<-lm(y~x)$coefficients;emv
```

La recta estimada resulta, nuevamente, $\hat{costes} = 323{,}6 + 131{,}7\, edad$.

Ejercicio 5.2. *Comprobar, manualmente a partir de la expresión para la log-verosimilitud, que la estimación máximo-verosímil conduce a las ecuaciones normales* (5.6)*. Para ello, derivar e igualar a cero la log-verosimilitud.*

5.5. Propiedades del ajuste de la recta de regresión.

Las propiedades más relevantes y básicas del ajuste de la recta de regresión son las siguientes:

1. Las estimación de la respuesta para un valor de $\mathbf{x} = x$ concreto según el modelo de regresión lineal simple (5.1) se obtiene de la recta de regresión ajustada:
$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x. \tag{5.11}$$

2. La suma de los residuos de una recta de regresión con término de interceptación β_0 es cero,
$$e_i = y_i - \hat{y} \rightsquigarrow \sum_i e_i = 0.$$

3. La media de los valores observados y_i coincide con la media de los valores predichos $\hat{y}_i$,
$$\frac{1}{n}\sum_i y_i = \frac{1}{n}\sum_i \hat{y}_i. \tag{5.12}$$

4. La recta de regresión pasa por el centroide de medias $(\bar{x}, \bar{y})$.

5. La suma de los residuos ponderados por el valor correspondiente de la variable predictora $\mathbf{x}$ es cero,
$$\sum_i x_i e_i = 0.$$

6. La suma de los residuos ponderados por el valor ajustado por la recta $\hat{\mathbf{y}}$ es cero,

$$\sum_i \hat{y}_i e_i = 0.$$

5.6. Estimación de σ^2.

La varianza σ^2 de los errores es una medida de la variabilidad (heterogeneidad) entre los individuos respecto a la media cuando el modelo RLS describe adecuadamente la tendencia entre las variables y y x, o lo que es lo mismo, de la dispersión de las observaciones respecto de la recta de regresión. Así pues, da una medida de bondad de ajuste del modelo de regresión a los datos observados. Cuando el modelo de regresión (5.1) es bueno, es posible conseguir una estimación de la varianza σ^2 a partir de la *suma de cuadrados residual* SSE, también llamada *suma de cuadrados debida al error*:

$$SSE = \sum_i (y_i - \hat{y}_i)^2 = S_{yy} - \hat{\beta}_1 S_{xy}.$$

SSE da una medida de la desviación entre las observaciones y_i y las estimaciones que proporciona la recta de regresión (5.11), $\hat{y}_i$. Puesto que en el modelo de regresión lineal simple se estiman 2 parámetros, los grados de libertad asociados a SSE son $n-2$. Se define pues el *cuadrado medio residual*, MSE, como un estimador de σ^2, que además resulta ser insesgado (esto es, su valor esperado es σ^2):

$$s^2 = MSE = \frac{SSE}{n-2}. \tag{5.13}$$

El error estándar residual viene dado por $s = \sqrt{MSE}$.

Ejemplo 5.4 (Variabilidad de errores en el ajuste para 'Tractores'). *¿Cuál es la variabilidad residual del modelo de regresión para los datos de Tractores? ¿Es posible reducir dicha variabilidad planteando otros modelos de regresión basados en transformaciones de los datos originales?*

```
# Sobre el modelo ajustado
x<-edad; y<-costes
fit<-lm(y~x)
# la estimación de la varianza se consigue a partir del
# resumen del ajuste, que se obtiene con el comando
sfit<-summary(fit)
# el error estándar residual es de 283.5
# y la varianza, su cuadrado:
sfit$sigma^2    #=80360.47
# que resulta considerablemente alta.

# Sin embargo, si consideramos una transformación de los datos
# con el logaritmo de los costes (cuyo rango es grande)
# en busca de mejorar la linealidad y por tanto la calidad del
# ajuste, tenemos:
cor.test(x,log(y))      #=0.7353647, que incrementa la cor(x,y)=0.69
fit.log<-lm(log(y)~x)
summary(fit.log)
# reduce considerablemente el error de los residuos a 0.4375, y
# por consiguiente la varianza del error:
summary(fit.log)$sigma^2    #=0.1913944
```

5.7. Inferencia sobre $\hat{\beta}_0$ y $\hat{\beta}_1$

Los estimadores de mínimos cuadrados $\hat{\beta}_0$ y $\hat{\beta}_1$ son insesgados y de mínima varianza de entre todos los estimadores insesgados. El hecho de especificar una distribución normal sobre los errores para la estimación máximo-verosímil, permite derivar de forma directa la distribución de dichos estimadores, que resulta también normal:

$$\begin{aligned} \hat{\beta}_0 &\sim N\left(\beta_0, \frac{\sum_{i=1}^n x_i^2}{nS_{xx}}\ \sigma^2\right) \\ \hat{\beta}_1 &\sim N\left(\beta_1, \frac{\sigma^2}{S_{xx}}\right), \end{aligned} \tag{5.14}$$

con S_{xx} dado en (5.8). Cuando el modelo de regresión es adecuado, podemos estimar las varianzas de dichas distribuciones sustituyendo σ^2 por s^2 en (5.13). De ahí podemos construir los **estadísticos** t para inferir sobre los parámetros:

$$\begin{aligned} t_0 &= \frac{\hat{\beta}_0 - \beta_0}{s\ \sqrt{\sum_i x_i^2/nS_{xx}}} \\ t_1 &= \frac{\hat{\beta}_1 - \beta_1}{s/\sqrt{S_{xx}}}. \end{aligned} \tag{5.15}$$

Ambos estadísticos se distribuyen según una distribución t con $n-2$ grados de libertad, que nos permite inferir (estimar y resolver contrastes de hipótesis) sobre los coeficientes del modelo, y en particular contestar a preguntas sobre la relación entre las variables respuesta y explicativa.

5.7.1. Estimación puntual y en intervalos

Las estimaciones puntuales de β_0 y β_1 las obtenemos directamente de (5.7).

Los **intervalos de confianza** al nivel de confianza $(1-\alpha)10\,\%$ para β_0 y β_1 se construyen a partir de los estadísticos t (5.15) y resultan:

$$IC(\beta_0; 1-\alpha) = \hat{\beta}_0 \pm t_{\left(n-2,1-\frac{\alpha}{2}\right)} \sqrt{\frac{\sum_{i=1}^n x_i^2}{nS_{xx}}\, s^2} \tag{5.16}$$
$$IC(\beta_1; 1-\alpha) = \hat{\beta}_1 \pm t_{\left(n-2,1-\frac{\alpha}{2}\right)} \sqrt{\frac{s^2}{S_{xx}}},$$

donde $t_{\left(n-2,1-\frac{\alpha}{2}\right)}$ es el cuantil $1-\alpha/2$ de una distribución t con $n-2$ grados de libertad (los correspondientes a s^2).

5.7.2. Contrastes de Hipótesis

Si queremos **contrastar hipótesis** sobre los coeficientes de la regresión:

$$\begin{aligned} &H_0 : \beta_i = \beta^* \\ &H_1 : \beta_i \neq \beta^*, \quad i = 0, 1 \end{aligned} \tag{5.17}$$

basta con considerar los estadísticos correspondientes (5.15), sustituyendo el valor β_i por el que se pretende contrastar, β^*. Estos estadísticos, bajo H_0, tienen una distribución t con $n-2$ grados de libertad. La resolución del contraste consiste en calcular el p-valor asociado al valor absoluto de la estimación, $|t_0|$ o $|t_1|$, según el caso, esto es, $p-valor = Pr[t_{n-2} > |t_i|]$. El contraste se resuelve de la siguiente forma:

- se rechaza H_0 a nivel α cuando $p-valor \leq \alpha$,
- si $p-valor > \alpha$, se dice que los datos no proporcionan suficientes evidencias en contra de la hipótesis nula y ésta no se puede rechazar.

Cuando el contraste propuesto sobre β_0 o β_1 tiene $\beta^* = 0$, en realidad se está contrastando, respectivamente, si la recta de regresión tiene interceptación o pendiente nula. Contrastar $\beta_1 = 0$ es equivalente a contrastar correlación nula $\rho_{xy} = 0$ entre las variables **x** e **y**, esto es, ausencia de relación lineal. Si conseguimos rechazar esta hipótesis con significatividad, concluiremos que la variable **x** está relacionada linealmente con **y** y por lo tanto se puede utilizar para predecir **y** a través de la recta de regresión ajustada.

Ejemplo 5.5 (Inferencia con el ajuste RLS para 'Tractores')**.** *Con el mejor modelo RLS obtenido hasta el momento para predecir los costes en función de la edad del camión:*

1. *Construir intervalos de confianza al 99 % para β_0 y β_1. ¿Qué podemos decir de la relación entre dichas variables?*

2. *Concluir sobre los contrastes $\beta_0 = 0$ y $\beta_1 = 0$. Comprobar también que el último contraste $\beta_1 = 0$ es equivalente al contraste de correlación nula entre las variables del modelo.*

```
x<-edad;y<-costes
fit<-lm(log(y)~x)
# Los coeficientes estimados se obtienen con:
coef(fit)
# (Intercept)           x
#   5.7008492   0.2309455

# Los errores estándar asociados a los coeficientes son:
sfit<-summary(fit)
sfit$coefficients[,2]   #= 0.22676720  0.05495161 respect.
# Los intervalos de confianza para los parámetros se obtienen con:
confint(fit)

# Los p-valores asociados al contraste beta=0 para cada coeficiente
# se visualizan en la tabla que proporciona
summary(fit)

# El test de correlación entre x y log(y) da como resultado:
cor.test(x,log(y))   # p-valor=0.0007687
# similar al p-valor asociado a beta1 = 0.000769.
```

La recta ajustada es

$$\widehat{log(costes)} = 5{,}7 + 0{,}231\ edad.$$

Los intervalos de confianza para β_0 y β_1 al nivel de confianza del 95 % resultan:

$$IC(\beta_0, 99\,\%) = (5{,}03, 6{,}369)$$
$$IC(\beta_0, 99\,\%) = (0{,}069, 0{,}393),$$

ninguno de los cuales incluye al cero, lo que habla positivamente de su significatividad estadística, esto es, predecimos el logaritmo de los costes con la edad de los vehículos a través de una recta con interceptación y pendientes (significativamente) distintas de cero. De hecho, la relación entre el logaritmo de los costes y la edad es directa, como se concluye del signo y la magnitud del coeficiente estimado β_1.

Puestos a resolver el contraste $\beta_i = 0$, para $i = 0, 1$, obtenemos los siguientes valores para los estadísticos t (5.15) y sus p-valores asociados:

$$t_0 = 25{,}140 \qquad p - valor = 1{,}12e - 13$$
$$t_1 = 4{,}203 \qquad p - valor = 0{,}000769,$$

lo que concluye contundentemente sobre la significatividad de ambos a favor de que son distintos de cero (se rechaza H_0), como ya habíamos comentado a partir de los intervalos de confianza. En particular, la edad explica significativamente el logaritmo de los costes a través del modelo lineal ajustado.

Efectivamente, el test de correlación cero y el correspondiente a $\beta_1 = 0$ están basados en el mismo estadístico t, $t_1 = 4{,}203$, proporcionando por tanto el mismo p-valor de 0,000769.

5.7.3. Estimación de la respuesta media

Si queremos estimar el valor medio de la variable y cuando la variable x toma un valor concreto x = x_0 dentro del rango de valores observados, basta con sustituir dicho valor en la recta de regresión ajustada:

$$\hat{y}_{x_0} = \hat{\beta}_0 + \hat{\beta}_1 x_0.$$

Un intervalo de confianza para la estimación del valor esperado de la respuesta para un x_0 dado es:

$$IC(E(\bar{\mathrm{y}}_n|x_0); 1-\alpha) = \hat{y}_{x_0} \pm t_{\left(n-2,1-\frac{\alpha}{2}\right)}\ s\ \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}, \tag{5.18}$$

con S_{xx} dado en (5.8) y s en (5.13).

5.7.4. Predicción de nuevas observaciones

Predeciremos una futura observación de la variable y para cierto valor de x = x_0, con

$$\hat{y}_{x_0} = \hat{\beta}_0 + \hat{\beta}_1 x_0 = \bar{y} + \hat{\beta}_1(x_0 - \bar{x}),$$

y el intervalo de confianza vendrá dado por:

$$IC(y_{x_0}; 1-\alpha) = \hat{y}_{x_0} \pm t_{\left(n-2,1-\frac{\alpha}{2}\right)}\ s\ \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}, \tag{5.19}$$

con S_{xx} dado en (5.8) y s en (5.13).

Notar que tanto la estimación de la respuesta media como la predicción coinciden, si bien difieren en cuanto al grado de incertidumbre de la misma. Como es de esperar, predecir un hecho puntual en el futuro conlleva más incertidumbre que estimar en términos medios qué va a a ocurrir.

Ejemplo 5.6 (Estimación media y predicción con el ajuste de 'Tractores')**.** *Estimar la respuesta media y la predicción de una nueva observación para una secuencia de 20 puntos equidistantes en el rango de valores observados para la variable explicativa. Representar las bandas de confianza tanto en la escala de la variable respuesta del modelo como de la variable original.*

```
y<-costes; x<-edad
fit<-lm(log(y)~x)
# Queremos estimar y predecir con la recta de regresión
# los costes de manutención para diversas edades:
x0<-seq(min(x),max(x),length=20)

# La estimación completa de la respuesta media se obtiene con:
pred.m<-predict(fit,data.frame(x=x0),interval="confidence",se.fit=T)

# La predicción completa de una nueva observación se obtiene con:
pred.p<-predict(fit,data.frame(x=x0),interval="prediction",se.fit=T)

# Dibujamos los intervalos de confianza para estimación y predicción,
#en la escala del log(costes):
par(mfrow=c(1,2))
matplot(x0,cbind(pred.m$fit,pred.p$fit[,-1]),lty=c(1,2,2,3,3),
col=c("black","red","red","blue","blue"),type="l",xlab="edad",
ylab="log(costes)")
legend(1,8.1,c("Estim.Media","Predicción"),lty=c(2,3),
col=c("red","blue"))
points(x,log(y))

# Repetimos las bandas en la escala de costes:
matplot(x0,cbind(exp(pred.m$fit),exp(pred.p$fit[,-1])),
lty=c(1,2,2,3,3),col=c("black","red","red","blue","blue"),
type="l",xlab="edad",ylab="costes")
legend(1,3300,c("Estim.Media","Predicción"),lty=c(2,3),
col=c("red","blue"))
points(x,y)
```

En el ancho de las bandas de confianza de la Figura 5.5 se aprecia la diferente certidumbre que se obtiene sobre la estimación de la respuesta media y sobre la predicción de una nueva observación. Puesto que la relación entre costes y log(costes) es biunívoca, es posible deshacer la transformación para obtener estimaciones e intervalos de confianza de la estimación de los costes en función de la edad de los camiones. La recta ajustada entre log(costes) y edad da lugar a una curva de estimación de costes en función de edad.

5.8. Bondad del Ajuste

Cuando hemos realizado el ajuste de un modelo de regresión lineal, hemos de verificar que efectivamente dicho modelo proporciona un buen ajuste a la hora de explicar (predecir) la variable respuesta. Básicamente la bondad del ajuste la cuantificamos con el tanto por ciento de variabilidad explicada por el modelo

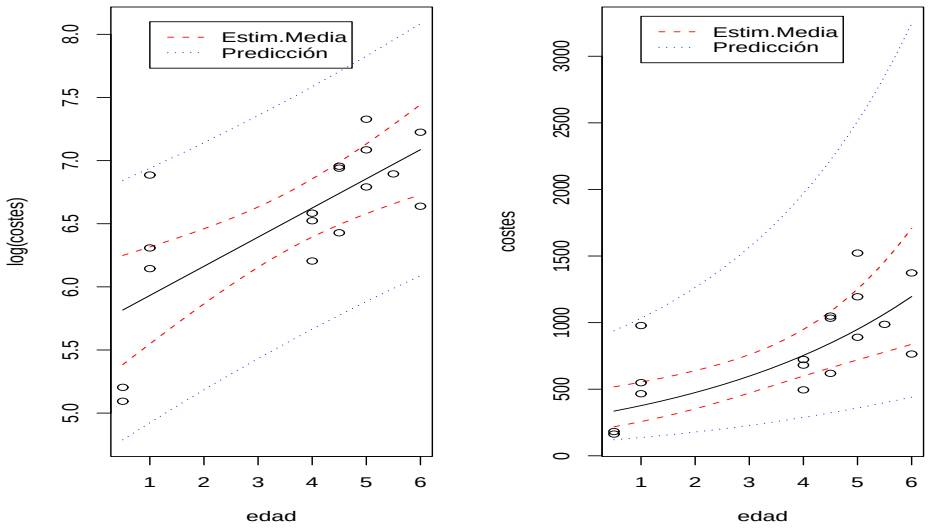


Figura 5.5: Estimación de la respuesta media y predicción de una futura observación para el ajuste de los Tractores. La recta de regresión relaciona log(costes) con edad.

sobre la variable respuesta. Para ello contamos con varios tipos de medidas que cuantifican esta variabilidad de diversos modos. Como medidas fundamentales de bondad de ajuste contamos con:

- el error residual estimado $s = \hat{\sigma}$;
- el test F de bondad de ajuste que se obtiene de la Tabla de Anova;
- el coeficiente de determinación R^2.

Con todo, la prueba más reconocida para concluir sobre la bondad del ajuste, por la obtención de significatividad estadística, es la correspondiente al test F derivado de la Tabla de Anova. Superado este test, el modelo se da por bueno. La siguiente prueba a superar será la del diagnóstico y validación del modelo, o verificación de las hipótesis del modelo RLS y de la capacidad predictiva del mismo. De entre todos los modelos propuestos para predecir una respuesta y que hayan superado la bondad del ajuste, el diagnóstico y la validación, podremos optar por el mejor según algún criterio preferido de comparación y selección de modelos. Pero estos criterios ya los presentaremos en el Tema 6.

5.8.1. Error residual estimado

Es una medida de bondad del ajuste relativa a la escala de medida utilizada. En general, se prefieren modelos con menor error residual estimado s, donde s^2 denotaba la estimación de la varianza σ^2 del modelo, dada en (5.13).

5.8.2. Descomposición de la varianza: Anova

Una medida de lo bueno que resulta un modelo para ajustar unos datos pasa por cuantificar cuánta de la variabilidad contenida en éstos ha conseguido ser explicada por dicho modelo. Un modelo es bueno si la variabilidad explicada es mucha, o lo que es lo mismo, si las diferencias entre los datos y las predicciones según el modelo son pequeñas.

Construir la tabla de ANOVA o Análisis de la Varianza consiste en:

- descomponer la variabilidad de los datos en la parte que es explicada por el modelo y la parte que se deja sin explicar, es decir, la variabilidad de los residuos, y
- compararlas y valorar estadísticamente si la variabilidad explicada por el modelo ajustado es suficientemente grande.

Si partimos de la identidad:

$$y_i - \bar{y} = (y_i - \hat{y}_i) + (\hat{y}_i - \bar{y})$$

y el hecho de que $\sum_i (y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) = 0$, tenemos que:

$$\sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$

es decir,

$$\underbrace{\text{Variabilidad datos}}_{SST} = \underbrace{\text{Variabilidad residuos}}_{SSE} + \underbrace{\text{Variabilidad recta}}_{SSR}$$

Las abreviaturas SST, SSE y SSR provienen del inglés para suma de cuadrados (*Sum of Squares*): Total, debida al Error (o residual) y debida a la Regresión, respectivamente. A partir de ellas se calculan las sumas de cuadrados medias dividiendo por los correspondientes grados de libertad asociados, $MST = SST/(n-1)$, $MSE = SSE/(n-2)$ y $MSR = SSR/1$.

Contrastar la bondad del ajuste de la recta de regresión significa resolver el contraste:

$$H_0 \text{ :el modelo lineal NO explica bien la respuesta}$$
$$H_1 \text{ :el modelo lineal explica bien la respuesta,}$$

que resulta equivalente a contrastar $H_0 : \beta_1 = 0,\ vs.\ H_1 : \beta_1 \neq 0$, esto es, contrastar si la variable predictora x explica suficientemente bien la variable respuesta `y` a través del modelo lineal propuesto. El estadístico de bondad de ajuste de la regresión está basado en comparar la variabilidad explicada por el modelo con la que queda sin explicar, esto es, en el cociente de las sumas de cuadrados medias MSE y MSR, que resulta tener una distribución F con 1 y $n-2$ grados de libertad cuando el modelo es correcto (Fisher, 1922):

$$F = \frac{SSR/\sigma^2}{\frac{SSE/\sigma^2}{n-2}} = \frac{MSR}{MSE} \sim F_{1,n-2}. \tag{5.20}$$

En el modelo RLS, el estadístico F es igual al estadístico t asociado a β_1, elevado al cuadrado. Ya hemos dicho antes que el contraste de bondad de ajuste es equivalente al de $\beta_1 = 0$.

Concluiremos que la recta de regresión es significativa al nivel $(1-\alpha)100\,\%$ para predecir la respuesta `y`, cuando el valor que obtenemos para el estadístico F supera el valor crítico que se corresponde con el cuantil $1-\alpha$ de una distribución F con 1 y $n-2$ grados de libertad. O lo que es lo mismo, cuando el p-valor asociado a él resulta inferior a α. En otro caso, diremos que no hemos obtenido evidencias suficientes para rechazar que el modelo lineal no es útil para predecir la variable `y` a través de x.

Todas estas sumas de cuadrados y estadísticos se suelen presentar en una tabla de Anova, cuya apariencia suele tener la forma de la Tabla 5.1.

Ejemplo 5.7 (Tabla de Anova en el ajuste de Tractores)**.** *Obtener la Tabla de Anova para el ajuste obtenido con los datos de Tractores en el Ejemplo 3.1. Concluir sobre la bondad del ajuste.*

Fuente	gl	SS	MS	estadístico F	p-valor
Regresión	1	SSR	$MSR = \frac{SSR}{1}$	$F = \frac{MSR}{MSE}$	$Pr(F_{1,n-2} > F)$
Error	$n-2$	SSE	$MSE = \frac{SSE}{n-2}$		
Total	$n-1$	S_{yy}			

Tabla 5.1: Tabla de Análisis de la Varianza en Regresión Lineal Simple.

```
x<-edad;y<-costes
fit<-lm(log(y)~x)
# La tabla de Anova se consigue con:
anova(fit)
# que proporciona estadístico F y p-valor:
#
#Analysis of Variance Table
#Response: log(y)
#          Df Sum Sq Mean Sq F value    Pr(>F)
#x          1 3.3806  3.3806  17.663 0.0007687 ***
#Residuals 15 2.8709  0.1914

# El estadístico F de la Tabla Anova es igual al estadístico t
# de beta1, elevado al cuadrado, y por tanto sus p-valores coinciden:
sfit$coefficients
coef(sfit)["x",]        # t=4.2027

coef(sfit)["x",3]^2     #t^2=17.66275
sfit$fstatistic         # F=17.66275
```

Simplemente observando la Tabla de Anova, vemos que la variabilidad explicada por la recta (en términos de sumas de cuadrados), SSR=3.3806, es superior a la que queda por explicar, SSE=2.8709. El estadístico F valora si dicha diferencia es suficientemente grande como para poder concluir que efectivamente la recta explica la mayor parte de la variabilidad existente. El valor del estadístico F es 17.66275, que para una F con 1 y 15 grados de libertad da un p-valor de 0.0007687. La conclusión es que podemos rechazar $H_0 : \beta_1 = 0$, o lo que es lo mismo, H_0:el modelo no explica los datos, a favor de que la edad resulta útil para predecir el logaritmo de los costes a través de un modelo de regresión lineal.

5.8.3. El coeficiente de determinación

Otro estadístico útil para chequear la bondad del ajuste de la recta de regresión es el *coeficiente de determinación* R^2. Éste se define como la proporción de la varianza que es explicada por la recta de regresión:

$$R^2 = \frac{SSR}{SST} = \frac{S_{xy}^2}{S_{xx}S_{yy}}. \tag{5.21}$$

De hecho, en RLS, R es la correlación lineal (4.1) entre la respuesta `y` y el predictor x, $R^2 = r_{xy}^2$.

Puesto que $0 \leq R^2 \leq 1$, un valor cercano a 1 (entre 0.6 y 1) implicará que buena parte de la varianza es explicada por la recta de regresión, y $R^2 \approx 0$ significará que prácticamente toda la variabilidad de los datos queda sin explicar por la recta. Sin embargo, R^2 no sirve para medir la idoneidad del modelo de regresión para describir los datos. De hecho, R^2 puede resultar grande a pesar de que la relación entre x e `y` no sea lineal (de hecho tiene la misma interpretación que un coeficiente de correlación, válido para cuantificar la relación lineal sólo cuando ésta existe). Siempre ha de ser utilizado con cautela. Así por ejemplo, la magnitud de R^2 depende del rango de variabilidad de la variable explicativa. Cuando el modelo de regresión es adecuado, la magnitud de R^2 aumenta, o disminuye, con la dispersión de x. Por otro lado, podemos obtener un valor muy pequeño de R^2 debido a que el rango de variación de x es demasiado pequeño, y entonces impedirá que se detecte su relación con `y`.

Ejemplo 5.8 (Coeficiente de determinación para el ajuste de Tractores)**.** *Obtener el coeficiente de determinación del ajuste conseguido para Tractores (Ejemplo 3.1). Comprobar que dicho coeficiente coincide con el coeficiente de correlación al cuadrado. Concluir sobre la bondad del ajuste en base a él.*

```
x<-edad;y<-costes
fit<-lm(log(y)~x)

# El coeficiente de determinación se obtiene a partir de:
sfit<-summary(fit);sfit
# identificado como Multiple R-Squared:
sfit$r.squared  #= 0.5407612

# Corroboramos que su valor es igual a la correlación al cuadrado:
cor(x,log(y))^2 #= 0.5407612
```

El valor que obtenemos para el coeficiente de determinación (Multiple R-Squared) es de $R^2 = 0{,}541$*, esto es, alrededor del* $54\,\%$ *de la variabilidad de los log-costes es*

explicada por la recta ajustada. No es un valor especialmente alto. De hecho, ya apreciábamos en el gráfico de los datos, Figura 5.5, que la tendencia lineal más marcada provenía de los vehículos que superaban los cuatro años. Los costes de los camiones muy nuevos han sido muy variables y es, cuanto menos arriesgado, hablar de una relación lineal con la edad del tractor.

5.9. Diagnóstico Gráfico del Modelo. Análisis de los Residuos

Una vez ajustado un modelo y habiendo superado las pruebas de bondad de ajuste pertinentes, fundamentalmente el test F de Anova, el diagnóstico del modelo consiste en verificar si satisface las hipótesis básicas del modelo de regresión, que son:

- linealidad entre las variables x e y;
- para los errores del modelo, ϵ_i:
    1. media cero
    2. varianza constante
    3. incorrelación
    4. normalidad.

El análisis de los residuos nos permitirá detectar deficiencias en la verificación de estas hipótesis, así como descubrir observaciones anómalas o especialmente influyentes en el ajuste. Una vez encontradas las deficiencias, si existen, cabrá considerar el replanteamiento del modelo, bien empleando transformaciones de las variables, bien proponiendo modelos alternativos al de RLS, que trataremos con detalle en el Tema 6.

El diagnóstico del modelo se lleva a cabo fundamentalmente a partir de la inspección de los residuos del modelo. Éstos sólo son buenos estimadores de los errores cuando el modelo ajustado es bueno. Aun así, es lo más aproximado con lo que contamos para indagar qué ocurre con los errores y si éstos satisfacen las hipótesis del modelo. El análisis de los residuos habitual es básicamente gráfico, si bien existen varios tests estadísticos útiles para detectar inadecuaciones del modelo, que no trataremos aquí (ver Draper y Smith, 1998).

Definimos los residuos de un modelo lineal como las desviaciones entre las observaciones y los valores ajustados:

$$e_i = y_i - \hat{y}_i, \quad i = 1, \ldots, n. \tag{5.22}$$

En ocasiones, es preferible trabajar con los residuos estandarizados, que tienen media cero y varianza aproximadamente unidad, y facilitan la visualización de las hipótesis:

$$d_i = \frac{e_i}{\sqrt{MSE}}, \quad i = 1, \ldots, n. \tag{5.23}$$

En cualquier caso, si detectamos violaciones en algunas de las hipótesis del modelo, cabe proceder con la revisión del modelo, formulación o inclusión de nuevas variables, replanteamiento del modelo o incluso de uno alternativo fuera de la RLS. Algunas de las posibilidades de corrección las repasaremos en el Tema 6.

5.9.1. Gráfico qq-plot e histograma de los residuos

Para verificar la normalidad de los errores disponemos de gráficos qq-plot de normalidad, en los que se representan los residuos ordenados $e_{[i]}$ (cuantiles empíricos) versus los cuantiles correspondientes de una normal estándar, $\Phi^{-1}[(i-1)/n]$. Si es cierta la normalidad de los residuos, los puntos han de estar alineados con la diagonal, como en el Gráfico (a) de la Figura 5.6. Desviaciones de la diagonal más o menos severas en las colas (Gráficos (c) y (d)) e incluso en el centro de la distribución (Gráfico (b)) delatan desviaciones de normalidad. La hipótesis de normalidad se puede chequear también con histogramas de los residuos cuando el tamaño muestral es grande.

Los residuos estandarizados también son útiles para detectar desviaciones de la normalidad. Si los errores se distribuyen según una normal, entonces aproximadamente el 68 % de los residuos estandarizados quedarán entre -1 y $+1$, y el 95 % entre -2 y $+2$.

5.9.2. Gráfico de residuos versus valores ajustados $\hat{y}_i$

Permiten detectar varios tipos de deficiencias del modelo ajustado. Si los residuos están distribuidos alrededor del cero y el gráfico no presenta ninguna

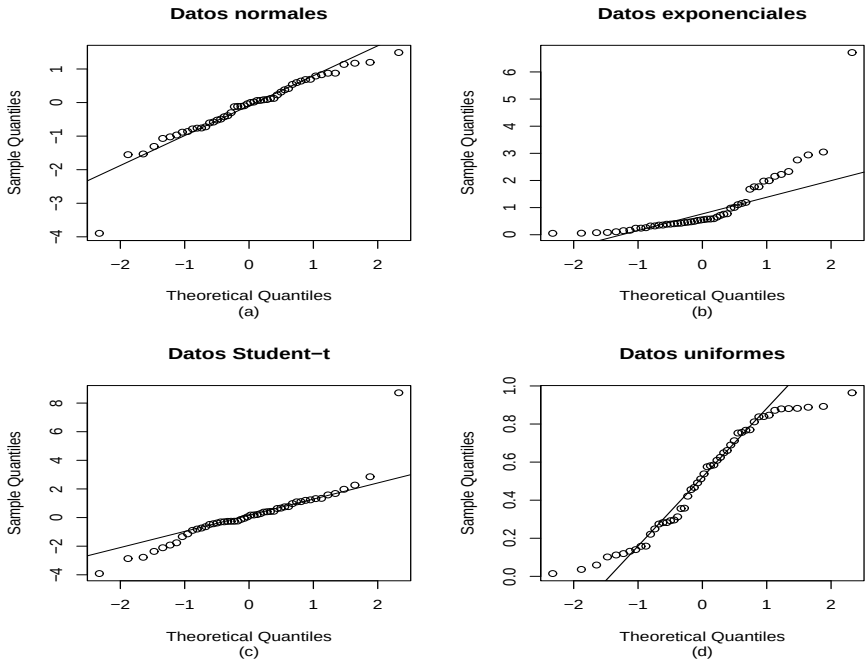


Figura 5.6: Gráficos qqplot para verificar normalidad.

tendencia (Figura 5.7 (a)), entonces el modelo se considera adecuado. Cuando aparece alguna tendencia como una forma de embudo (Figura 5.7 (b)) o un abombamiento (Figura 5.7 (c)), etc., podemos tener algún problema con la hipótesis de varianza constante para los errores (heterocedasticidad). En el caso b) la varianza aumenta con la magnitud de la respuesta y en el caso c) las observaciones con magnitud media tienen mayor dispersión que las extremas. Cuando se aprecia alguna tendencia, hablamos de violación de la hipótesis de linealidad: el modelo lineal ha sido incapaz de capturar una tendencia no lineal apreciada en los residuos, posiblemente debido a que existen otras variables explicativas adicionales no consideradas en el modelo, o a que la variable predictora explica la respuesta de un modo más complejo (quizás polinómico, etc.) al considerado en el modelo lineal (Figura 5.7 (d)).

5.9.3. Gráfico de residuos versus valores de la variable predictora x_i

Son útiles para apreciar tendencias en los residuos que han quedado sin explicar por el modelo ajustado. Básicamente se interpretan como los gráficos de

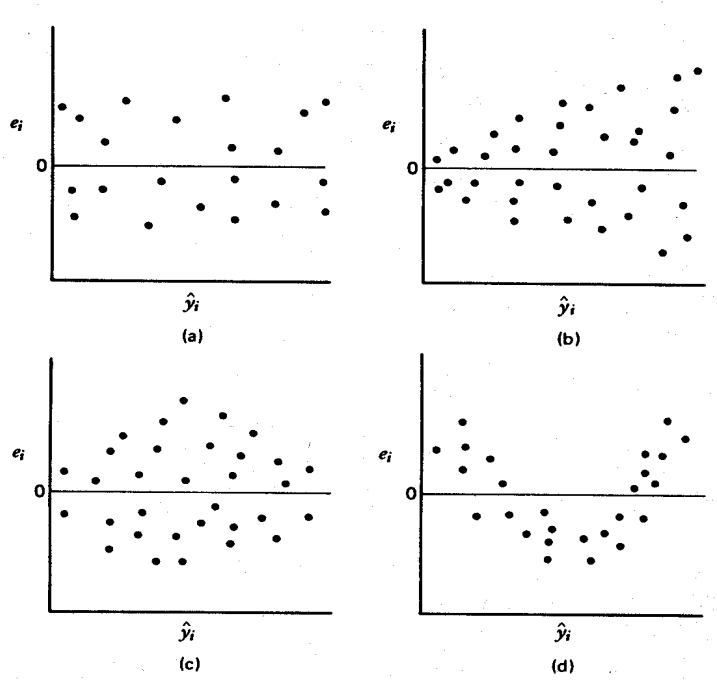


Figura 5.7: Gráficos de residuos versus valores ajustados: (a) Adecuación del modelo. (b) y (c) Heterocedasticidad. (d) Falta de linealidad.

residuos versus valores ajustados $\hat{y}_i$. Es deseable que los residuos aparezcan representados en una banda horizontal sin tendencias alrededor del cero. Por ejemplo, si hay tendencias de tipo cuadrático, posiblemente hayamos de incorporar la variable x^2 en el modelo, o bien abordar algún tipo de transformación que linealice la relación entre predictor y respuesta.

5.9.4. Gráfico de residuos versus otros posibles regresores

Representar los residuos versus otras variables observadas que puedan hacer el papel de predictores puede revelar la necesidad de incluirlos para conseguir explicar algo más de la respuesta. Los gráficos que manifiesten algún tipo de tendencias identificarán a otros regresores potenciales que ayuden a la mejora del modelo.

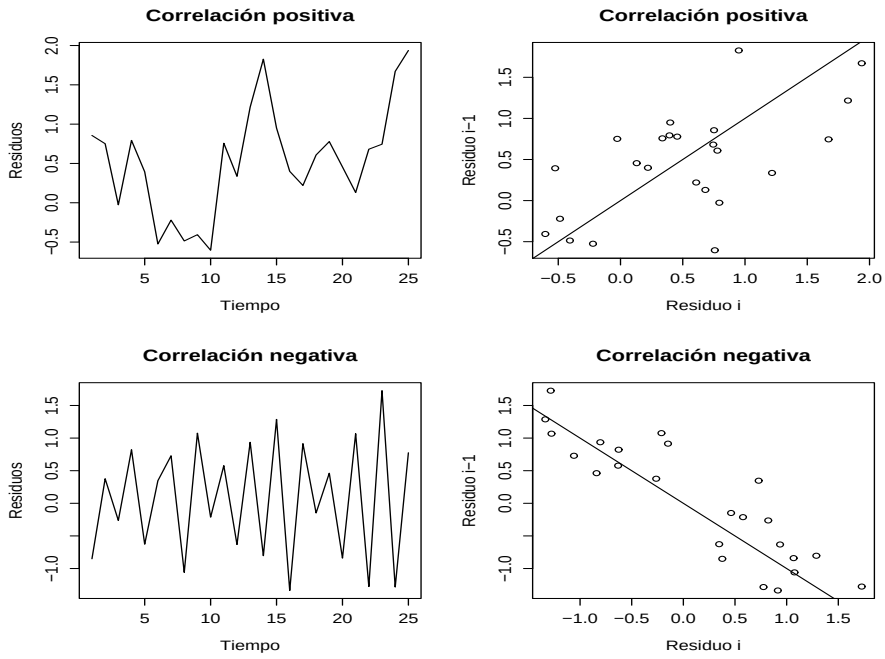


Figura 5.8: Autocorrelación de los residuos.

5.9.5. Gráfico secuencial de los residuos

La correlación entre los datos es un proceso intrínseco al muestreo; saber cómo se ha llevado a cabo éste da información, generalmente suficiente, para poder hablar de correlación o incorrelación. En todo caso, los gráficos secuenciales de residuos sirven para detectar problemas de correlación de éstos (*autocorrelación*), o de inestabilidad de la varianza a lo largo del tiempo. También son útiles para esto los gráficos en que se representa un residuo versus el anterior en la secuencia en que han sido observados; si hay correlación se apreciará tendencia. Detectar autocorrelación llevará a considerar otro tipo de modelos distintos (autocorrelados: modelos de series temporales). En la Figura 5.8 mostramos diversas situaciones de correlación de los residuos, detectadas a través de los dos tipos de gráficos propuestos: secuenciales (G1) y un residuo versus el anterior (G2). La correlación positiva se aprecia gracias a que residuos grandes suelen ser seguidos de residuos grandes, y los mismo ocurre con los pequeños; en el gráfico G2 se aprecia una tendencia lineal con pendiente positiva. La correlación negativa implica que residuos grandes van seguidos de residuos pequeños en G1, o bien que la tendencia en G2 es lineal con pendiente negativa.

Ejemplo 5.9 (Diagnóstico del modelo RLS en el ajuste de Tractores)**.** *Calcular*

los residuos asociados al ajuste de los Tractores y llevar a cabo el diagnóstico del modelo. Concluir sobre el mismo.

```
x<-edad;y<-costes
fit<-lm(log(y)~x)


# Los residuos e(i) se obtienen con:
fit$residuals
# o bien
e<-residuals(fit); e

# Los residuos estandarizados se obtienen con:
sfit<-summary(fit)
d<-e/sfit$sigma; d
# o bien con una función de la librería MASS
library(MASS)
stdres(fit)


# R proporciona un diagnóstico gráfico del modelo ajustado con
par(mfrow=c(2,2))
plot(fit)
```

En la Figura 5.9 tenemos el diagnóstico gráfico que proporciona $\mathcal{R}$ por defecto. En el gráfico de residuos versus valores ajustados (Residuals vs Fitted) apreciamos mayor dispersión en los datos con menor coste previsto y cierta tendencia en los datos con mayores costes predichos, si bien dicha tendencia prácticamente desaparece al utilizar los residuos estandarizados, por lo que el comportamiento ahí no nos preocupa. La diferente variabilidad entre observaciones con costes predichos mayores y menores sí que es preocupante, y cabría plantear alguna corrección de la misma.

La normalidad de los residuos (gráfico Normal-QQ) no es especialmente respetuosa con la diagonal, y en concreto la observación 15 se escapa considerablemente de la misma.

El último gráfico de Residuals vs Leverage se utiliza, como veremos en el Tema 6 para identificar observaciones alejadas o influyentes en el ajuste (posibles outliers u observaciones deficientes). De momento no los comentamos.

Procedemos a continuación a dibujar los gráficos de diagnóstico descritos y no proporcionados por defecto por $\mathcal{R}$.

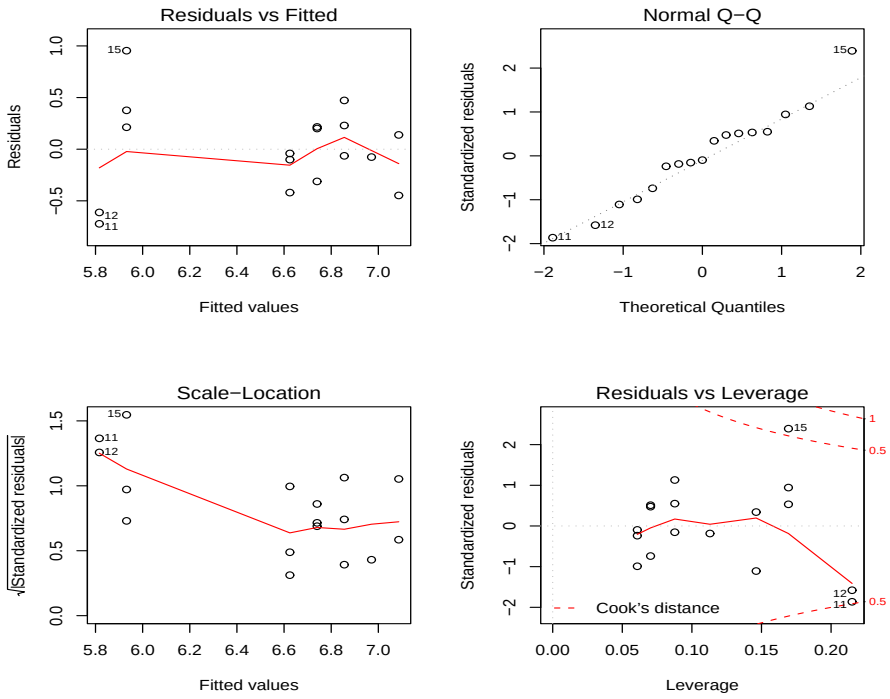


Figura 5.9: Diagnóstico Gráfico del modelo RLS ajustado para Tractores.

```
par(mfrow=c(2,2))
# el histograma de los residuos, superponiendo una densidad normal:
hist(d,probability=T,xlab="Residuos estandarizados",main="",
xlim=c(-3,3))
d.seq<-seq(-3,3,length=50)
lines(d.seq,dnorm(d.seq,mean(d),sd(d)))

# Conseguir los gráficos qqplot de normalidad con los comandos:
# qqnorm(d)
# qqline(d)

# El gráfico de residuos versus el predictor:
plot(x,d,xlab="Edad Tractor",ylab="Residuos Estandarizados")
# y capturamos con una curva suavizada la tendencia:
lines(lowess(x,d),col="red")

# La secuencia temporal de residuos se obtiene con:
plot(d,type="b",ylab="Residuos estandarizados")

# y el gráfico de residuos versus el anterior será:
n<-length(d)
plot(d[1:(n-1)],d[2:n],xlab="Residuo i",ylab="Residuo i-1")
# y capturamos con una curva suavizada la tendencia:
lines(lowess(d[1:(n-1)],d[2:n]),col="red")
```

En la Figura 5.10 tenemos el resultado de los gráficos ejecutados. Como en el gráfico qq-plot, el histograma *de los residuos muestra cierta desviación de normalidad, especialmente crítica en la cola derecha. La diferente variabilidad de los residuos es notoria en el gráfico de* Residuos versus Edad*, así como cierta tendencia cuadrática en los residuos explicada por la Edad en los datos correspondientes a los tractores mayores de 4 años. El gráfico secuencial de residuos no muestra ninguna estructura patente de correlación, y tampoco el de residuos consecutivos.*

La conclusión es que detectamos faltas de normalidad, homocedasticidad y de linealidad, especialmente debida a la distinción entre dos grupos de tractores: los de edad inferior a un año y los de edad superior a 4. Para resolver esto, podríamos crear una nueva variable que identifique estos dos grupos y emplearla en el ajuste. Esto es, proponer un nuevo modelo de predicción que permita ajustar una recta de regresión para los camiones modernos y otra para los más viejos. Este tipo de modelos más complejos (Ancova) los tratamos en el Tema 6.

Cuando, una vez realizado el diagnóstico del modelo, detectamos deficiencias, cabe reconsiderar la modelización, posiblemente con transformaciones de los datos que corrijan los defectos encontrados. Propuesto un nuevo modelo corrector,

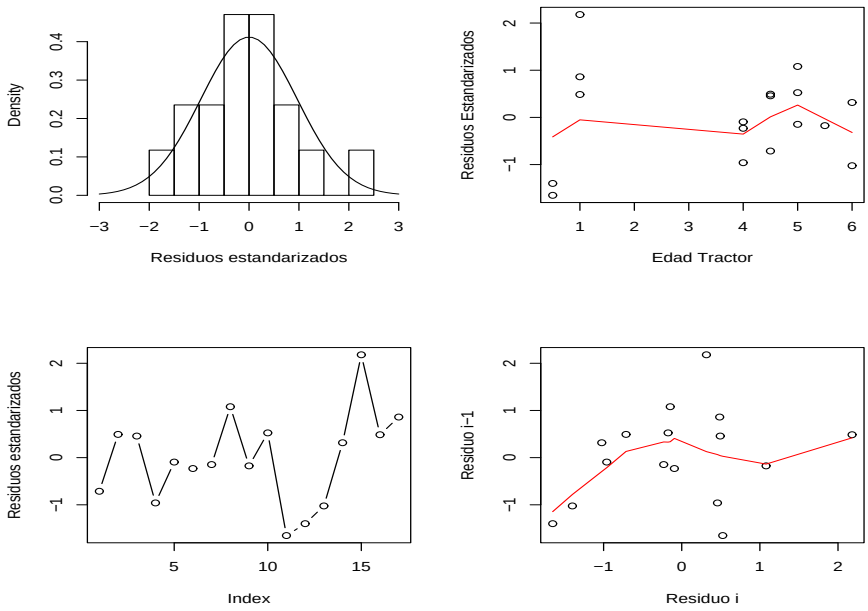


Figura 5.10: Diagnóstico Gráfico del modelo RLS ajustado para Tractores.

los pasos a dar vuelven a ser los mismos: ajuste, bondad del ajuste y diagnosis. En el tema siguiente hablaremos más de las soluciones a problemas en el diagnóstico, y de otro tema importante cuando un modelo ajustado cumple con todas las hipótesis del modelo RLS, como es el de validación.

5.10. Ejercicios

Ajuste de una recta de regresión.

1. Considera los datos `cars` en $\mathcal{R}$ (*data(cars)*), que provienen de datos recopilados en la década de 1920, sobre velocidad de los coches y distancias de frenada. La velocidad (*speed*) viene dada en millas por hora, y la distancia de frenado (*dist*), en piés.

   a) Investiga la relación lineal entre las dos variables y justifica la idoneidad de un modelo de regresión lineal simple para describir dicha relación. ¿Se puede hablar de una relación de causalidad entre las dos variables? Justifica la respuesta.

   b) Identifica justificadamente y según el contexto de los datos, la variable respuesta, y, y la explicativa x.

   c) Ajusta el modelo de regresión lineal correspondiente, al que nos referiremos en adelante por *ajuste1*. Interpreta todas las salidas.

   d) Construye intervalos de confianza para los coeficientes de la recta y resuelve el contraste de regresión con (5.15). Comprueba la equivalencia entre el contraste $\beta_1 = 0$, el F de bondad del ajuste y el de correlación $\rho = 0$ a partir de los estadísticos t (5.15), F (5.20) y el t de correlación de Pearson (4.3).

   e) Construye y representa gráficamente la estimación y las bandas de estimación de la respuesta media $E(\mathsf{y})$ para el rango de valores de x comprendido entre el mínimo y el máximo de $\{x_1, \ldots, x_n\}$.

   f) Construye y representa gráficamente la predicción y las bandas de predicción de una nueva observación de y para el rango de valores de x comprendido entre el mínimo y el máximo de $\{x_1, \ldots, x_n\}$. Compáralo con el resultado de la estimación de la respuesta media. Comenta y justifica las diferencias.

   g) Consigue la tabla de Anova (5.1) y comenta la significatividad del ajuste. Expresa las conclusiones en un lenguaje no técnico.

Calibración del modelo.

En la calibración de un modelo lineal (para ampliar, ver Osborne, 1991), el interés es, a partir de una recta de regresión de y sobre x, determinar valores de x a partir de respuestas observadas de y. Esto es, determinar los valores de la variable explicativa que predicen/estiman una determinada respuesta.

h) Escoge al azar un valor de `y` en su rango de variación y llámalo y^*. Da una estimación del valor de x, x^*, que predeciría como respuesta media $E(y|x^*) = y^*$. Construye un intervalo de confianza para dicho x^*, utilizando la banda de estimación obtenida en el apartado $1e$).

i) Construye ahora la recta de regresión de x sobre `y` y llámala *ajuste2*. Para el valor elegido en el apartado anterior, $\mathtt{y} = y^*$, estima $E(\mathrm{x}|\mathtt{y} = y^*)$ y una banda de confianza y compara los resultados con los que obtuviste antes. Comenta las diferencias entre *ajuste1* y *ajuste2*.

Diagnóstico gráfico del modelo.

2. Representa gráficamente el ajuste obtenido y comenta la calidad del mismo. ¿Qué conclusiones se pueden extraer de los gráficos?

   a) Valores $y_1, \ldots, y_n$ versus $x_1, \ldots, x_n$, y la recta ajustada superpuesta.

   b) Los residuos $\{e_i = y_i - \hat{y}_i\}_{i=1}^n$ secuencialmente.

   c) Residuos versus valores predichos $\hat{y}_1, \ldots, \hat{y}_n$.

   d) Residuos versus valores $x_1, \ldots, x_n$.

Influencia de observaciones extremas y raras.

3. Investiga la influencia de las observaciones en los extremos del rango de estimación. Para ello, excluye el dato correspondiente al valor de x más extremo y reajusta el modelo de regresión (llámalo *ajuste3*); compáralo con *ajuste1* y comenta cómo varía la recta.

4. Identifica gráficamente los valores más influyentes en el *ajuste1*, exclúyelos y reajusta el modelo de regresión (llámalo *ajuste4*). Compáralo con *ajuste1* y comenta cómo varía la recta.

Otros ajustes.

1. Considera los datos `bestbuy.dat`, que provienen de información recogida por la empresa americana Best Buy Co., Inc. sobre potencia de computación necesitada por sus ordenadores (medida en MIPS - Millions of Instructions Per Second) y número de almacenes repartidos por el país. Ajusta la recta de regresión. Realiza el diagnóstico de residuos. Presta especial atención al gráfico de residuos frente a la variable explicativa. Prueba con alguna transformación de la variable respuesta y chequea su idoneidad.

2. Fichero `diamond.dat`. El archivo contiene los precios (en dolares de Singapur)de una serie de anillos de señora y el peso del diamante que llevan incrustado. El peso de los diamantes se mide en quilates. Un quilate equivale a 0.2 gramos. Los anillos se hacen con oro de 20 quilates de pureza y son montados, cada uno de ellos, con una única piedra de diamante. Ajusta el modelo de regresión. Profundiza sobre el hecho de que el intercepto estimado tome valores negativos. Plantea algún tipo de solución al problema.

3. Archivo `PCB.dat`. Descripción de los datos: concentración de PCBs en diferentes bahías para los años 1984 y 1985. Las variables del fichero son: BAY (el nombre de la bahía), PCB84 (la concentración en partes por cada mil millones) y PCB85 (la concentración en partes por cada mil millones). Se desea relacionar PCB84 con PCB85 a través de una recta de regresión. Evalúa la bondad del ajuste.

Recursos en la red.

Visitando la web `http://bcs.whfreeman.com/ips4e/pages/bcs-main.asp?s=00010&n=99000&i=99010.01&v=categor` tenemos la posibilidad de mejorar nuestra intuición sobre conceptos estadísticos como correlación o regresión linea, a través del uso de unos applets de java. En particular, nos interesa consultar los ejemplos “Two Variable Calculator (Correlation and Regression)” y “Correlation and Regression Demo”.

Tema 6

El modelo lineal general

6.1. Introducción, conceptos y particularizaciones

El modelo lineal general es un modelo que relaciona de modo lineal una variable respuesta continua y con los coeficientes del modelo β, que cuantifican el peso de las variables regresoras consideradas a la hora de explicar la respuesta. Si vamos a observar una muestra aleatoria de la variable y, y los valores correspondientes de ciertas $p-1$ variables explicativas controladas que suponemos relacionadas con la primera, $\mathrm{x}_1, \ldots, \mathrm{x}_{p-1}$, el modelo lineal general se formula en términos matriciales según:

$$\mathbf{y}|X = X\beta + \epsilon, \tag{6.1}$$

donde

> $\mathbf{y}$ es el vector de observaciones, $n \times 1$
> X es la *matriz de diseño*, de dimensión $n \times p$, contiene una primera columna de unos, correspondientes a un efecto global común a todos los datos, y en las restantes $p-1$ columnas, las observaciones de las variables explicativas $\mathrm{x}_1, \ldots, \mathrm{x}_{p-1}$
> β es el vector de parámetros, de dimensión $p \times 1$
> ϵ es el vector de errores aleatorios, de dimensión $n \times 1$,

y responde a las hipótesis de normalidad, incorrelación, igualdad de varianza y media cero para los errores, esto es,

$$\epsilon \sim N(\mathbf{0}, \sigma^2 I).$$

Con esto, la distribución condicionada de la respuesta individual y_i dadas las observaciones de las variables explicativas observadas, $\mathrm{x}_1 = x_{i1}, \ldots, \mathrm{x}_{p-1} = x_{i,p-1}$, es:

$$\mathsf{y}_i|\mathbf{x}_i \overset{iid}{\sim} N(\mathbf{x}_i'\beta, \sigma^2),\ i = 1, \ldots, n$$

con $\mathbf{x}_i' = (1, x_{i1}, \ldots, x_{i,p-1})$, esto es, la i-ésima fila de la matriz de diseño X.

En el modelo lineal general el objetivo básico es describir la relación entre una variable respuesta continua y ciertas variables explicativas, a través del modelo (6.1). Dichas variables explicativas pueden ser tanto continuas como de tipo categórico. Como ya comentamos, a las primeras nos referiremos habitualmente como *covariables*, mientras que a las segundas aludiremos como *factores*.

Nuestro propósito en este tema es presentar y estudiar de un modo unificado diversos modelos lineales sujetos a las hipótesis del modelo lineal general, que difieren por la naturaleza de las variables explicativas y de los objetivos del análisis. Los que trataremos son:

1. Modelos lineales con todas las variables explicativas de tipo continuo, generalmente con fines predictivos:
   - Regresión lineal simple.
   - Regresión polinómica.
   - Regresión lineal múltiple.
2. Modelos lineales con variables explicativas categóricas, o con categóricas y continuas conjuntamente, que añaden objetivos de comparación de grupos:
   - Modelos de Anova.
   - Modelos de Ancova.

Veamos a continuación cómo particularizar la formulación del modelo lineal general (6.1) a cada uno de estos modelos. Para ello, recordemos primero los ejemplos que ya presentamos en el Tema 3, y formulemos el modelo lineal general sobre cada uno de ellos.

6.1.1. Regresión lineal simple

Para escribir el modelo de regresión lineal simple (5.1) en términos matriciales, basta considerar la matriz de diseño $X = (\mathbf{1}, \mathbf{x})$, cuya fila i-ésima es

$\mathbf{x}_i' = (1, x_i)$, $i = 1, \ldots, n$:

$$\mathsf{y} = \beta_0 + \beta_1 \mathrm{x} + \epsilon = X\beta + \epsilon, \tag{6.2}$$

El vector de parámetros del modelo es $\beta = (\beta_0, \beta_1)'$ y el vector de errores aleatorios, $\epsilon' = (\epsilon_1, \ldots, \epsilon_n)$.

6.1.2. Regresión lineal múltiple

Hablamos de un modelo de regresión lineal múltiple cuando tenemos $p-1$ variables explicativas de tipo continuo, $\mathrm{x}_1, \ldots, \mathrm{x}_{p-1}$, relacionadas con una variable respuesta y a través del modelo:

$$\mathsf{y} = \beta_0 + \beta_1 \mathrm{x}_1 + \beta_2 \mathrm{x}_2 + \ldots + \beta_{p-1} \mathrm{x}_{p-1} + \epsilon, \tag{6.3}$$

donde ϵ representa un error aleatorio normal, centrado en cero y con varianza constante.

Para unos datos dados, obtener la expresión matricial (6.1) pasa de nuevo por definir una matriz de diseño X cuya primera columna es un vector de unos y las siguientes contienen las observaciones de las variables explicativas $\mathrm{x}_1, \ldots, \mathrm{x}_{p-1}$, esto es,

$$X = (\mathbf{1}, \mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{p-1}),$$

con la fila i-ésima $\mathbf{x}_i' = (1, x_{i1}, \ldots, x_{in})$.

El vector de coeficientes es $\beta' = (\beta_0, \beta_1, \ldots, \beta_{p-1}) \in \mathbb{R}^p$, tal que cada coeficiente β_i cuantifica el peso que tiene la variable x_i explicando la respuesta media de y. En concreto, para un valor x_i dado de cierta variable x_i, si se observara $x_i + 1$ en dicho predictor y los mismos valores en los restantes regresores (especificados), entonces la predicción de la respuesta media $E(\mathsf{y})$ se incrementaría en β_i.

Un modelo de regresión múltiple puede ser útil para describir de un modo sencillo complejas relaciones entre la variable respuesta y y ciertas variables explicativas observadas. De hecho, se trata de describir dicha relación a través de una superficie, lineal en las variables explicativas, lo más próxima posible a los valores observados de la respuesta.

Los datos del Ejemplo 3.2 son modelizables con un modelo de regresión múltiple.

La idea básica de la regresión múltiple cuando tenemos, por ejemplo, dos predictores x_1 y x_2 para explicar una respuesta y, es que tratamos de explicar de y todo lo posible con x_1,y a continuación, intentamos utilizar la información adicional que aporta x_2 y que no está contenida en x_1 para completar la predicción sobre y.

6.1.3. Regresión polinómica

El modelo de regresión lineal simple (6.2) es un modelo polinómico de orden 1 con una sola variable predictora x. El modelo de regresión lineal múltiple (6.3) es un modelo polinómico de orden 1 con $p-1$ variables predictoras $\mathrm{x}_1, \mathrm{x}_2, \ldots, \mathrm{x}_{p-1}$.

En general, los modelos polinómicos son útiles cuando se aprecia una tendencia curvilínea entre los predictores y la respuesta. Asimismo, a veces constituyen una aproximación sencilla (por serie de Taylor) a modelos complejos e incluso no-lineales. Los modelos polinómicos de orden mayor a 1 se denominan *superficies de respuesta*.

Los datos del Ejemplo 3.5 son apropiados para utilizar un modelo polinómico de orden 2.

El modelo ajustado sobre los datos anteriores es un modelo polinómico de orden 2 con una sola variable explicativa x, dependiente de $p = 3$ coeficientes y con la forma:

$$\mathsf{y} = \beta_0 + \beta_1 \mathrm{x} + \beta_2 \mathrm{x}^2 + \epsilon = X\beta + \epsilon,$$

donde la matriz de diseño es en este caso de dimensión $n \times 3$, $X = (\mathbf{1}, \mathbf{x}, \mathbf{x}^2)$, y $\beta = (\beta_0, \beta_1, \beta_2)' \in \mathbb{R}^3$.

Un modelo polinómico de orden 2 con dos variables predictoras x_1 y x_2 exige la estimación de $p = 6$ coeficientes y tiene la forma:

$$\mathsf{y} = \beta_0 + \beta_1 \mathrm{x}_1 + \beta_2 \mathrm{x}_2 + \beta_3 \mathrm{x}_1^2 + \beta_4 \mathrm{x}_2^2 + \beta_5 \mathrm{x}_1 \mathrm{x}_2 + \epsilon, \tag{6.4}$$

donde la matriz de diseño es de dimensión $n \times 6$, y sus columnas son $X = (\mathbf{1}, \mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_1^2, \mathbf{x}_2^2, \mathbf{x}_1\mathbf{x}_2)$, con $\beta \in \mathbb{R}^6$.

A la hora de ajustar un modelo polinómico, siempre serán preferibles modelos con órdenes pequeños antes que grandes (principio de parsimonia o simplicidad). La elección del orden del polinomio a ajustar puede decidirse a través de algún procedimiento secuencial de selección de modelos, que estudiaremos en la Sección 6.6.

En ocasiones sin embargo, es preferible evitar los modelos polinómicos si se puede conseguir linealidad con alguna transformación de las variables predictoras. Son transformaciones habituales el logaritmo, la inversa, la raíz cuadrada. Además, al carecer de error las variables predictoras, transformarlas no afecta a la distribución del error en el modelo lineal a ajustar.

6.1.4. Modelos de Anova

El objetivo básico de un modelo Anova es el de demostrar diferencias entre grupos de individuos. Estos grupos de individuos surgen a partir de los factores controlados que clasifican a los mismos en función de las condiciones de experimentación a las que han sido expuestos. En el ejemplo, el factor de clasificación es el grupo al que fue asignado el individuo al recibir un tratamiento o un placebo.

Consideramos la situación general en que hemos medido la respuesta $\mathbf{y}$ en función de una variable controlada categorizada o *factor*, a varios niveles fijos $i = 1, 2, \ldots, I$. En el ejemplo, tenemos dos niveles: tratamiento y placebo. Conseguimos varias mediciones de $\mathbf{y}$ para cada nivel i del factor, $y_{i1}, y_{i2}, \ldots, y_{in_i}$. El objetivo de un modelo de Anova es estimar la respuesta media θ_i en cada nivel del factor, con el fin último de comparar todas las medias y poder concluir sobre si dicha variable afecta a la respuesta provocando diferente respuesta media en cada nivel. Este modelo se escribe de la forma:

$$y_{ij} = \theta_i + \epsilon_{ij}, \quad \text{para } i = 1, 2, \ldots, I \ \text{ y } \ j = 1, 2, \ldots, n_i, \tag{6.5}$$

o lo que es equivalente,

$$y_{ij} = \mu + \alpha_i + \epsilon_{ij}, \quad \text{para } i = 1, 2, \ldots, I \ \text{ y } \ j = 1, 2, \ldots, n_i, \tag{6.6}$$

donde se asume que existe una respuesta media global μ y un efecto adicional por cada nivel que representa las diferencias respecto de la media μ, representado por α_i. ϵ_{ij} es el error aleatorio que contiene las diferencias entre el individuo j y la respuesta media del grupo i al que pertenece.

Para expresar el modelo de Anova como un modelo lineal general (6.1), hay que definir variables ficticias x_i de ceros y unos (variables dummy). Para evitar problemas de estimabilidad con la matriz de diseño X, consideramos una construcción sencilla que consiste en crear $I - 1$ variables *dummy*, más una variable toda de unos para representar el efecto global:

$\mathrm{x}_0 = 1$ para todas las respuestas,
$\mathrm{x}_1 = 1$ para todas las respuestas al nivel 2 del factor, y 0 para el resto,
$\mathrm{x}_2 = 1$ para todas las respuestas al nivel 3 del factor, y 0 para el resto,
...
$\mathrm{x}_{I-1} = 1$ para todas las respuestas al nivel I del factor, y 0 para el resto.

Así, podemos expresar el modelo (6.6) según:

$$\mathsf{y} = X\beta + \epsilon,$$

donde la matriz de diseño, salvando repeticiones en cada nivel (n_i para el nivel i), tiene la forma:

	$\mathbf{1}$	$\mathbf{x}_1$	$\mathbf{x}_2$	...	$\mathbf{x}_{I-1}$
Nivel 1 con n_1 individuos (filas)	1	0	0	...	0
Nivel 2 con n_2 individuos	1	1	0	...	0
Nivel 3 con n_3 individuos	1	0	1	...	0
.					
.					
Nivel I-1 con n_{I-1} individuos	1	0	0	...	0
Nivel I con n_I individuos	1	0	0	...	1

$$\mathbf{y} = (y_{11}, y_{12}, \ldots, y_{1n_1}, y_{21}, y_{22}, \ldots, y_{2n_2}, \ldots, y_{I1}, y_{I2}, \ldots, y_{In_I})'$$
$$\beta = (\mu, \alpha_2, \alpha_3, \ldots, \alpha_I)' \in \mathbb{R}^I.$$

Esta construcción de las variables dummy nos lleva a modelizar la respuesta media para el primer nivel del factor explicativo con:

$$E(\bar{\mathsf{y}}_1|\mathbf{x}_1) = \mu,$$

y para el resto de niveles según:

$$E(\bar{\mathsf{y}}_i|\mathbf{x}_i) = \mu + \alpha_i, \qquad i = 2, \ldots, I.$$

El número de variables dummy (grados de libertad) asociadas a un factor de clasificación con I niveles, y sin contar con el término de interceptación (o media global), es igual al número de niveles menos uno, $I - 1$. La justificación de las variables dummy proviene de ciertos problemas de estimabilidad en el modelo (6.6), que viene parametrizado en términos de $I + 1$ parámetros cuando

sólo hay I grupos de observaciones diferenciados. Surge entonces, la necesidad de añadir restricciones de identificabilidad. De hecho, nuestra formulación del modelo de Anova es equivalente a utilizar como restricción de identificabilidad $\alpha_1 = 0$. Igualmente podríamos haber considerado como categoría de referencia cualquier otra y para ella haber asumido el α correspondiente igual a cero. Otras restricciones de identificabilidad dan lugar a otras construcciones de las variables dummy.

Cuando tenemos más de un factor para predecir una respuesta y, hablamos de los modelos Anova de dos o más vías, con alguna complejidad añadida que no comentaremos aquí.

6.1.5. Modelos de Ancova

Consideremos ahora la situación en que, para explicar una variable respuesta y disponemos de variables explicativas continuas y categóricas. El objetivo es construir un modelo útil con el que expliquemos cómo varía la respuesta cuando varían las covariables continuas, y cómo afectan a estas variaciones las variables de clasificación.

Concentrémonos en un caso sencillo para ilustrar los posibles modelos de predicción: consideramos una variable explicativa continua z y un factor de clasificación F con I niveles de clasificación. En nuestro ejemplo, la velocidad es la covariable y el tipo de herramienta es el factor de clasificación.

Si ambas variables son significativas para predecir la respuesta y a través de un modelo lineal, z estará relacionada linealmente con y en cada uno de los niveles de clasificación del factor F. Es decir, para cada nivel tendrá sentido ajustar una recta de regresión. Si la covariable y el factor actúan independientemente sobre la respuesta, dichas rectas serán paralelas (en principio, sólo con interceptación distinta), esto es, la relación entre la covariable y la respuesta es la misma en todos los grupos de clasificación del factor. Si interactúan entre sí, tendremos que una misma variación de z afecta de forma diferente a la respuesta en función del nivel del factor al que pertenece, con lo que estaremos hablando de dos rectas distintas (con distinta interceptación y pendiente).

Para unos datos dados, un gráfico de dispersión de $\mathsf{y} \sim \mathrm{z}$, distinguiendo el nivel de clasificación del factor, ayudará a decidir qué modelo elegir. Si los subíndices ij identifican valores observados para el individuo j en el grupo de clasificación (o nivel) i del factor, los modelos que hemos mencionado se formulan

según:

1. Rectas paralelas:

$$y_{ij} = \mu + \alpha_i + \beta z_{ij} + \epsilon_{ij}, \quad i = 1, \ldots, I; \; j = 1, \ldots, n_i. \tag{6.7}$$

2. Rectas distintas:

$$y_{ij} = \mu + \alpha_i + (\beta + \gamma_i) z_{ij} + \epsilon_{ij}, \quad i = 1, \ldots, I; \; j = 1, \ldots, n_i. \tag{6.8}$$

Los datos del Ejemplo 3.8 son modelizables con un modelo de Ancova. En la Figura 6.1 observamos el resultado de considerar un ajuste de rectas paraleles, sin interacción entre tiempo de vida y tipo de cortadora (A/B), y otro con rectas distintas, esto es, con dicha interacción incluida en el modelo.

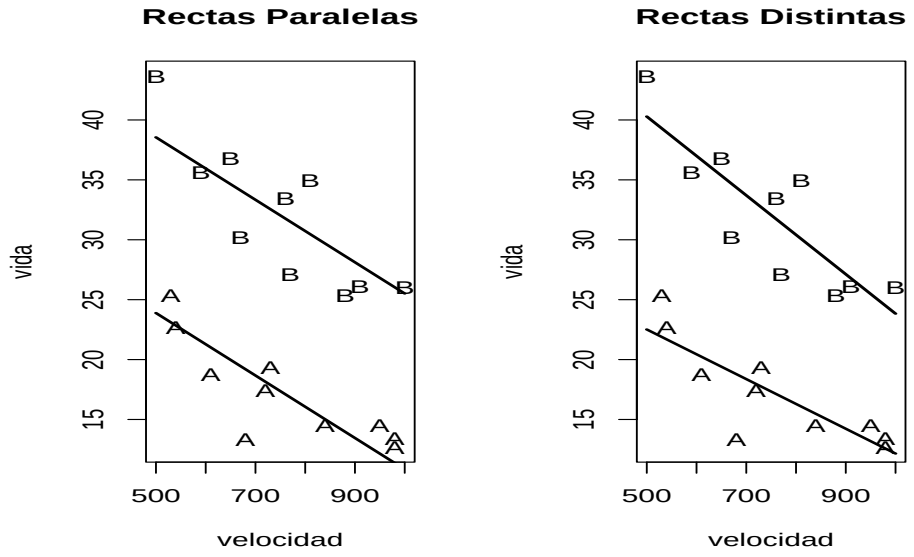


Figura 6.1: Modelo de Ancova.

De nuevo, si queremos expresar los modelos (6.7) y (6.8) como modelos lineales generales, de la forma $\mathsf{y} = X\beta + \epsilon$, hemos de construir variables dummy asociadas al factor, de forma similar a como lo hicimos en el modelo de Anova. La matriz de diseño X, salvo repeticiones (n_i en el nivel i) tendrá la forma:

1. Modelo sin interacción (rectas paralelas): la primera columna de unos, las siguientes $I - 1$ columnas definidas por las variables dummy asociadas a F

y la última columna, integrada por las observaciones de la covariable continua, z; cada una de las filas que se muestran en la tabla que sigue aparece repetida tantas veces como número de individuos n_i se hayan observado en cada nivel i.

	$\mathbf{1}$	$\mathbf{x}_1$	$\mathbf{x}_2$	...	$\mathbf{x}_{I-1}$	$\mathbf{z}$
Nivel 1 con n_1 individuos	1	0	0	...	0	z_1
Nivel 2 con n_2 individuos	1	1	0	...	0	z_2
Nivel 3 con n_3 individuos	1	0	1	...	0	z_3
.						
.						
Nivel I-1 con n_{I-1} individuos	1	0	0	...	0	z_{I-1}
Nivel I con n_I individuos	1	0	0	...	1	z_I

El vector de coeficientes para el modelo (6.7) es

$$\beta' = (\mu, \alpha_2, \alpha_3, \ldots, \alpha_I, \beta) \in \mathbb{R}^I.$$

2. Modelo con interacción (rectas distintas): las primeras $I+1$ columnas de X vienen como en el modelo (6.7), y las siguientes columnas, representando la interacción z : F, se construyen con variables dummy que surgen al multiplicar la covariable z por las dummy asociadas a F, $\mathrm{x}_1, \mathrm{x}_2, \ldots, \mathrm{x}_{I-1}$, esto es, $\mathrm{z} \cdot \mathrm{x}_1, \mathrm{z} \cdot \mathrm{x}_2, \ldots, \mathrm{z} \cdot \mathrm{x}_{I-1}$:

	$\mathbf{1}$	$\mathbf{x}_1$	$\mathbf{x}_2$	...	$\mathbf{x}_{I-1}$	$\mathbf{z}$	$\mathbf{x}_{I+1}$	$\mathbf{x}_{I+2}$	...	$\mathbf{x}_{I+(I-1)}$
Nivel 1 con n_1 individuos	1	0	0	...	0	z_1	0	0	...	0
Nivel 2 con n_2 individuos	1	1	0	...	0	z_2	z_2	0	...	0
Nivel 3 con n_3 individuos	1	0	1	...	0	z_3	0	z_3	...	0
.										
.										
Nivel I con n_I individuos	1	0	0	...	1	z_I	0	0	...	z_I

De nuevo, cada una de estas filas se repite en la matriz de diseño X tantas veces como individuos observados en cada nivel i, n_i para $i = 1, \ldots, I$. El vector de coeficientes para el modelo (6.8) es entonces:

$$\beta' = (\mu, \alpha_2, \alpha_3, \ldots, \alpha_I, \beta, \gamma_2, \gamma_3, \ldots, \gamma_I) \in \mathbb{R}^{2I}.$$

Esta construcción de X es equivalente a asumir como restricción de identificabilidad, $\alpha_1 = \gamma_1 = 0$, de modo que la respuesta media en el primer nivel del factor se predice a través del efecto global y el efecto de la covariable:

$$E(\mathsf{y}_1) = \mu + \beta x,$$

y los siguientes niveles ya contienen el efecto diferencial (añadido) respecto de la respuesta para el primero.

La generalización del modelo de Ancova a dos o más covariables y dos o más factores de clasificación se complica algo pero es inmediata.

6.1.6. Ajuste del modelo

Las hipótesis sobre los errores de incorrelación, varianza constante y media cero son suficientes para obtener el ajuste por mínimos cuadrados. La normalidad es necesaria para obtener las inferencias y concluir sobre su fiabilidad.

Para estimar β seguimos el criterio de minimizar la suma de cuadrados debida al error, esto es,

$$min_{\beta}\ \epsilon'\epsilon, \tag{6.9}$$

con

$$\epsilon'\epsilon = (\mathbf{y} - X\beta)'(\mathbf{y} - X\beta) = \mathbf{y}'\mathbf{y} - 2\beta' X'\mathbf{y} + \beta' X' X\beta. \tag{6.10}$$

Tras derivar (6.10) respecto de β e igualarlo a cero, se obtiene el estimador de mínimos cuadrados de β para el modelo (6.1), $\hat{\beta}$, resolviendo las p *ecuaciones normales*:

$$X'X\beta = X'\mathbf{y}. \tag{6.11}$$

A la hora de resolver (6.11), se pueden presentar dos situaciones:

- Las p ecuaciones normales que resultan de (6.11) no son independientes y por lo tanto no existe la inversa de $X'X$. Esto ocurre cuando las variables explicativas no son independientes entre sí. Entonces el modelo ha de expresarse en términos de menos parámetros (modificarse) o han de incorporarse restricciones adicionales sobre los parámetros para dar una matriz no singular.

  Por ejemplo, en el modelo de Anova (6.6), la definición que hicimos de las variables dummy evita el problema de inversión de la matriz $X'X$.

  Cuando $(X'X)$ es singular, el estimador de β se obtiene a partir de una matriz inversa generalizada $X'X$, $(X'X)^{-}$, como:

$$\hat{\beta} = (X'X)^{-}X'\mathbf{y}. \tag{6.12}$$

La *inversa generalizada* A^- de una matriz A existe siempre, no es única, y es tal que verifica $AA^-A = A$ (ver Apéndice A).

Así, diferentes elecciones de la inversa generalizada $(X'X)^-$ producen diferentes estimaciones de β. Sin embargo, el modelo ajustado es el mismo, esto es, $\hat{\mathbf{y}} = X\hat{\beta}$ es invariante a la inversa generalizada elegida.

- Las p ecuaciones normales son independientes, con lo que $X'X$ es no singular y existe su inversa. El estimador de mínimos cuadrados resulta:

$$\hat{\beta} = (X'X)^{-1}(X'\mathbf{y}). \tag{6.13}$$

Este estimador de mínimos cuadrados coincide con el máximo verosímil, ya que bajo la hipótesis de normalidad de los errores aleatorios, la verosimilitud conjunta tiene la forma:

$$L(\beta;\mathbf{y}) \propto f(\mathbf{y};\beta) \propto \left(\frac{1}{\sigma^2}\right)^{n/2} exp\left\{-\frac{(\mathbf{y}-X\beta)'(\mathbf{y}-X\beta)}{2\sigma^2}\right\}, \tag{6.14}$$

y maximizar la verosimilitud es equivalente a minimizar la log-verosimilitud cambiada de signo, que coincide con la suma de cuadrados del error (6.9) para un valor fijo de σ^2.

Ejemplo 6.1. *Obtener el ajuste de un modelo de regresión lineal múltiple para los datos del Ejemplo 3.2.*

```
# Cargamos primero los datos a partir del Apéndice.

# Queremos explicar VOL(volumen de madera) con las restantes variables.
# El ajuste lineal por máxima-verosimilitud lo obtenemos con:
fit<-lm(VOL~DBH+D16+HT,data=bosque); fit
# que coincide con la estimación por mínimos cuadrados:
fit.ls<-lsfit(cbind(DBH,D16,HT),VOL); fit.ls$coefficients
```

El ajuste obtenido nos permite predecir el volumen de madera (VOL) de un árbol en función de las restantes mediciones (no destructivas) observadas (DBH, D16 y HT), a través del modelo lineal:

$$\hat{VOL} = -108{,}5758 + 1{,}6258\, DBH + 5{,}6714\, D16 + 0{,}6938\, HT.$$

Ejemplo 6.2. *Obtener el ajuste del modelo de regresión polinómica de orden 2 propuesto para los datos del Ejemplo 3.5.*

```
# Cargamos los datos  del Apéndice y ajustamos el modelo lineal:
fit<-lm(tension~madera+I(madera^2)); fit
```

El ajuste obtenido predice la tensión del papel obtenido (variable 'tension') en función de la concentración de madera en la pulpa con la que se elabora el mismo (variable 'madera') viene dado por el modelo polinómico de orden 2:

$$\hat{tensi}on = -6{,}6742 + 11{,}7640\, madera - 0{,}6345\, madera^2.$$

Ejemplo 6.3. *Obtener el ajuste del modelo de Anova propuesto en el Ejemplo 3.6.*

```
# Cargamos los datos del Apéndice  y ajustamos el modelo,
# asegurándonos de que la variable explicativa 'tratamiento' se ha
# definido como factor:
fit<-lm(datos~tratamiento,x=T); fit
# (Intercept)  tratamientoPlacebo
#     1.568              -1.239
# y además podemos recuperar y visualizar la matriz de diseño X:
fit$x
```

El modelo ajustado permite estimar la reducción media en la tensión arterial tras tomar el fármaco o el placebo. Así, la reducción media estimada para los que tomaron el fármaco fue de:

$$\bar{reduccion}.hipertension = 1{,}568,$$

mientras que en grupo de los que tomaron placebo, dicha reducción media resultó de:

$$\bar{reduccion}.hipertension = 1{,}568 - 1{,}239 = 0{,}329.$$

Claramente se aprecia una mayor reducción en la hipertensión para los individuos que tomaron el fármaco que para los que tomaron el placebo. Faltará comprobar que dicha diferencia es estadísticamente significativa.

Ejemplo 6.4. *Obtener el ajuste de los dos modelos propuestos en el Ejemplo 3.8: rectas paralelas y rectas distintas.*

```
# Cargamos los datos del Apéndice.

# Ajustamos el modelo de rectas paralelas (sin interacción)
fit1<-lm(vida~velocidad+herramienta,x=T); fit1
# y el de rectas distintas (con interacción)
fit2<-lm(vida~velocidad*herramienta,x=T); fit2
```

Planteábamos dos posibilidades para estimar el tiempo de vida de una pieza cortadora en función de la velocidad del torno en el que se integra y del tipo de pieza (A o B).

El primer modelo propuesto pretendía asumir la misma relación velocidad ∼ tpo.vida para los dos tipos de piezas, esto es, la estimación de dos rectas paralelas, que resultan:

$$\begin{aligned} &\textit{Tipo A} && \hat{tpo.vida} = 36{,}92916 - 0{,}02608\, velocidad \\ &\textit{Tipo B} && \hat{tpo.vida} = 51{,}595 - 0{,}02608\, velocidad, \end{aligned}$$

dado que para el Tipo B la interceptación se construye a través del término de interceptación 36.92916, más el efecto incremental sobre éste del nivel 'HerramientaB', 14.66583.

El segundo modelo era menos estricto y daba la posibilidad de considerar un efecto diferente de la 'velocidad' sobre el tiempo de vida para cada uno de los tipos de piezas. Esto equivale a considerar un término de interacción 'velocidad:tipo'. El modelo resultante es el siguiente:

$$\begin{aligned} &\textit{Tipo A} && \hat{tpo.vida} = 32{,}87453 - 0{,}02072\, velocidad \\ &\textit{Tipo B} && \hat{tpo.vida} = 56{,}74535 - 0{,}03291\, velocidad. \end{aligned}$$

Las estimaciones de interceptación y el coeficiente de 'velocidad' respectan pues a la predicción para el tipo de cortadora A. El efecto incremental en la interceptación viene dado por la estimación para el nivel 'herramientaB' del factor 'herramienta', 23.87082, y el efecto diferencial en el coeficiente de la 'velocidad' viene contenido en la estimación de la interacción 'velocidad:herramientaB', -0.01219.

A primera vista, sí que parecen existir diferencias claras en el tiempo de vida medio según el tipo de cortadora, A o B, 23.87082, pero la diferencia en el efecto de la 'velocidad' sobre el tiempo de vida para los dos tipos de piezas es muy pequeña, -0.01219. Quedará por probar si dichas diferencias son significativas, para optar entonces por un modelo más simple (rectas paralelas) o preferir el más complejo (rectas distintas).

6.2. Propiedades del ajuste por mínimos cuadrados

Cuando prescindimos de la hipótesis de normalidad de los errores, obtenemos la estimación por mínimos cuadrados a partir de (6.9), que tiene las siguientes propiedades:

1. El estimador de mínimos cuadrados $\hat{\beta}$ minimiza $\epsilon'\epsilon$, independientemente de la distribución de los errores. La hipótesis de normalidad se añade para justificar las inferencias basadas en estadísticos t o F.

2. Los elementos de $\hat{\beta}$ son funciones lineales de las observaciones $y_1, \ldots, y_n$ y son estimadores insesgados de mínima varianza, sea cual sea la distribución de los errores. Así tenemos:

$$E(\hat{\beta}) = \beta \quad \text{y} \quad Var(\hat{\beta}) = \sigma^2(X'X)^{-1}. \tag{6.15}$$

3. Los estimaciones/predicciones de la variable respuesta y se obtienen con el modelo lineal ajustado:

$$\hat{\mathbf{y}} = X\hat{\beta}. \tag{6.16}$$

4. Los residuos $\mathbf{e} = \mathbf{y} - X\hat{\beta}$ verifican:

   - $\sum_{i=1}^n e_i\hat{y}_i = 0 \;\Leftrightarrow\; \mathbf{e}'\hat{\mathbf{y}} = \hat{\mathbf{y}}'\mathbf{e} = 0$
   - La ortogonalidad entre los vectores de estimaciones y de residuos, $\hat{\mathbf{y}}$ y $\mathbf{e}$ respectivamente, implica el teorema de Pitágoras:

$$|\mathbf{y}|^2 = |\hat{\mathbf{y}}|^2 + |\mathbf{e}|^2 \;\Leftrightarrow\; \sum_{i=1}^n y_i^2 = \sum_{i=1}^n \hat{y}_i^2 + \sum_{i=1}^n e_i^2.$$

   - $\sum_{i=1}^n e_i = 0 \;\Leftrightarrow\; \mathbf{e}'\mathbf{1} = \mathbf{1}'\mathbf{e} = 0$

6.3. Inferencia y predicción

Para hacer inferencia y predicción ya es preciso incorporar la hipótesis de normalidad de los errores. De ella podemos obtener la distribución de los estadísticos y estimadores involucrados en el proceso de inferencia con el modelo lineal ajustado.

6.3.1. Estimación de la varianza del modelo

Podemos obtener un estimador de σ^2 basado en la variabilidad que ha quedado sin explicar por el modelo, cuantificada por lo que llamamos suma de cuadrados

residual SSE:

$$SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \mathbf{e}'\mathbf{e} \tag{6.17}$$
$$= \mathbf{y}'\mathbf{y} - 2\hat{\beta}'X'\mathbf{y} + \hat{\beta}'X'X\hat{\beta}$$
$$= \mathbf{y}'\mathbf{y} - \hat{\beta}'X'\mathbf{y}.$$

Puesto que en el modelo lineal propuesto (6.1) se estiman p parámetros, la suma de cuadrados residual SSE tiene asociados $n - p$ grados de libertad (el número de datos menos el de coeficientes del modelo). El cociente entre SSE y sus grados de libertad, $n - p$, es el estimador de mínimos cuadrados de σ^2, y es además, un estimador insesgado:

$$\hat{\sigma}^2 = s^2 = MSE = \frac{SSE}{n-p}. \tag{6.18}$$

Asumiendo que el modelo (6.1) es cierto, su distribución es proporcional a una χ^2 con $n - p$ grados de libertad,

$$\frac{(n-p)s^2}{\sigma^2} \sim \chi^2_{n-p}. \tag{6.19}$$

Ejemplo 6.5 (Regresión múltiple de 'Bosque')**.** *Determinar el error residual y los grados de libertad asociados al error en el ajuste de regresión múltiple obtenido en el Ejemplo 6.1.*

```
# Cargamos los datos del Apéndice.

# Para el modelo de regresión múltiple ajustado
fit<-lm(VOL~DBH+D16+HT, data=bosque);fit
# la estimación de la varianza del modelo se obtiene del sumario:
sfit<-summary(fit); sfit
sfit$sigma^2    # = 9.581294
# y los grados de libertad asociados con
sfit$df[2]      #= 16
# o alternativamente con:
fit$df.residual
```

El error residual del modelo es de 3.095, con lo que al elevarlo al cuadrado obtenemos la estimación de la varianza residual $\hat{\sigma}^2 = 9{,}5813$.

6.3.2. Inferencia sobre los coeficientes del modelo

Bajo la hipótesis de normalidad de los errores, tenemos que el estimador máximo-verosímil $\hat{\beta}$ tiene una distribución normal:

$$\hat{\beta} \sim N(\beta, \sigma^2(X'X)^{-1}). \tag{6.20}$$

Esto implica que la distribución marginal de cada uno de los coeficientes de la regresión, $\hat{\beta}_i$, también es normal,

$$\hat{\beta}_i \sim N(\beta_i, \sigma^2 C_{ii}^X), \quad i = 1, \ldots, n, \tag{6.21}$$

con C_{ii}^X el i-ésimo elemento de la diagonal de la matriz $(X'X)^{-1}$.

En consecuencia, para construir intervalos de confianza o resolver contrastes sobre cada uno de los coeficientes del modelo, individualmente, podemos utilizar estadísticos t que se distribuyen con una distribución *t de Student* con $n - p$ grados de libertad:

$$\frac{\hat{\beta}_i - \beta_i}{\sqrt{s^2 C_{ii}^X}} \sim t_{n-p}, \quad i = 1, \ldots, n, \tag{6.22}$$

construidos a partir del estimador de σ^2, s^2, obtenido en (6.18).

Así, un intervalo de confianza para un coeficiente de interés β_i al nivel $(1 - \alpha)100\,\%$ viene dado por:

$$\hat{\beta}_i \pm t_{(n-p,1-\alpha/2)} \sqrt{s^2\ C_{ii}^X}, \tag{6.23}$$

donde $t_{(n-p,1-\alpha/2)}$ es el cuantil $1 - \alpha/2$ de una distribución t con $n - p$ grados de libertad.

El contraste $H_0 : \beta_i = 0$ se resolverá con el rechazo de H_0 a nivel $1 - \alpha$ si

$$|\hat{\beta}_i| > t_{(n-p,1-\alpha/2)} \sqrt{s^2\ C_{ii}^X}. \tag{6.24}$$

Cuando se pretende obtener intervalos de confianza para varios coeficientes del modelo a la vez, es recomendable ser más conservador. Hay diversas soluciones propuestas para realizar "comparaciones múltiples", esto es, testar todos los coeficientes a la vez, y obtener regiones de confianza conjuntas. Quizá el más conocido es el ajuste de Bonferroni, basado en sustituir el cuantil $t_{(n-p,1-\alpha/2)}$ en (6.23), por $t_{(n-p,1-\alpha/2q)}$, si q es el número de coeficientes para los que se desea

una estimación en intervalo. Se obtendrán entonces unos intervalos de confianza 'ensanchados' respecto a los intervalos de confianza individuales. Si no tenemos ninguna prioridad particular sobre determinados coeficientes, lo lógico será obtener conjuntamente los intervalos de confianza para todos los coeficientes del modelo, esto es, $q = p$.

Otra opción para la estimación en intervalo es construir una *región de confianza conjunta* para todos los parámetros β del modelo, determinando los puntos β de la elipse definida por:

$$(\beta - \hat{\beta})' X' X (\beta - \hat{\beta}) = (p+1)\ s^2\ F_{(p,n-p,1-\alpha)}, \tag{6.25}$$

donde $F_{(p,n-p,1-\alpha)}$ es el cuantil $1-\alpha$ de una distribución F con p y $n-p$ grados de libertad.

Es posible construir regiones de confianza conjuntas del tipo (6.25) para cualquier subconjunto de coeficientes del modelo. Bastará variar adecuadamente los grados de libertad p y $n-p$. Estas regiones acaban siendo complicadas de interpretar, especialmente cuando la dimensión de β es grande.

Ejemplo 6.6 (Inferencia en la Regresión múltiple de 'Bosque')**.** *Calcular la significatividad de los coeficientes del modelo de regresión lineal múltiple ajustado en el Ejemplos 6.1. Construir intervalos de confianza individuales y conjuntos a través de la corrección de Bonferroni y basados en la distribución de las elipses dada en* (6.25)*. Proporcionar y comentar los resultados para la estimación de los coeficientes de las variables DBH y D16.*

```
# Cargamos los datos del Apéndice.

# Para el modelo de regresión múltiple ajustado
fit<-lm(VOL~DBH+D16+HT,data=bosque); fit
# se consiguen los estadísticos t y su significatividad con:
sfit<-summary(fit); sfit
# y en concreto
sfit$coefficients
# donde tenemos los coeficientes (columna Estimate), errores estándar
# (columna Std.Error), los estadísticos t (columna t-value)
# y su significatividad asociada (columna  Pr(>|t|)).

# Directamente, podemos obtener los intervalos de confianza con:
ic.ind<-confint(fit,level=0.95);ic.ind
# o con
library(gmodels)
ci(fit)
# que además visualiza los coeficientes, estadísticos t y p-valores.
```

Así por ejemplo, el p-valor asociado al coeficiente de la variable D16 resulta de 0.0002325, por lo que dicha variable resulta claramente significativa (al 5 %) en el ajuste. Su estimación era 5.6713954 y el intervalo de confianza al 95 % es (3.1227,8.2201). No podemos decir lo mismo sobre DBH, con un p-valor de 0.1326.

```
# Calculamos la corrección de Bonferroni para los intervalos de
# confianza de los p coeficientes estimados en el modelo:
alpha<-0.05
estim<-fit$coefficients
error<-sfit$coefficients[,2]

p<-length(estim)
t.alpha<-qt(1-alpha/(2*p),df.residual(fit))
ic.bonf4<-cbind(inf=estim-t.alpha*error,sup=estim+t.alpha*error)
ic.bonf4

# Tenemos una función para calcular directamente los p-valores
# corregidos por Bonferroni:
p.adjust(sfit$coefficients[,4],method="bonferroni")

# Si queremos aplicar la corrección de Bonferroni exclusivamente
# a los dos coeficientes de interés (DBH y D16), entonces
p<-2
t.alpha<-qt(1-alpha/(2*p),df.residual(fit))
ic.bonf2<-cbind(inf=estim-t.alpha*error,sup=estim+t.alpha*error)
ic.bonf2

p.adjust(sfit$coefficients[2:3,4],method="bonferroni")
```

La corrección del p-valor por Bonferroni nos da, para DBH y D16, los resultados que aparecen en la Tabla 6.1, junto con los resultados individuales. Cuantos más sean los coeficientes para los que queremos hacer la corrección conjunta, más se ensancha el intervalo de confianza y mayor se hace el p-valor correspondiente.

Dibujemos a continuación los intervalos de confianza individuales para los coeficientes de DBH y D16, el conjunto de Bonferroni (con $p = 2$) y el proporcionado por la elipse.

	P-valor ajustado	Intervalo de confianza
	Individuales	
DBH	0.1326	(-0.5492,3.8007)
D16	0.0002	(3.1227 , 8.2201)
	Bonferroni para p=2	
DBH	0.2652	(-0.9113 , 4.1628)
D16	0.0005	(2.6984 , 8.6444)
	Bonferroni para p=4	
DBH	0.5304	(-1.2603,4.5118)
D16	0.0009	(2.2894,9.0534)

Tabla 6.1: P-valores e intervalos de confianza obtenidos individualmente y utilizando correcciones de Bonferroni para los coeficientes de DBH y D16 en el ajuste del modelo de regresión múltiple para 'Bosque'.

```
# Dibujamos
library(ellipse)

# La elipse de confianza para dichos coeficientes se obtiene con:
plot(ellipse(fit,2:3),type="l",main="Regiones de Confianza")
# con las estimaciones en el punto:
points(fit$coef[2],fit$coef[3])

# sobre el que superponemos los intervalos de confianza individuales:
abline(v=ic.ind[2,])
abline(h=ic.ind[3,])

# y los intervalos de confianza de Bonferroni para p=2:
abline(v=ic.bonf2[2,],lty=2,col="red")
abline(h=ic.bonf2[3,],lty=2,col="red")

# y para p=4
abline(v=ic.bonf4[2,],lty=3,col="blue")
abline(h=ic.bonf4[3,],lty=3,col="blue")

legend(2,8,c("Individual","Bonferroni(p=2)","Bonferroni(p=4)"),lty=1:3,
col=c("black","red","blue"))
```

En la Figura 6.2 se aprecia el ensanchamiento del intervalo de confianza a medida que imponemos más confianza conjunta sobre un mayor número de coeficientes. La elipse es más restrictiva con los valores confiables posibles, si bien se encaja dentro del recinto delimitado por la corrección más severa de Bonferroni.

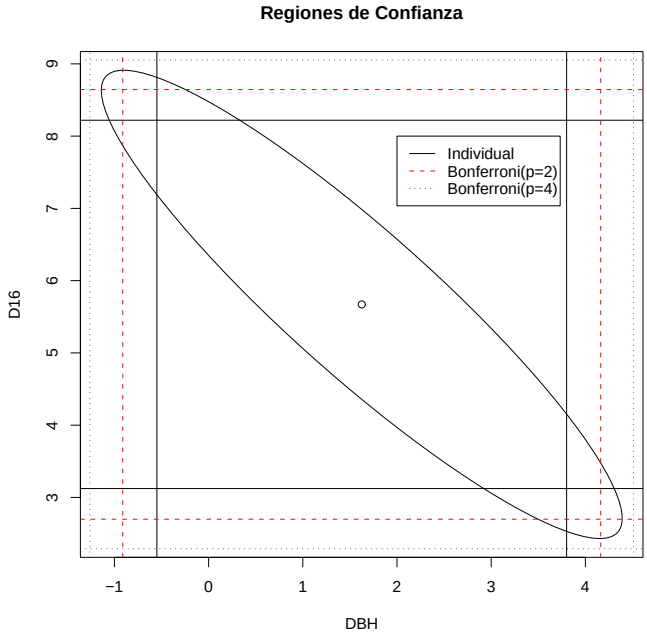


Figura 6.2: Intervalos de confianza individuales y conjuntos para los coeficientes de las variables DBH y D16 en el modelo de regresión múltiple para 'Bosque'.

6.3.3. Estimación de la respuesta media y predicción

Si $X_0 \in \mathbb{R}^p$ representa un vector fijo de valores de las variables explicativas contenidas en la matriz de diseño X, podemos predecir la respuesta $\mathbf{y}$ en X_0 a través del modelo ajustado con

$$\hat{\mathbf{y}} = X_0\hat{\beta},$$

pero el error asociado a la estimación depende de la situación que estemos prediciendo:

Estimación de la respuesta media. La varianza asociada a dicha estimación viene dada por:

$$Var[\hat{E}(\mathbf{y}|X_0)] = \sigma^2 X_0(X'X)^{-1}X_0'. \tag{6.26}$$

Un intervalo de confianza a nivel $1-\alpha$ está basado en la distribución $t-$*Student*:

$$\hat{y}_{X_0} \pm t_{(n-p,1-\alpha/2)}\ \sqrt{s^2 X_0(X'X)^{-1}X_0'}, \tag{6.27}$$

siendo $t_{(n-p,1-\alpha/2)}$ el cuantil $1-\alpha/2$ de una distribución $t-Student$ con $n-p$ grados de libertad, con p el número de coeficientes en el modelo y n el número de datos.

Ejemplo 6.7 (Estimación de la respuesta media en la regresión múltiple de 'Bosque')**.** *Calcular la estimación de la respuesta media para los valores medios de todas las variables explicativas. Representar cómo evoluciona dicha estimación en función de D16, fijando los valores de las restantes variables explicativas en sus medias.*

```
# Cargamos  los datos del Apéndice.

# Partimos del ajuste del modelo de regresión múltiple
fit<-lm(VOL~DBH+D16+HT); fit
# Queremos estimar el volumen esperado de madera cuando en las
# variables explicativas tenemos los valores medios
dbh.0<-mean(DBH)
d16.0<-mean(D16)
ht.0<-mean(HT)

# La estimación de la respuesta media viene dada por
fit.d16.0<-predict(fit,data.frame(DBH=dbh.0,D16=d16.0,HT=ht.0),
interval="confidence");fit.d16.0

# Si queremos estimar la respuesta media  en función de D16,
# dando el correspondiente intervalo de confianza, mantenemos
# fijas las demás y creamos una secuencia de valores para D16
# donde estimamos VOL:
d16<-seq(min(D16),max(D16),length=50)
fit.d16<-predict(fit,data.frame(DBH=dbh.0,D16=d16,HT=ht.0),
interval="confidence")

opar<-par(mfrow=c(1,1))
plot(d16,fit.d16[,1],type="l",xlab="D16",ylab="VOL",
main="Estimación Respuesta Media")
lines(x=rep(d16.0,2),y=c(fit.d16.0[2],fit.d16.0[3]))
lines(d16,fit.d16[,2],lty=2,col="red")
lines(d16,fit.d16[,3],lty=2,col="red")
par(opar)
```

Predicción de nuevas observaciones. La predicción de la respuesta $\mathbf{y}$ para un determinado valor X_0 de las variables explicativas involucra más incertidumbre que la estimación de un promedio. En este caso, la varianza asociada a la predicción es:

$$Var(\hat{y}|X_0) = \sigma^2(1 + X_0(X'X)^{-1}X_0'). \tag{6.28}$$

Un intervalo de confianza a nivel $1-\alpha$ para dicha predicción viene dado por:

$$\hat{y}_{X_0} \pm t_{(n-p,1-\alpha/2)} \sqrt{s^2\ (1 + X_0'(X'X)^{-1}X_0)}. \tag{6.29}$$

Ejemplo 6.8 (Predicción de una futura observación en la regresión múltiple de 'Bosque')**.** *Predecir el volumen de madera de un árbol cuyas mediciones en las variables DBH, D16 y HT corresponden a los valores medios de los datos observados. Representar la evolución de esta predicción (con su intervalo de confianza) cuando varía D16 y se mantienen fijas en los valores medios las otras variables explicativas. Comparar los intervalos de confianza obtenidos con la estimación de la respuesta media y la predicción de una nueva observación.*

```
# La predicción (en intervalo) de una nueva observación
# en las medias de las variables explicativas se obtiene con:
fit.d16.0.p<-predict(fit,data.frame(DBH=dbh.0,D16=d16.0,HT=ht.0),
interval="prediction"); fit.d16.0.p

# Para comparar la predicción con la estimación de la respuesta media
# cuando varía D16, calculamos la predicción en intervalo:
fit.d16.p<-predict(fit,data.frame(DBH=dbh.0,D16=d16,HT=ht.0),
interval="prediction");fit.d16.p
opar<-par(mfrow=c(1,1))
limits<-c(min(fit.d16.p[,2]),max(fit.d16.p[,3]))
plot(d16,fit.d16[,1],type="l",xlab="D16",ylab="VOL",
main="Estimación y Predicción",ylim=limits)

lines(d16,fit.d16[,2],lty=2,col="red")
lines(d16,fit.d16[,3],lty=2,col="red")
lines(x=rep(d16.0,2),y=c(fit.d16.0[2],fit.d16.0[3]),col="red",lty=2)

lines(d16,fit.d16.p[,2],lty=3,col="blue")
lines(d16,fit.d16.p[,3],lty=3,col="blue")
lines(x=rep(d16.0,2),y=c(fit.d16.0.p[2],fit.d16.0.p[3]),lty=3,
col="blue")

legend(10,90,c("Estimación","IC.Resp.Media","IC Predicción"),lty=1:3,
col=c("black","red","blue"))
par(opar)
```

En la Figura 6.3 apreciamos el resultado de la variabilidad extra en la predicción de nuevas observaciones, respecto de la estimación de la respuesta media.

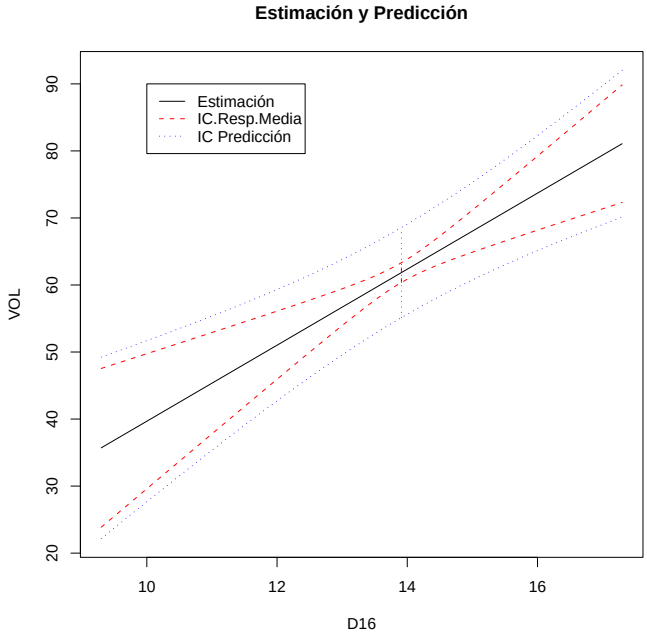


Figura 6.3: Estimación de la respuesta media y predicción en función de D16 para el modelo de regresión múltiple para 'Bosque'.

6.4. Descomposición de la variabilidad: Tabla de Anova y coeficiente de determinación

Habitualmente, la primera forma de juzgar la calidad del ajuste obtenido consiste en valorar la variabilidad de la respuesta que se ha podido explicar con el modelo propuesto. En lo que sigue, asumiremos que el modelo ajustado es $\hat{y} = X\hat{\beta}$, donde la matriz de diseño X tiene por columnas todas las variables explicativas consideradas, sean continuas o dummies definidas para representar algún factor. En todo caso, suponemos que hemos estimado p coeficientes, esto es, $\hat{\beta} \in \mathbb{R}^p$.

Descomponemos pues la variabilidad de las observaciones $\mathbf{y}$, en la parte explicada por el modelo ajustado y corregida por la media de los datos (suma de cuadrados de la regresión), SSR, y la parte residual (suma de cuadrados debida al error) que ha quedado sin explicar, SSE:

$$\underbrace{(\mathbf{y} - \bar{y}\mathbf{1})'(\mathbf{y} - \bar{y}\mathbf{1})}_{S_{\mathbf{yy}}} = \underbrace{(\hat{\mathbf{y}} - \bar{y}\mathbf{1})'(\hat{\mathbf{y}} - \bar{y}\mathbf{1})}_{SSR} + \underbrace{\mathbf{e}'\mathbf{e}}_{SSE},$$

donde $\bar{y} = \sum_i y_i/n$.

Los grados de libertad asociados a SSR son $p-1$, pues se pierde un parámetro al corregir la estimación $\hat{\mathbf{y}}$ (obtenida a partir de p parámetros) por la media $\bar{y}$. La suma de cuadrados del error SSE tiene asociados $n-p$ grados de libertad, esto es, el número de datos menos el número de parámetros estimados en el modelo. Al dividir las sumas de cuadrados por sus grados de libertad respectivos, obtenemos los cuadrados medios correspondientes, $MSR = SSR/(p-1)$ y $MSE = SSE/(n-p)$, que nos resultan útiles para valorar la bondad del ajuste. El test de bondad de ajuste propone el contraste:

$$H_0 : \beta_1 = \beta_2 = \ldots = \beta_{p-1} = 0, \qquad H_1 : \text{ algún } \beta_i \neq 0.$$

Cuando el modelo es bueno, MSR y MSE siguen sendas distribuciones proporcionales a chi-cuadrados independientes (con la misma constante de proporcionalidad σ^2), con $p-1$ y $n-p$ grados de libertad respectivamente; de ahí que su cociente (libre ya de la constante desconocida σ^2) resulta tener una distribución F con $p-1$ y $n-p$ grados de libertad:

$$F = \frac{SSR/(p-1)}{SSE/(n-p)} = \frac{MSR}{MSE} \sim F_{(p-1,n-p)}. \tag{6.30}$$

Así, con dicho estadístico F contrastamos si la variabilidad explicada por el modelo ajustado es suficientemente grande comparada con la que queda sin explicar (la de los residuos); en otras palabras, si el modelo ajustado es significativo para explicar la variabilidad de los datos. Si el p-valor asociado al estadístico F es inferior a la significatividad considerada (generalmente 0.05), rechazamos que el modelo propuesto no explique conjuntamente la respuesta, y concluimos a favor de que algunas de las covariables contienen información significativa para predecir la respuesta, esto es, a favor de la bondad del ajuste. En otro caso, no podemos garantizar significativamente la bondad del modelo propuesto.

La Tabla de Anova es la forma habitual de presentar toda la información de las sumas, medias de cuadrados, estadísticos F y p-valores asociados al contraste de bondad de ajuste del modelo, y tiene la forma habitual de la Tabla 6.2.

Fuente	g.l.	SS	MS	estadístico F	p-valor
Regresión	$p-1$	SSR	MSR	$F = MSR/MSE$	$Pr(F_{p-1,n-p} > F)$
Error	$n-p$	SSE	MSE		
Total	$n-1$	$S_{\mathbf{yy}}$			

Tabla 6.2: Tabla de Análisis de la varianza en el modelo lineal general.

La salida de la Tabla Anova que proporciona R no es exactamente la anterior. En dicha tabla, en lugar de contrastar globalmente el ajuste a través de la suma de cuadrados asociada a la regresión, se contrasta secuencialmente la significatividad de cada una de las covariables a la hora de explicar la variable respuesta en presencia de las variables que ya han sido incorporadas al modelo (las que quedan por encima en la salida). Sin embargo, con dicha salida es posible calcular el test F de bondad de ajuste.

Ejemplo 6.9 (Tabla de Anova para la regresión múltiple de 'Bosque')**.** *Interpretar la Tabla Anova que genera $\mathcal{R}$ para el ajuste obtenido sobre los datos del 'Bosque'. Obtener, a partir de ella, la tabla de Anova global tal y como se presenta en la Tabla 6.2 y resolver el test F de bondad de ajuste.*

```
# Cargamos los datos del Apéndice.

# Para la regresión múltiple
fit<-lm(VOL~DBH+D16+HT)
# la tabla de Anova se obtiene con
aov(fit)
# y dicha tabla junto con los p-valores calculados, con:
afit<-anova(fit); afit
# o con
summary(aov(fit))
```

Así, tenemos que a la hora de predecir el volumen de madera (VOL), la variable DBH es significativa frente a predecir dicha respuesta exclusivamente con su media (p-valor de $5{,}051e-12$*). Una vez incluida DBH para predecir VOL, el siguiente contraste de la tabla resuelve si D16 aporta información "adicional" sobre VOL de una manera significativa; la conclusión es que sí es significativa, dado su p-valor de* $2{,}303e-05$*. El último contraste es el que plantea si, una vez incorporadas para la predicción las variables DBH y D16, la variable HT explica algo más sobre VOL con significatividad; de nuevo un p-valor de* $0{,}0006056$ *nos permite afirmar que dicha variable es significativa en presencia de las restantes para predecir la respuesta.*

Para obtener el test F de bondad de ajuste a partir de esta tabla, basta con sumar las sumas de cuadrados de los términos de la regresión y sus correspondientes grados de libertad.

	TABLA DE ANOVA				
	gl	SS	MS	F	p-valor
[h] Regresión	3	3591.0577	1197.0192	124.9329	0
Error	16	153.3007	9.5813		
Total	19	3744.3585			

Tabla 6.3: Tabla de Anova para la regresión múltiple de 'Bosque'.

```
# Utilizamos 'afit'
is.list(afit)
# Determinamos primero el número de efectos en la tabla, siendo
# el último el SSE residual; hemos de sumar los restantes:
nterminos<-length(afit[[1]])
gl.ssr<-sum(afit[[1]][-nterminos]);gl.ssr  # afit[[1]]=g.l.
ssr<-sum(afit[[2]][-nterminos]);ssr     # afit[[2]]=SS
gl.sse<-afit[[1]][nterminos]
sse<-afit[[2]][nterminos]
mse<-afit[[3]][nterminos] ;mse          # afit[[3]]=MS
# y calculamos el estadístico F y sus g.l.:
f<-(ssr/gl.ssr)/mse;f       #=124.9329
gl.f<-c(gl.ssr,gl.sse);gl.f #=(3,16)
# y podemos calcular también el p-valor asociado
p.f<-1-pf(f,gl.f[1],gl.f[2])
# que coinciden con los que proporcionaba 'summary'
summary(fit)

# Pintamos la tabla Anova en formato 'tradicional'
n<-length(VOL)
sink("anova.tabla")
cat("-------------------------------------------------------------\n")
cat("TABLA DE ANOVA \n")
cat("           gl        SS         MS           F          p-valor \n")
cat(paste("Regresión  ",gl.ssr,"   ",round(ssr,4),"   ",
round(ssr/gl.ssr,4),"   ",round(f,4),"   ",
round(,4),"\n"))
cat(paste("Error     ",gl.sse,"   ",round(sse,4),"     ",round(mse,4),"\n"))
cat(paste("Total     ",n-1,"   ",round((n-1)*var(VOL),4),"\n"))
sink()
```

La Tabla 6.3 de Anova resultante ya contiene la información completa de una tabla de Anova tradicional:

El coeficiente de determinación múltiple R^2 se define como la parte proporcional de la variabilidad de los datos que es explicada por el modelo ajustado:

$$R^2 = \frac{SSR}{S_{\mathbf{yy}}} = 1 - \frac{SSE}{S_{\mathbf{yy}}}. \tag{6.31}$$

Por definición tenemos que $0 \leq R^2 \leq 1$. Un ajuste perfecto de los datos produciría $R^2 = 1$. Si ninguna de las variables predictoras $\mathrm{x}_1, \ldots, \mathrm{x}_{p-1}$ es útil para explicar la respuesta $\mathbf{y}$, entonces $R^2 = 0$.

Siempre es posible conseguir R^2 suficientemente grande, simplemente añadiendo más términos en el modelo. Por ejemplo, si hay más de un valor de $\mathbf{y}$ para un mismo x observado, un polinomio de grado $n-1$ proporcionará un ajuste "perfecto" ($R^2 = 1$) para n datos. Cuando esto no ocurre y hay únicamente un valor de $\mathbf{y}$ por cada x, R^2 nunca puede ser igual a 1 porque el modelo no puede explicar la variabilidad debida al *error puro*.

Aunque R^2 siempre aumenta cuando añadimos una variable explicativa al modelo, esto no significa necesariamente que el nuevo modelo sea superior al antiguo, es decir, que dicha variable sea útil para explicar mejor los datos. A pesar de que la suma de cuadrados residual SSE del nuevo modelo se reduce por una cantidad igual al anterior MSE, el nuevo modelo tendrá un MSE mayor debido a que pierde un grado de libertad. Por lo tanto, el nuevo modelo será de hecho, peor que el antiguo.

En consecuencia, algunos analistas prefieren utilizar una versión ajustada del estadístico R^2. El R^2 *ajustado* penaliza los modelos que incorporan variables innecesarias dividiendo las sumas de cuadrados en (6.31) por sus grados de libertad, esto es,

$$R_a^2 = 1 - \frac{SSE/(n-p)}{S_{\mathbf{yy}}/(n-1)} = 1 - (1 - R^2)\left(\frac{n-1}{n-p}\right). \tag{6.32}$$

R_a^2 es preferible a R^2 cuando sus valores difieren mucho. Su interpretación tiene algún problema debido a que puede tomar valores negativos; esto ocurre cuando el estadístico F toma valores inferiores a 1 (o produce p-valores mayores que 0.5).

La relación entre el coeficiente de determinación R^2 y el estadístico F, en (6.30), viene dada por:

$$R^2 = \frac{(p-1)F}{(p-1)F + (n-p)} \quad \Leftrightarrow \quad F = \frac{(n-p)R^2}{(p-1)(1-R^2)}. \tag{6.33}$$

La distribución de R^2 es una $Beta((p-1)/2,(n-p)/2)$. Así pues, para contrastar conjuntamente la significatividad de todas las variables dependientes en la predicción de la respuesta $\mathbf{y}$ a través del modelo lineal (6.1), podemos utilizar igualmente el estadístico R^2 y su distribución $Beta$, o el estadístico F y su distribución correspondiente. Ambos resultados son equivalentes.

Ejemplo 6.10 (Bondad de ajuste con R^2 para la regresión múltiple de 'Bosque')**.** *Calcular el coeficiente de determinación y su versión ajustada para el ajuste de regresión lineal múltiple y concluir con ellos sobre la idoneidad del ajuste. Resolver el contraste de bondad de ajuste para la regresión múltiple de los datos de 'Bosque' a través de la distribución de R^2 y verificar su equivalencia con el test F de la Tabla Anova.*

```
# Nos aseguramos de tener los datos (del Apéndice)
# y el ajuste

fit<-lm(VOL~DBH+D16+HT)

# El coeficiente de determinación se obtiene con:
sfit<-summary(fit)
sfit$r.squared      #=0.9590582
# y su versión ajustada con:
sfit$adj.r.squared  #=0.9513816

# Calculamos los parámetros de la distribución Beta de R2:
n<-length(VOL)                  # número de datos
p<-length(fit$coefficients)     # nº parámetros estimados
par.r2<-c((p-1)/2,(n-p)/2)
# el p-valor asociado al R2 observado es
1-pbeta(sfit$r.squared,par.r2[1],par.r2[2])   #=2.587e-11
# que coincide con el p-valor para el estadístico F
sfit$fstatistic
```

Podemos concluir que el ajuste explica un 95.9 % de la variabilidad de los datos, y además es muy pequeña la diferencia con el valor del coeficiente de determinación ajustado. En base al coeficiente de determinación obtenido , el ajuste es muy bueno. Asimismo, dicha bondad viene corroborada estadísticamente con un p-valor de $2{,}5810^{-11}$*, que es justamente el p-valor que obteníamos para el estadístico F de la Tabla Anova.*

R^2 es el cuadrado del *coeficiente de correlación múltiple* entre $\mathbf{y}$ y el conjunto de variables explicativas consideradas $\mathrm{x}_1,\ldots,\mathrm{x}_{p-1}$. Esto es, da una medida conjunta de la relación lineal entre la variable respuesta y las predictoras. Asimismo,

R^2 representa el cuadrado de la correlación entre el vector de observaciones $\mathbf{y}$ y el de valores ajustados $\hat{\mathbf{y}}$.

Por otro lado, el coeficiente de correlación simple permite cuantificar la relación lineal entre cada predictor y la respuesta en términos absolutos, esto es, sin considerar el resto de variables. Para medir realmente la relación entre la respuesta y un predictor incluido en el ajuste de un modelo de regresión lineal múltiple, es preciso tener en cuenta la relación con el resto de covariables. Para ello podemos utilizar el coeficiente de correlación parcial, introducido ya en el Tema 4.

Para dos variables predictoras x_1 y x_2, la relación entre el coeficiente de correlación parcial y los coeficientes de correlación simple viene dada por:

$$r_{y2,1} = \frac{r_{y2} - r_{y1}r_{21}}{\sqrt{(1 - r_{y1}^2)(1 - r_{21}^2)}}. \tag{6.34}$$

Cuando $\mathbf{x}_1$ no está correlada ni con $\mathbf{y}$ ni con $\mathbf{x}_2$, el coeficiente de correlación parcial entre $\mathbf{x}_2$ e $\mathbf{y}$ dado $\mathbf{x}_1$, $r_{y2,1}$ coincide con el coeficiente de correlación simple r_{y2}.

La relación entre el coeficiente de correlación parcial y el coeficiente de correlación múltiple $R \equiv r_{12y}$, viene dada entonces por:

$$1 - R^2 = (1 - r_{y1}^2)(1 - r_{y2,1}^2), \tag{6.35}$$

donde R^2 es el coeficiente de determinación de la regresión múltiple.

Más tarde veremos que el coeficiente de correlación parcial ayudará a detectar problemas de *multicolinealidad*, esto es, de relación entre las variables explicativas.

Ejemplo 6.11 (Relación entre correlación múltiple y parcial en el ajuste de 'Bosque')**.** *Para el ajuste de regresión múltiple con los datos de 'Bosque', obtener los coeficientes de correlación parcial y múltiple entre VOL y las variables explicativas. Comentar los resultados.*

```
# Cargamos los datos del Apéndice
bosque<-data.frame(VOL=VOL,DBH=DBH,D16=D16,HT=HT)
# y calculamos las correlaciones parciales, junto con las simples
library(ggm)
correlations(bosque)
# El coeficiente de correlación múltiple se obtiene de la regresión
fit<-lm(VOL~DBH+D16+HT,data=bosque)
R2<-summary(fit)$r.squared;
sqrt(R2)    #=0.9793152
```

En particular, la correlación parcial entre VOL y DBH resulta 0.3683119, un valor algo pequeño, claramente distante de la correlación simple, 0.9078088. Esto es, cuando consideramos las restantes variables explicativas, DBH no explica tanto el volumen de madera como parecía a partir de un simple análisis de correlación simple.

Obtener a continuación la correlación parcial entre DBH y VOL a partir del coeficiente de determinación de regresiones parciales, siguiendo los siguientes pasos:

1. *Conseguir los residuos $eVOL$ del modelo de regresión que explica VOL con las variables D16 y HT.*
2. *Conseguir los residuos $eDBH$ del modelo de regresión que explica DBH con las variables D16 y HT.*
3. *Calcular el coeficiente de determinación del modelo de regresión que predice la parte de VOL que ha quedado sin explicar por las covariables D16 y HT, $eVOL$, con la información contenida en DBH y no explicada (compartida) por D16 y HT, $eDBH$.*

```
fit.vol<-lm(VOL~D16+HT,data=bosque)
fit.dbh<-lm(DBH~D16+HT,data=bosque)
fit.e<-lm(residuals(fit.vol)~residuals(fit.dbh))
# el coeficiente de determinación de dicha regresión es:
R2.e<-summary(fit.e)$r.squared; R2.e
# y su raíz cuadrada resulta el coeficiente de correlación parcial
sqrt(R2.e)      #=0.3683119
```

6.5. Contrastes lineales

Una vez ajustado un modelo lineal, con frecuencia interesa resolver ciertos contrastes "lineales" sobre los coeficientes, esto es, concluir sobre si determinadas combinaciones lineales de los coeficientes son iguales a cierto valor, $H_0 : a_0\beta_0 + a_1\beta_1 + \ldots + a_{p-1}\beta_{p-1} = b$. De hecho, el contraste de bondad de ajuste es un contraste de tipo lineal, así como el contraste de significatividad de cada predictor. Cuando existen predictores categóricos (factores), una vez comprobado que contribuyen significativamente a la explicación de la respuesta, suele ser interesante identificar entre qué niveles de clasificación existen diferencias significativas en la respuesta media; también éstos son contrastes de tipo lineal.

Un contraste de tipo lineal sobre los coeficientes del modelo siempre lo vamos a poder formular matricialmente de la forma:

$$H_0 : \ C\beta = 0, \tag{6.36}$$

donde C es una matriz de dimensión $m \times p$, y rango q, esto es, tal que sólo q de las m ecuaciones son independientes.

Por ejemplo, para el contraste de bondad de ajuste, $H_0 = \beta_1 = \ldots = \beta_{p-1} = 0$:

$$C = \begin{pmatrix} 0 & 0 & \ldots & 0 \\ 1 & 0 & \ldots & 0 \\ 0 & 1 & \ldots & 0 \\ . & . & \ldots & . \\ 0 & 0 & \ldots & 1 \end{pmatrix}$$

para el contraste de significatividad de un predictor x_i, $C = (0 \ldots \underbrace{1}_{i} \ldots 0)$. Para la comparación de dos niveles de un factor en un modelo de Anova, identificados por las variables dummies x_i y x_j, la matriz de contraste resulta $C = (0 \ldots \underbrace{1}_{i} 0 \ldots \underbrace{-1}_{j} 0 \ldots)$.

Así, si H_0 es cierta, podremos utilizar las q ecuaciones independientes para expresar q coeficientes β en función de los $p-q$ restantes. Veamos cómo resolver este tipo de contrastes.

El modelo completo, en términos de la matriz de diseño X, $\mathbf{y} = X\beta + \epsilon$, producía una suma de cuadrados residual SSE con $n - p$ grados de libertad asociados:

$$SSE = \mathbf{y}'\mathbf{y} - \hat{\beta}' X'\mathbf{y}.$$

El modelo si H_0 es cierta, esto es, asumiendo como cierta la restricción $C\beta = 0$, origina un modelo reducido de la forma:

$$E[\mathbf{y}|C\beta = 0] = \mathbf{Z}\gamma, \tag{6.37}$$

donde $\mathbf{Z}$ es una matriz $n \times (p - q)$, y γ un vector de $p - q$ coeficientes, tal que $\mathbf{Z}\gamma$ resulta de sustituir en $X\beta$ los β's dependientes.

El nuevo vector de coeficientes del modelo (6.37), γ, será estimado (si $\mathbf{Z}'\mathbf{Z}$ es no singular) según:

$$\hat{\gamma}|\{C\beta = 0\} = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{y},$$

y la suma de cuadrados asociada a este modelo resulta:

$$SSE_{C\beta=0} = \mathbf{y}'\mathbf{y} - \hat{\gamma}'\mathbf{Z}'\mathbf{y}, \tag{6.38}$$

con $n-(p-q)$ grados de libertad asociados.

La forma de contrastar (6.36) consiste simplemente en comparar la suma de cuadrados residual del modelo reducido, (6.38) con la del modelo completo (6.17). Surge así un estadístico F:

$$F_{C\beta=0} = \frac{(SSE_{C\beta=0} - SSE)/((n-p)-(n-(p-q)))}{SSE/(n-p)} = \frac{(SSE_{C\beta=0} - SSE)/q}{SSE/(n-p)}, \tag{6.39}$$

cuya distribución bajo H_0 es una F con q y $n-p$ grados de libertad. El contraste se resuelve con el rechazo de $H_0 : C\beta = 0$ a un nivel de confianza $1-\alpha$ si $F_{C\beta=0}$ supera el valor crítico definido por el cuantil $1-\alpha$ de una distribución F con q y $n-p$ grados de libertad, $F_{(1-\alpha,q,n-p)}$.

Ejemplo 6.12 (Contrastes lineales en el ajuste de 'Bosque')**.** *Considerar el ajuste lineal de los datos de 'Bosque'. Plantear y resolver como contrastes lineales los siguientes:*

1. *La variable DBH, ¿es significativa para predecir el volumen de madera cuando ya consideramos DB16 y HT? La matriz de contraste para $H_0 : \beta_1 = 0$, viene dada por $C = (0, 1, 0, 0)$, de dimensión 1×4.*

```
# Nos aseguramos de tener cargados los datos del Apéndice
# y consideramos el ajuste obtenido:
fit<-lm(VOL~.,data=bosque); fit

#Cargamos una librería de utilidades para contrastes
library(gregmisc)
# construimos la matriz de contraste:
matriz.C<-matrix(c(0,1,0,0),nrow=1)
# y el contraste lo resolvemos directamente con:
fit.c<-glh.test(fit,matriz.C,d=0); fit.c
# donde 'd' proviene del contraste general 'C beta=d'.


# En particular, la resolución de estos contrastes
# sobre la significatividad de los predictores
# mediante el test F, los proporciona la función:
drop1(fit,test="F")
```

El resultado del test F correspondiente es un p-valor de 0,1326*, lo que no permite aceptar que DBH aporte, con significatividad, información adicional para predecir el volumen de madera en presencia de las otras dos*

variables explicativas. El mismo p-valor habíamos obtenido con el test t, dada su equivalencia con el $F = \sqrt{t}$.

2. *Reajustar un modelo significativo y contrastar si el efecto de D16 para predecir el volumen de madera es 10 veces el de HT.*

   Reajustamos pues, un nuevo modelo sólo con DB16 y HT, tal que:

   $$E(VOL) = \beta_0 + \beta_1 D16 + \beta_2 HT.$$

   Además, queremos contrastar sobre este nuevo modelo si $\beta_1 = 10\beta_2$, *que expresado en forma de contraste lineal genera la matriz de contraste* $C = (0, 1, -10)$.

```
# El nuevo modelo es:
fit.2<-update(fit,.~.-DBH)
# la nueva matriz de contraste es:
matriz.C<-matrix(c(0,1,-10),nrow=1)
# y resolvemos con:
glh.test(fit.2,matriz.C,d=0)
```

   de modo que, de nuevo, con un p-valor de 0.7467, no podemos rechazar que el efecto de DB16 sobre VOL sea 10 veces superior al efecto de HT.

3. *Contrastar la bondad del ajuste planteándola como un contraste lineal. Para resolver como contraste lineal el contraste de bondad de ajuste,* $\beta_1 = \beta_2 = 0$, *utilizaremos la matriz de contraste de rango 2:*

   $$C = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

```
# Construimos la matriz de contraste:
matriz.C<-matrix(c(0,1,0,0,0,1),nrow=2,byrow=T)
# y resolvemos con:
glh.test(fit.2,matriz.C,d=0)
# que resulta igual al test F de bondad de ajuste:
summary(fit.2)
```

   Con un p-valor de 5.515e-12 concluimos a favor del modelo propuesto y rechazamos que ambas variables no predigan conjuntamente el volumen de madera.

6.6. Comparación y selección de modelos

La modelización de datos es siempre una faena tediosa debido a la innumerable cantidad de alternativas posibles. Está por determinar el tipo de modelo, las transformaciones más adecuadas, identificar las variables más relevantes, descartar las innecesarias, y posteriormente abordar la diagnosis y validación del modelo, que trataremos en las Secciones siguientes. Si el modelo está mal especificado, las estimaciones de los coeficientes pueden resultar considerablemente sesgadas. Una buena especificación del modelo es un trabajo, en general, complicado de obtener.

Si se ha optado por la modelización lineal de una respuesta en función de una serie de posibles variables predictoras, y el objetivo es seleccionar el mejor subconjunto de predictores para explicar la respuesta, el planteamiento es siempre el de obtener "buenas" predicciones. Sin embargo, sabemos que cuantos más predictores incluyamos en el modelo, mejores predicciones tendremos (menos sesgo), pero a la vez menos precisión sobre ellas (ya que la varianza es proporcional al número de variables predictoras en el modelo). Para la selección del "mejor" modelo habremos de llegar a un compromiso entre estos dos propósitos. Tratamos pues la selección de variables como un problema de comparación y selección de modelos.

Vamos a presentar diversos criterios para comparar modelos y seleccionar el mejor modelo de entre dos alternativas. En ocasiones todos darán los mismos resultados, pero generalmente no, por lo que habrá de ser el analista el que decida qué criterio utilizar en función de sus intereses prioritarios. La selección del modelo la podemos basar en:

1. la significatividad de los predictores;
2. el coeficiente de determinación ajustado;
3. el error residual del ajuste, s^2;
4. el estadístico C_p de Mallows;
5. el estadístico AIC (*Akaike Information Criteria*);
6. el error de predicción $PRESS$.

Una vez seleccionado el "mejor" modelo según el criterio elegido, habremos de proseguir la confirmación del mismo realizando la diagnosis y la validación

del modelo, que puede fallar en algún paso, lo que nos conduciría de nuevo a la reformulación del modelo (y todos los pasos que le siguen), optando por aquellas correcciones y/o transformaciones de variables sugeridas en el diagnóstico. La consecución del *mejor modelo* será pues, un procedimiento iterativo, basado en selección y valoración de la calidad del ajuste, diagnóstico y validación.

6.6.1. Comparación de modelos basada en las sumas de cuadrados

Supongamos que tenemos ajustado un modelo definido por p coeficientes. Si queremos valorar la contribución que hacen al ajuste de los datos un subconjunto de q variables predictoras adicionales, estamos planteando el contraste

$$H_0 : \mathbf{y} = X_p\beta_p + \epsilon, \quad vs. \quad H_1 : \mathbf{y} = X_{p+q}\beta_{p+q} + \epsilon. \tag{6.40}$$

Para resolverlo a través de un test de Anova basado en un estadístico F, basta con ajustar sendos modelos y calcular sus sumas de cuadrados respectivas, $SSE(p)$ y $SSE(p+q)$. Su diferencia representa la reducción del error debida a la inclusión de los q regresores adicionales, y bajo H_0 tienen una distribución chi-cuadrado, independiente de $SSE(p)$. Se puede definir entonces un estadístico F para realizar la comparación de modelos y resolver el contraste (6.40), dado por:

$$F_q = \frac{(SSE(p) - SSE(p+q))/q}{SSE(p)/(n-p)} \sim F_{q,n-p}. \tag{6.41}$$

Las q variables adicionales se consideran relevantes (significativas) en la explicación de la respuesta, si F_q tiene asociado un p-valor significativo.

6.6.2. Coeficiente de determinación ajustado

Para evitar los problemas de interpretación de R^2, no válido para comparar modelos con diferente número de predictores, podemos utilizar el coeficiente de determinación ajustado, R^2_a, definido para un modelo con p coeficientes, por:

$$R^2_a(p) = 1 - \left(\frac{n-1}{n-p}[1 - R^2(p)]\right), \tag{6.42}$$

donde $R^2(p)$ es el coeficiente de determinación para un modelo con p regresores,

$$R(p)^2 = \frac{SSR(p)}{S_{\mathbf{yy}}} = 1 - \frac{SSE(p)}{S_{\mathbf{yy}}}.$$

$R_a^2(p)$ no aumenta necesariamente con p, aunque es frecuente. De hecho, si se añaden q regresores al modelo, $R_a^2(p+q) > R_a^2(p)$ si y sólo si el estadístico F parcial en (6.41) con el que se contrasta la significatividad de las q variables añadidas es mayor que 1. Un criterio para seleccionar un modelo es elegir aquel que tiene máximo R_a^2.

Ejemplo 6.13 (Comparación de modelos con el test F y R^2 en el ajuste de 'Bosque')**.** *Comparar los modelos en que predecimos el volumen de madera (VOL) con las variables DBH, D16 y HT, y en el que prescindimos de DBH. Utilizar los criterios de suma de cuadrados y coeficiente de determinación ajustado.*

```
# Nos aseguramos de tener cargados los datos.
# Los dos modelos propuestos son:
fit.3<-lm(VOL~DBH+D16+HT)
fit.2<-update(fit.3,.~.-DBH)
# El test F para comparar los dos modelos se obtiene con
anova(fit.2,fit.3)
# donde hay que introducir los modelos ordenados
# de menor a mayor complejidad.

# También 'drop1' resuelve este contraste comparando
# el modelo ajustado con todos los que resultan de
# prescindir de una variable predictora:
drop1(fit,test="F")

# Calculamos el coeficiente de determinación ajustado:
cda.fit.3<-summary(fit.3)$adj.r.squared; cda.fit.3 #=0.95138
cda.fit.2<-summary(fit.2)$adj.r.squared; cda.fit.2 #=0.94706
```

Efectivamente, el test F (p-valor=0.1326) nos lleva a NO rechazar con significatividad el modelo más sencillo, o lo que es lo mismo, a no aceptar el modelo con los tres regresores DBH, D16 y HT. Sin embargo, el coeficiente de determinación ajustado da preferencia al modelo completo con los tres regresores; por poco que añade DBH, las sumas de cuadrados de la regresión salen ganando con un regresor más. Apreciamos pues, discrepancia entre los criterios.

6.6.3. C_p de Mallows

Mallows (1973) presentó un criterio de selección basado en el error cuadrático medio de los valores ajustados $\hat{y}_i$. Supongamos ajustado un modelo completo con todos los regresores disponibles, en el que se han estimado $p+q$ coeficientes, y que ha dado lugar a una estimación del error del modelo de s^2. El estadístico C_p

para un modelo en el que se estiman sólo p coeficientes se define como:

$$C_p = \frac{SSE(p)}{s^2} - (n - 2p), \tag{6.43}$$

donde $SSE(p)$ es la suma de cuadrados correspondiente al error en el modelo con p coeficientes.

Así pues, para el modelo completo tendremos un valor del estadístico igual al número de parámetros estimados $p + q$:

$$C_{p+q} = \frac{(n - (p + q))s^2}{s^2} - (n - 2(p + q)) = p + q.$$

Es decir, esencialmente se está comparando el error del modelo con p coeficientes con el del modelo completo con todos los regresores observados.

Kennard (1971) demostró que el estadístico C_p está relacionado con R_a^2 y R^2. Si un modelo con p parámetros es adecuado, entonces $E(SSE(p)) = (n - p)\sigma^2$ y además $E(s^2) = \sigma^2$, de donde aproximadamente $E(C_p) = p$. Así pues, cuando el modelo es adecuado, $C_p \approx p$. Representar C_p versus p puede ayudar a decidir por qué modelo optar, prefiriendo siempre aquel cuyo valor del C_p quede más próximo a p y preferiblemente más pequeño (los modelos con sesgo significativo suelen producir valores del C_p por encima de p; valores pequeños del C_p van relacionados con errores de estimación menores).

Ejemplo 6.14 (Selección de variables en el ajuste de 'Bosque' con el criterio Cp)**.** *Calcular y representar el estadístico C_p en función del número de parámetros p. Elegir en consecuencia el mejor modelo.*

```
# En R disponemos de varias librerías para calcular
# o visualizar el Cp en los mejores modelos.
#.....................................
# Posibilidad 1: cálculo y visualización
library(wle)
cp<-mle.cp(VOL~DBH+D16+HT)
# que nos da el mejor modelo y su Cp
cp
# (Intercept)  DBH   D16   HT   cp
#      1        1     1    1     4
# así como el resto de modelos posibles con sus Cp
cp$cp
# que son representados con:
plot(cp) # y pinta en azul la solución.
# Gana pues el modelo con los tres regresores.

#.....................................
# Posibilidad 2: cálculo en los r mejores modelos.
library(leaps)
fit.leaps<-leaps(x=bosque[,-1],y=bosque[,1],method='Cp')
fit.leaps
cbind(fit.leaps$which,p=fit.leaps$size,Cp=fit.leaps$Cp)

# También facilita los coeficientes de determinación
# para decidir sobre el mejor modelo:
leaps(x=bosque[,-1],y=bosque[,1],method="adjr2")
leaps(x=bosque[,-1],y=bosque[,1],method="r2")

#.....................................
# Posibilidad 3: visualización
library(faraway)
Cpplot(fit.leaps)
```

Los resultados dan como favorito el modelo con los tres regresores ($p = 4 = C_p$). Sin embargo, tampoco el modelo con los regresores $DBHyHT$ acaba mal colocado ($p = 3, C_p = 4{,}511$). De hecho, en la Figura 6.4 se aprecia la supremacía de los modelos con los tres regresores (1111) y con sólo DH16 y HT (1011), cuyos valores del estadístico C_p quedan relativamente cerca de sus p.

6.6.4. Los estadísticos AIC y BIC

El criterio de información de Akaike (Akaike, 1973) está basado en la función de verosimilitud e incluye una penalización que aumenta con el número de parámetros estimados en el modelo. Premia pues, los modelos que dan un buen

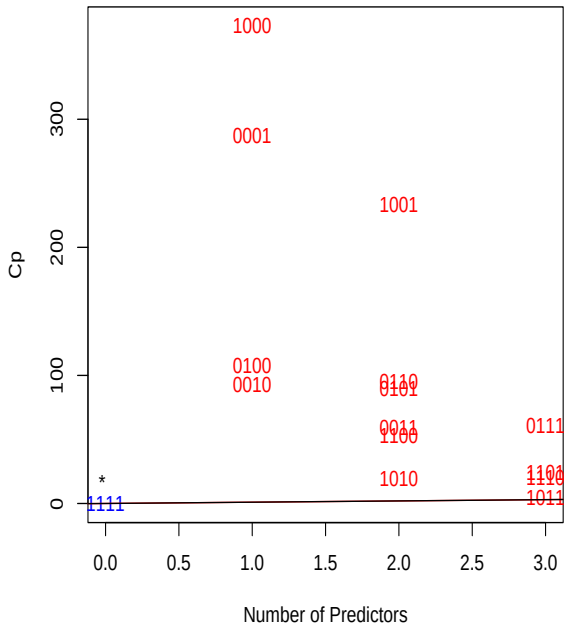


Figura 6.4: Gráficos de valores para el estadístico C_p en función de p. El mejor modelo está marcado con un asterisco (se visualiza en azul en $\mathcal{R}$). Ajuste de regresión para 'Bosque'

ajuste en términos de verosimilitud y a la vez son parsimoniosos (tienen pocos parámetros).

Si $\hat{\beta}$ es el estimador máximo-verosímil del modelo (6.1), de dimensión p, y $l(\theta)$ denota el logaritmo (neperiano) de la verosimilitud (6.14), el estadístico AIC se define por:

$$AIC = -2\, l(\hat{\beta}) + 2p. \tag{6.44}$$

Una versión del AIC que tiene en cuenta también el número de datos utilizados en el ajuste, es el *Schwarz's Bayesian criterion* (Schwarz, 1978), conocido como BIC, y definido por:

$$BIC = -2\, l(\hat{\beta}) + log(n)\, p. \tag{6.45}$$

Valores pequeños de ambos criterios identifican mejores modelos.

Ejemplo 6.15 (AIC y BIC en el ajuste de 'Bosque'). *Calcular los estadísticos AIC y BIC para comparar los modelos con tres regresores (DBH, D16 y HT) y con sólo dos (D16 y HT).*

```
# Obtenemos el AIC para sendos modelos:
AIC(fit.3)  #=107.4909
AIC(fit.2)  #=108.4066

# y el BIC con:
AIC(fit.3,k=log(nrow(bosque)))
AIC(fit.2,k=log(nrow(bosque)))

# o bien con:
library(stats4)
BIC(fit.3)    #=112.4696
BIC(fit.2)  #=112.3895
```

Si bien las diferencias entre los BIC o entre los AIC son muy pequeñas, según el AIC nos quedaríamos con un modelo con tres regresores, y según el BIC, que tiene en cuenta tanto el número de parámetros como el número de datos, el mejor modelo resultante es el que utiliza como predictores sólo D16 y HT.

6.6.5. Procedimientos secuenciales de selección de variables

Un procedimiento secuencial de selección de variables puede estar basado en cualquiera de los criterios presentados. La idea básica es partir de un modelo con cierto número de regresores, y secuencialmente moverse hacia modelos mejores (según el criterio elegido) con más o menos regresores de entre todos los observados. Una vez elegido el criterio para la selección, distinguimos básicamente entre los siguientes procedimientos secuenciales, en función de cuál es el punto (modelo) de partida y la forma de ir considerando modelos alternativos:

hacia adelante, se parte del modelo más simple y se van incluyendo una a una las variables que satisfacen el criterio de inclusión;

hacia atrás, se parte del modelo más complejo y se van excluyendo una a una las variables que satisfacen el criterio de exclusión;

paso a paso, se suele partir del modelo más simple y en cada paso se incluye o excluye la variable que satisface el criterio de inclusión/exclusión.

Los procedimientos hacia adelante y hacia atrás los hemos de llevar a cabo en $\mathcal{R}$ de forma manual. El procedimiento paso a paso es automático para el criterio del AIC.

Ejemplo 6.16 (Selección de variables secuencial en el ajuste de 'Bosque'). *Realizar sendos procedimientos secuenciales de selección de variables, el primero basado en el test F y el segundo en el criterio AIC. Concluir sobre cuál resulta el mejor modelo siguiendo cada procedimiento.*

```
# El test F secuencial hemos de hacerlo manualmente con las
# funciones 'drop1' y 'add1', partiendo del modelo completo:
fit<-lm(VOL~.,data=bosque)

drop1(fit,test="F")
# sale no significativa DBH, con p-valor=0.1326
# La quitamos y reajustamos
fit.1<-update(fit,.~.-DBH)
drop1(fit.1,test="F")
# Ya todas las variables salen significativas. Paramos!!
#-------------------------------------------------
# La selección secuencial con el AIC es automática en R:
step(fit)
```

Como ya habíamos visto, según el criterio F eliminamos del modelo DBH, pero lo preservamos si seleccionamos en función del criterio AIC.

6.7. Multicolinealidad

La multicolinealidad es un problema relativamente frecuente en regresión lineal múltiple, y en general en análisis con varias variables explicativas, entre cuyas soluciones se halla la selección de variables. Cuando los regresores no están relacionados linealmente entre sí, se dice que son *ortogonales*. Que exista multicolinealidad significa que las columnas de X no son linealmente independientes. Si existiera una dependencia lineal total entre algunas de las columnas, tendríamos que el rango de la matriz $X'X$ sería menor a p y $(X'X)^{-1}$ no existiría. El hecho de que haya multicolinealidad, esto es, una relación casi lineal entre algunos regresores, afecta a la estimación e interpretación de los coeficientes del modelo.

La multicolinealidad no es un problema de violación de hipótesis; simplemente es una situación que puede ocasionar problemas en las inferencias con el modelo de regresión. Nos ocupamos a continuación de examinar las causas de la

multicolinealidad, algunos de los efectos que tiene en las inferencias, los métodos básicos para detectar el problema y algunas formas de tratarlo.

6.7.1. Causas de la multicolinealidad

Montgomery y Peck (1992) comentan que la colinealidad puede surgir por el método de recogida de datos, restricciones en el modelo o en la población, especificación y sobreformulación del modelo (consideración de más variables de las necesarias); en modelos polinómicos, por ejemplo, se pueden presentar problemas serios de multicolinealidad en la matriz de diseño X cuando el rango de variación de los predictores es muy pequeño. Obviamente, modelos con más covariables son más propicios a padecer problemas de multicolinealidad.

6.7.2. Efectos de la multicolinealidad

Los principales efectos de la multicolinealidad son los siguientes:

• Una multicolinealidad fuerte produce varianzas y covarianzas grandes para los estimadores de mínimos cuadrados. Así, muestras con pequeñas diferencias podrían dar lugar a estimaciones muy diferentes de los coeficientes del modelo. Es decir, las estimaciones de los coeficientes resultan poco fiables cuando hay un problema de multicolinealidad. De hecho, dichos coeficientes vienen a explicar cómo varía la respuesta cuando varía la variable independiente en cuestión y todas las demás quedan fijas; si las variables predictoras están relacionadas entre sí, es inviable que al variar una no lo vayan a hacer las demás y en consecuencia puedan quedar fijas. La multicolinealidad reduce la efectividad del ajuste lineal si su propósito es determinar los efectos de las variables independientes.

• A consecuencia de la gran magnitud de los errores estándar de las estimaciones, muchas de éstas no resultarán significativamente distintas de cero: los intervalos de confianza serán 'grandes' y por tanto, con frecuencia contendrán al cero.

• La multicolinealidad tiende a producir estimaciones de mínimos cuadrados $\hat{\beta}_j$ muy grandes en valor absoluto.

• Los coeficientes del ajuste con todos los predictores difieren bastante de los que se obtendrían con una regresión simple entre la respuesta y cada variable explicativa.

• La multicolinealidad no afecta al ajuste global del modelo (medidas como la R^2, etc.) y por lo tanto no afecta a la habilidad del modelo para estimar puntualmente la respuesta o la varianza residual. Sin embargo, al aumentar los errores estándar de las estimaciones de los coeficientes del modelo, también lo hacen los errores estándar de las estimaciones de la respuesta media y de la predicción de nuevas observaciones, lo que afecta a la estimación en intervalo.

Ejemplo 6.17 (Multicolinealidad en el ajuste para 'Pizza')**.** *Verificar los efectos de la multicolinealidad (ya patentes en la Figura 3.4) en el ajuste de regresión lineal múltiple para los datos de Pizza, donde se pretendía predecir las ventas en función del número de anuncios publicitarios contratados y el coste agregado de éstos (Ejemplo 3.4).*

```
# Para los datos del Pizza Shack, en el Apéndice,
# ajustamos el modelo de regresión múltiple:
fit<-lm(sales~ads+cost,data=pizza,x=T)
# que da como resultado
summary(fit)
#Coefficients:
#            Estimate Std. Error t value Pr(>|t|)
#(Intercept)   6.5836     8.5422   0.771    0.461
#ads           0.6247     1.1203   0.558    0.591
#cost          2.1389     1.4701   1.455    0.180
```

Apreciamos, al menos para el término de interceptación, un coeficiente y un error estándar, considerablemente altos. Los restantes errores estándar no son en absoluto pequeños, teniendo en cuenta el tamaño de la estimación. Los p-valores resultantes son grandes, de manera que no se puede aceptar la significatividad de ninguna de las variables para la explicación de las ventas.

Veamos cómo difieren las estimaciones del modelo global con 'ads' y 'cost', de los modelos de regresión lineal simple que podemos construir con cada una de dichas variables explicativas:

```
# Ajustamos las regresiones simples:
fit.cost<-lm(sales ~ cost); summary(fit.cost)
fit.ads<-lm(sales ~ ads); summary(fit.ads)
```

Los modelos de regresión lineal simple obtenidos son:

$$\textit{M1: sales} = 4{,}173 + 2{,}872\,\textit{cost}$$
$$\textit{M2: sales} = 16{,}937 + 2{,}083\,\textit{ads},$$

cuyos coeficientes difieren considerablemente de los estimados para la regresión múltiple, especialmente la interceptación y el número de anuncios 'ads':

$$sales = 6{,}5836 + 0{,}6247\, ads + 2{,}1389 cost.$$

En lo que respecta al ajuste global del modelo, la R^2 no difiere demasiado para los tres modelos considerados: 0.684 para el modelo con los dos predictores, 0.6731 para la regresión simple con 'cost' y 0.6097 para la regresión simple con 'ads'. Tampoco difieren demasiado respecto al error estándar del modelo: 3.989 para el modelo completo, 3.849 para el de 'cost' y 4.206 para el de 'ads'.

6.7.3. Diagnósticos para multicolinealidad

Existen diversos diagnósticos propuestos para detectar problemas de multicolinealidad. Consideramos los más relevantes, que son:

1. Los gráficos entre variables explicativas son útiles para estudiar la relación entre las variables explicativas y su disposición en el espacio, y con ello detectar correlaciones o identificar observaciones muy alejadas del resto de datos y que pueden influenciar notablemente la estimación. Consisten en gráficos de dispersión entre un par de covariables continuas o un par de factores (a través de sus códigos), y gráficos de cajas cuando se trata de investigar la relación entre un factor y una covariable.

2. Una medida simple de multicolinealidad consiste en la inspección de los elementos fuera de la diagonal de la matriz $X'X$, es decir, las correlaciones simples r_{ij} entre todos los regresores. Si dos regresores x_i y x_j son casi linealmente dependientes, entonces $|r_{ij}| \approx 1$. Sin embargo, cuando la multicolinealidad involucra a varias variables, no hay garantías de detectarla a través de las correlaciones bivariadas.

3. Puesto que uno de los efectos principales de la multicolinealidad es la inflación de la varianza y covarianza de las estimaciones, es posible calcular unos factores de inflación de la varianza, FIV, que permiten apreciar tal efecto. En concreto, la varianza de $\hat{\beta}_j$ viene estimada por $Var(\hat{\beta}_j) = s^2\, C_{jj}$, donde C_{jj}^X son los elementos de la diagonal de la matriz $(X'X)^{-1}$, es decir,

$$C_{jj}^X = \frac{1}{(1 - R_j^2)\, S_{\mathrm{x}_j\mathrm{x}_j}}, \quad j = 1, 2, \ldots, p,$$

con R_j^2 el coeficiente de determinación múltiple para la regresión de x_j sobre las restantes $p-1$ covariables. Si hay una correlación muy alta entre x_j y los restantes regresores, entonces $R_j^2 \approx 1$. En particular, puesto que s^2 no varía ante un problema de multicolinealidad, si ésta existe, la varianza de $\hat{\beta}_j$ aumenta por un factor igual a $1/(1-R_j^2)$, que se define como el FIV para x_j:

$$FIV_j = 1/(1-R_j^2). \tag{6.46}$$

Generalmente, valores de un FIV superiores a 10 dan indicios de un problema de multicolinealidad, si bien su magnitud depende del modelo ajustado. Lo ideal es compararlo con su equivalente en el modelo ajustado, esto es, $1/(1-R^2)$, donde R^2 es el coeficiente de determinación del modelo. Los valores FIV mayores que esta cantidad implican que la relación entre las variables independientes es mayor que la que existe entre la respuesta y los predictores, y por tanto dan indicios de multicolinealidad.

4. Dado que la multicolinealidad afecta a la singularidad (rango menor que p) de la matriz $X'X$, sus valores propios $\lambda_1, \lambda_2, \ldots, \lambda_p$ pueden revelar multicolinealidad en los datos. De hecho, si hay una o más dependencias casi lineales en los datos, entonces uno o más de los valores propios será pequeño.

5. En lugar de buscar valores propios pequeños, se puede optar por calcular el número de condición de $X'X$, definido por:

$$\kappa = \lambda_{max}/\lambda_{min}, \tag{6.47}$$

que es una medida de dispersión en el espectro de valores propios de $X'X$. Generalmente, si el número de condición es menor que 100, no hay problemas de multicolinealidad. Números de condición entre 100 y 1000 implican multicolinealidad moderada, y mayores que 1000 implican multicolinealidad severa.

6. Los índices de condición de la matriz $X'X$ también son útiles para el diagnóstico de multicolinealidad y se definen por:

$$\kappa_j = \lambda_{max}/\lambda_j, \quad \mathrm{J} = 1, \ldots, p. \tag{6.48}$$

El número de índices de condición que son grandes (por ejemplo, ≥ 1000) es una medida útil del número de dependencias casi lineales en $X'X$.

7. Otra posibilidad de diagnóstico es a través de un análisis de componentes principales. Este tipo de análisis multivariante se plantea sobre conjuntos de variables relacionadas linealmente entre sí y tiene como finalidad la de

definir un conjunto menor de nuevas variables obtenidas como combinación lineal de las originales, y que a la vez resultan ortogonales entre sí. Si el análisis de componentes principales resulta significativo, estamos reconociendo multicolinealidad.

Ejemplo 6.18 (Diagnóstico de multicolinealidad en el ajuste de 'Pizza')**.** *Verificar la multicolinealidad en el ajuste de regresión múltiple para los datos de Pizza del Ejemplo 3.4.*

```
# Consideremos el ajuste de regresión
fit<-lm(sales~cost+ads,data=pizza,x=T)
# que tiene como matriz de diseño
x<-fit$x

# Inspeccionamos los elementos fuera de la diagonal de X'X
x.x<-t(x)%*%x;x.x
# que proporcionan las correlaciones entre 'ads' y 'cost',
# también calculadas directamente con:
cor(pizza)    #=0.8949

# Los valores FIV se calculan con:
library(faraway)
vif(fit)
# =  5.021806(cost)  5.021806(ads)
# que comparamos con 1/(1-R2):
1/(1-summary(fit)$r.squared)
#= 3.164693

# Los valores propios de X'X:
lambda<-eigen(x.x)$values;lambda
#= 2586.8330506   10.2242966    0.2126529

# Número de condición:
kappa<-max(lambda)/min(lambda); kappa
#= 12164.58

# Índices de condición:
kappa.j<-max(lambda)/lambda; kappa.j
#= 1.0000   253.0084 12164.5803


# Análisis de componentes principales
pr<-princomp(~ads+cost)
pr$loadings
# Componente 1: 0.804*ads-0.594*cost
summary(pr)
# Proportion of Variance 0.9513543
```

Encontramos indicios de multicolinealidad entre los regresores 'ads' y 'cost' respecto de todos los criterios de diagnóstico propuestos:

- *Ya con el gráfico de dispersión (Figura 3.4) apreciábamos una relación lineal directa entre las variables explicativas 'ads' y 'cost', ratificada por un coeficiente de correlación simple entre ellas de 0.8949.*
- *El* $FIV = 5{,}022$ *resulta mayor que su equivalente en el modelo,* $1/(1-R^2) = 3{,}165$.
- *Uno de los valores propios (el último,* $= 0{,}2126$*) es pequeño.*
- *El número de condición* $\kappa = 12164{,}58$ *es mayor a 1000.*
- *Hay un índice de condición superior a 1000: el último=* $12164{,}58$*, lo que justifica una relación de dependencia entre los regresores.*
- *El análisis de componentes principales da lugar a una combinación lineal de los dos regresores, dada por* $C = 0{,}804ads + 0{,}594cost$*, que explica un 95 % de la variabilidad de los datos (un resultado muy positivo).*

Ejercicio 6.1. *Verificar el problema de multicolinealidad existente en el ajuste de regresión lineal múltiple con los datos de Bosque (Ejemplo 3.2).*

6.7.4. Soluciones a la multicolinealidad

Una vez detectado un problema de multicolinealidad, es recomendable intentar aliviarlo (por sus efectos). Para ello disponemos de diversos recursos, y en función del objetivo del análisis, será más aconsejable uno u otro. Básicamente podemos distinguir como objetivos esenciales:

1. Estimar bien la respuesta media en función de un conjunto de variables explicativas, sin importar demasiado la contribución individual de cada una de esas variables.
2. Hacer un análisis de estructura, esto es, describir el efecto de las variables explicativas en la predicción de la respuesta. Las magnitudes y significatividades de los coeficientes son entonces de interés. Así, en un análisis de estructura es importante conseguir un buen modelo de ajuste para cuantificar bien la información que aportan las variables explicativas sobre la respuesta.

Hay tres aproximaciones básicas como remedio a la multicolinealidad:

Selección de variables (ver Sección 6.6). Respecto a la selección de variables, lo ideal ante un problema de multicolinealidad es seleccionar aquellas variables predictoras que son más significativas y contienen la mayor parte de la información sobre la respuesta. Sin embargo, hay que actuar con precaución, pues los métodos automáticos de selección de variables son bastante sensibles cuando existe relación entre los regresores y no está garantizado que el modelo resultante tenga menor multicolinealidad. Por otro lado, la capacidad predictiva del modelo puede verse seriamente menguada al reducir el número de covariables consideradas, de modo que este remedio iría más indicado cuando el objetivo del análisis es el 2.

Redefinición de variables. Otra alternativa es transformar las covariables. Para ello es importante identificar entre qué covariables hay relación, con el fin de utilizar transformaciones apropiadas. Si varias variables están relacionadas linealmente, a veces funciona considerar la más completa de ellas tal y como es, y transformaciones de las otras con cocientes o diferencias respecto de la más completa. Es decir, si x_i y x_j están relacionadas y x_i da una información más completa que x_j, se puede considerar un nuevo ajuste que involucre a las variables x_i y $\mathrm{x}_j/\mathrm{x}_i$, o bien a x_i y $\mathrm{x}_j - \mathrm{x}_i$.

Cuando la intuición o el conocimiento de las variables no sugiere ninguna transformación concreta, una opción es llevar a cabo un *análisis de componentes principales* con el fin de obtener nuevas variables, expresables como combinación lineal de las originales, ortogonales entre sí y que contengan toda la información disponible en las primeras. En ocasiones, las componentes que resultan tienen un significado intuitivo por la forma de asimilar la información de las variables originales, y en ocasiones no, en cuyo caso se puede proceder a la realización de un análisis factorial y a la búsqueda de alguna rotación geométrica que permita llegar a variables "interpretables".

Una vez obtenidas las componentes Z, se pueden seguir dos alternativas: i) plantear una regresión de la respuesta explicada por todas las componentes principales obtenidas, o ii)ajustar un modelo de regresión sólo las componentes más relevantes como variables predictoras (componentes principales incompletas). En el primer caso, a partir del modelo ajustado $\mathbf{y} = Z\gamma + \epsilon$, es posible recuperar el efecto de las variables originales sobre la respuesta sin más que deshacer el cambio. Esto no es posible para la segunda alternativa, pues las estimaciones que se consiguen están sesgadas; sin embargo, esta opción reduce la varianza de las estimaciones respecto del modelo original.

Estimación sesgada. Si uno de los efectos de la multicolinealidad es que au-

menta el error estándar de las estimaciones por mínimos cuadrados de los coeficientes del modelo, cabe la posibilidad de utilizar estimadores que, aun sesgados, produzcan estimaciones con menor error estándar y un error cuadrático medio inferior al de los estimadores de mínimos cuadrados (que son, de los insesgados, los de mínima varianza).

Hay varios procedimientos de estimación sesgada. Las *componentes principales incompletas* es uno de ellos. La *regresión Ridge* es otro método interesante.

La regresión Ridge (Hoerl y Kennard, 1970a,b) consiste en utilizar como estimador de β, el siguiente:

$$\hat{\beta}_k = (X'X + kI)^{-1}X'\mathbf{y}, \tag{6.49}$$

donde k es una constante pequeña arbitraria.

Cuando todos los predictores están estandarizados, tenemos que $X'X$ es la matriz de correlaciones, con unos en la diagonal. Así, la correlación "efectiva" que se consigue ahora entre x_i y x_j es $r_{ij}/(1+k)$. Es decir, todas las correlaciones se reducen artificialmente en un factor $1/(1+k)$, reduciendo entonces la multicolinealidad. Valores grandes de k reducen la multicolinealidad pero, como contraprestación, aumentan el sesgo de las estimaciones. Para determinar el valor de k a utilizar, se suelen considerar gráficos en los que se representa k versus las estimaciones del modelo (*ridge plots*). Para valores pequeños de k, las estimaciones de los coeficientes cambian mucho, mientras que a medida que k aumenta, las estimaciones parecen estabilizarse. Se dice que se consigue un valor óptimo para k cuando se da dicha estabilización en las estimaciones. Este procedimiento resulta pues, algo subjetivo, pero sin embargo ha resultado efectivo en la práctica.

Hay otros procedimientos sesgados de estimación propuestos en la literatura que alivian el problema de la multicolinealidad. Para ampliar, consultar por ejemplo, Montgomery y Peck (1992), pags.323-360, que proporciona numerosas fuentes bibliográficas.

Ejemplo 6.19 (Soluciones a la multicolinealidad en el ajuste de 'Pizza')**.** *Utilizar la selección de variables para resolver el problema de multicolinealidad sobre el ajuste de regresión múltiple con los datos de Pizza (Ejemplo 3.4).*

```
# Partimos del ajuste con los datos del Apéndice:
fit<-lm(sales~ads+cost,data=pizza)

# Procedemos a una selección automática de variables
fit.aic<-step(fit); fit.aic
# se excluye 'ads':  sales ~ cost

# Si hacemos la selección con el test F:
drop1(fit,test="F")
# también eliminamos 'ads' por tener mayor p-valor
# ads   p-valor=0.5907
# cost  p-valor=0.1797
# con lo que nos quedamos con el ajuste
fit.cost<-lm(sales ~ cost); summary(fit.cost)
```

En este caso, tanto por el procedimiento automático de selección basado en el AIC, como en función del test F, resolvemos que la variable a excluir es 'ads'. Así, el ajuste que resultaría al eliminarla proporciona como modelo de predicción:

$$sales = 4{,}17(7{,}109) + 2{,}87(0{,}633)cost.$$

Entre paréntesis figura el error estándar de la estimación de los correspondientes coeficientes. En este modelo, sin embargo, no es significativa la interceptación. Podemos optar por eliminarla:

```
fit.cost<-update(fit.cost,.~.-1)
summary(fit.cost)
# y probamos si podemos incluir ahora 'ads'
anova(fit.cost,update(fit.cost,.~.+ads),test="F")
```

Finalmente no incluimos ni interceptación ni 'ads' (p-valor=0.5907 para el modelo que la incluye, habiendo ya eliminado la interceptación). El modelo resultante, $sales = 3{,}24(0{,}096)cost$*, tiene un* $R^2 = 0{,}9904$ *considerablemente superior al del modelo inicial (*$R^2 = 0{,}684$*), si bien el error residual sale de orden similar,* $s = 3{,}733$*, con un p-valor asociado a la significatividad del modelo de* $p - valor(F) = 1{,}841e - 12$*.*

Ejemplo 6.20 (Soluciones a la multicolinealidad en el ajuste de 'Bosque')**.** *Implementar las soluciones que ofrecen las componentes principales incompletas y la regresión Ridge para aliviar el problema de multicolinealidad con los datos del 'Bosque' (Ejemplo 3.2). Comparar y comentar los resultados.*

```
# Cargamos los datos del Apéndice.
# Creamos nuevas variables ortogonales por componentes
# principales:
pr<-princomp(~DBH+HT+D16,data=bosque);pr
pr$loadings
# Las más relevantes son:
pr1<--0.1758344*DBH-0.9695333*HT-0.1705505*D16
pr2<- 0.7451155*DBH-0.2443003*HT+0.6205806*D16
# y ajustamos un nuevo modelo con ellas
# (componentes principales incompletas):
fit.pr<-lm(VOL~pr1+pr2)
summary(fit.pr)
```

El análisis de componentes principales nos proporciona dos considerablemente importantes (juntas explican un 99.28 % de la varianza). Esas componentes las utilizamos para ajustar un nuevo modelo de predicción que resulta significativo (p-valor del test F=1.313e-11):

$$VOL = -112{,}061 - 1{,}926pr1 + 4{,}561pr2.$$

A partir de él es posible recuperar la predicción en términos de las variables originales sin más que despejar $pr1$ y $pr2$ en términos de ellas.

Otra aproximación consiste en la regresión Ridge. Veamos qué resultados da:

```
library(MASS)
# Dibujamos el gráfico ridge para decidir el valor de k (lambda):
plot(lm.ridge(VOL~.,data=bosque,lambda=seq(0,3,0.01)))
title(main="Ridge Plot")
# y empleamos estimadores de dicho valor implementados en R:
select(lm.ridge(VOL~.,data=bosque))
#modified HKB estimator is 0.09691231
#modified L-W estimator is 0.05336196
#smallest value of GCV  at 0.65

# Ajustamos el modelo ridge con la estimación HKB:
fit.ridge.HKB<-lm.ridge(VOL~.,data=bosque,lambda=0.09691231)
fit.ridge.HKB
#                     DBH          D16           HT
#-108.406928    1.721365    5.541354    0.695394
# Y decidimos otro valor lambda a partir del gráfico ridge:
fit.ridge<-lm.ridge(VOL~.,data=bosque,lambda=1.5)
fit.ridge
                      DBH           D16            HT
-104.7096682    2.3458025    4.5627709    0.6976587
```

En el gráfico Ridge (Figura 6.5) apreciamos la tendencia de estabilización de los

coeficientes del modelo a partir del valor de $k > 1{,}5$ *(*λ *en el gráfico) aproximadamente. La estimación que nos proporcionan los procedimientos HKB y L-W sobre el valor de la constante k son de* 0,09691231 *y* 0,05336196 *respectivamente. Si ajustamos los modelos ridge con la estimación HKB y con el valor* $k = 1{,}5$*, observamos cómo los coeficientes que más varían son los relativos a las variables DBH y D16, que son las más correlacionadas entre sí (correlación simple de 0.9264282 y correlación parcial de 0.2686789).*

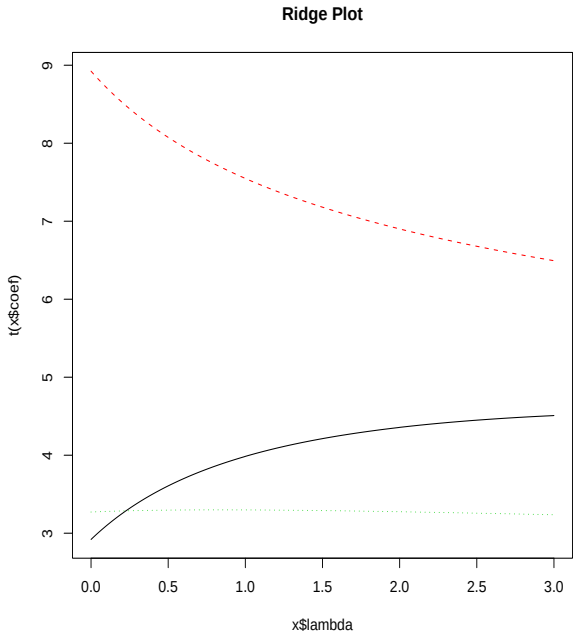


Figura 6.5: Gráfico Ridge para el ajuste de regresión múltiple con los datos de 'Bosque'.

6.8. Diagnóstico del modelo

La dinámica que proponemos para el tratamiento estadístico de un problema lineal es la siguiente:

1. la formulación del modelo en función del problema a resolver, los datos observados, objetivos e hipótesis que se asumen tras una revisión gráfica de los mismos;

2. la estimación de los parámetros del modelo y comprobación de la bondad del ajuste

3. comparación y selección de modelos y/o variables

4. el diagnóstico del modelo basado en los residuos

5. el análisis de influencia para detectar datos anómalos

6. la validación del modelo ajustado

7. la obtención de inferencias y respuestas a los problemas/objetivos inicialmente planteados.

Con todo, durante el ajuste de un banco de datos, habremos de rodar sucesivamente sobre los pasos 2 a 6 hasta dar con un modelo apropiado del que extraer finalmente conclusiones (paso 7). El diagnóstico del modelo es realmente valioso por cuanto nos permite corroborar que se cumplen (o no) cada una de las hipótesis asumidas para el ajuste del modelo y que dan credibilidad a las conclusiones que obtenemos. Este diagnóstico suele sugerir con frecuencia alguna modificación correctora del modelo que nos lleva de nuevo al paso 2 y nos obliga a repetir la dinámica de análisis hasta dar con una solución satisfactoria.

La herramienta básica para el diagnóstico del modelo es el análisis de los residuos, tanto a través de gráficos, como de tests que verifican la validez de las hipótesis asumidas en el ajuste del modelo lineal, y que son:

- $E(\epsilon_i) = 0,\ \forall i = 1, \ldots, n \ \rightsquigarrow$ bondad del ajuste.
- $Cov(\epsilon_i, \epsilon_j) = 0, \forall\ i \neq j \ \rightsquigarrow$ Independencia de los errores.
- $Var(\epsilon_i) = \sigma^2,\ \forall i \ \rightsquigarrow$ Varianza constante (homocedasticidad).
- $\epsilon \sim N(\mathbf{0}, \sigma^2 I) \ \rightsquigarrow$ Normalidad de los errores.

Si encontramos indicios de violación de alguna de ellas, en ocasiones podremos resolverlas a través de las soluciones que proponemos a continuación. Tanto las herramientas de diagnóstico como las soluciones propuestas para cuando encontramos problemas, son una ampliación del análisis de residuos que ya estudiamos para el modelo de regresión lineal simple. De hecho, primeramente presentamos los principales tipos de residuos sobre los que se realiza el diagnóstico.

6.8.1. Tipos de Residuos

Presentamos diversos tipos de residuos, útiles tanto para la diagnosis del modelo como para el análisis de influencia (detección de observaciones influyentes y/o raras o anómalas). Generalmente, los procedimientos de diagnóstico del modelo basados en residuos son gráficos, si bien en ocasiones disponemos de algunos tests basados en ellos.

Residuos comunes Los residuos comunes del modelo lineal $\mathbf{y} = X\beta + \epsilon$ consisten simplemente en las desviaciones entre los datos observados y_i y los predichos $\hat{y}_i$, esto es, los obtenidos de:

$$\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}} = y - X\hat{\beta} = \mathbf{y} - X\hat{\beta} = (I - X(X'X)^{-1}X')\mathbf{y} \tag{6.50}$$

cuando $X'X$ es no singular.

Surge así una matriz básica en la definición de los residuos, denominada *matriz gorro* y definida por:

$$H = X(X'X)^{-1}X', \tag{6.51}$$

que tiene su importancia en la interpretación y redefinición de nuevos tipos de residuos, como veremos. A sus elementos nos referiremos como h_{ij}. Esta matriz H es simétrica ($H' = H$) e idempotente ($HH = H$), de dimensión $n \times n$ y de rango $p = rang(X)$.

En términos de H, los residuos $\mathbf{e}$ se pueden escribir como:

$$\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}} = (I - H)\mathbf{y},$$

esto es,

$$e_i = (1 - \sum_{j=1}^{n} h_{ij})\, y_i = y_i - \hat{y}_i, \quad i = 1, \ldots, n. \tag{6.52}$$

Así, los residuos $\mathbf{e}_i$ son estimadores sesgados de los errores aleatorios ϵ_i, cuyo sesgo y varianza depende de la matriz H gorro, y por lo tanto de la matriz de diseño del experimento (X):

$$\begin{aligned} \mathbf{e} - E(\mathbf{e}) &= (I - H)(\mathbf{y} - X\beta) = (I - H)\epsilon, \\ Var(\mathbf{e}) &= (I - H)\ \sigma^2, \end{aligned} \tag{6.53}$$

de donde la varianza de cada residuo es:

$$Var(e_i) = (1 - h_{ii})\sigma^2, \quad i = 1, \ldots, n,$$

y la correlación entre los residuos e_i y e_j:

$$Cor(e_i, e_j) = \frac{-h_{ij}}{\sqrt{(1-h_{ii})(1-h_{jj})}}.$$

Residuos estandarizados. Son residuos de media cero y varianza aproximadamente unidad, definidos por:

$$d_i = \frac{e_i}{\sqrt{s^2}}, \qquad i = 1, \ldots, n, \tag{6.54}$$

donde s^2 es la estimación habitual de σ^2 que da el cuadrado medio residual (6.18).

Residuos (internamente) estudentizados. Son unos residuos estandarizados (con varianza aproximadamente igual a 1) que tratan de librar a éstos de la variabilidad que introduce la matriz de diseño:

$$r_i = \frac{e_i}{\sqrt{s^2(1-h_{ii})}}. \tag{6.55}$$

La distribución marginal de $r_i/(n-p)$ es una $Be(1/2, (n-p-1)/2)$. Así, cuando los grados de libertad del error, $n-p$, son pequeños, ningún residuo $|r_i|$ será demasiado grande. Se suelen utilizar para la detección de outliers u observaciones influyentes.

Residuos externamente estudentizados. Se trata de otro tipo de residuos estandarizados calculados según:

$$rt_i = \frac{e_i}{\sqrt{s^2_{(i)}(1-h_{ii})}}, \tag{6.56}$$

con $s^2_{(i)}$ la estima de la varianza en el ajuste sin la observación i-ésima,

$$s^2_{(i)} = \frac{(n-p)s^2 - e_i^2/(1-h_{ii})}{n-p-1}. \tag{6.57}$$

Estos residuos siguen una distribución t_{n-p-1} cuando los errores ϵ son normales. Proporcionan pues, un procedimiento más formal para la detección de outliers vía contraste de hipótesis. De hecho, se puede utilizar una corrección de Bonferroni para comparar todos los n valores $|rt_i|$ con el cuantil $t_{n-p-1,\alpha/2n}$, e identificar así los más "raros". Una ventaja de estos residuos externamente estudentizados es que si e_i es grande, la observación correspondiente aún destaca más a través del residuo rt_i.

Residuos parciales. Para una covariable x_j, estos residuos se obtienen cuando se prescinde de ella en el ajuste, y se calculan según:

$$e^*_{ij} = y_i - \sum_{k \neq j} \hat{\beta}_k \, x_{ik} = e_i + \hat{\beta}_j \, x_{ij}, \quad i = 1, \ldots, n, \tag{6.58}$$

Se utilizan, como veremos, para valorar la linealidad entre una covariable y la variable respuesta, en presencia de los restantes predictores.

Residuos de predicción PRESS. Estos residuos se utilizan para cuantificar el error de predicción, y se calculan a partir de la diferencia entre la respuesta observada y_i y la predicción que se obtiene ajustando el modelo propuesto sólo con las restantes $n-1$ observaciones, $\hat{y}_i^{(i)}$. Están relacionados con los residuos habituales según:

$$e_{(i)} = y_i - \hat{y}_i^{(i)} = \frac{e_i}{1 - h_{ii}}. \tag{6.59}$$

La varianza del residuo $e_{(i)}$ es:

$$Var(e_{(i)}) = \sigma^2/(1 - h_{ii}),$$

con lo que, la versión estandarizada del residuo de predicción utilizando como estimación de σ^2, s^2, coincide con el residuo estudentizado r_i en (6.55).

Estos residuos se utilizan en el análisis de influencia y en la validación del modelo.

Ejemplo 6.21 (Residuos en el ajuste para 'Bosque')**.** *Obtener la matriz gorro para el modelo de regresión lineal múltiple sobre los datos del Bosque. Calcular y representar los residuos comunes, estandarizados, los residuos de predicción y los estudentizados.*

```
# Nos aseguramos de tener cargados los datos del Apéndice.
fit<-lm(VOL~.,data=bosque,x=T)

# La matriz gorro completa se obtiene con:
x<-fit$x
hat<-x%*%solve(t(x)%*%x)%*%t(x); h
i<-diag(rep(1,length(VOL)))   # matriz identidad
# con la que se calculan los residuos comunes
(i-hat)%*%VOL
# obtenidos también directamente con:
residuals(fit)
e<-fit$res; e

# los residuos estandarizados, se consiguen de los
#  residuos comunes y la estima del error:
s<-summary(fit)$sigma
d<-e/s

# los residuos (inter.) estudentizados, con:
rstandard(fit)
# obtenidos de los residuos comunes, la matriz gorro
# y la estima del error:
h<-hatvalues(fit)  # es la diagonal de la matriz gorro
hat(fit$x)
r<-e/(s*sqrt(1-h));r

# los residuos ext. estudentizados son:
rt<-rstudent(fit);rt

# los residuos PRESS:
ei<-e/(1-h);ei

# Los pintamos todos:
par(mfrow=c(2,1))
plot(e,ylim=c(-7,7))
points(d,pch=2)
points(ei,pch=3)
legend(5,6.5,c("R.común","R.estandar","R.press"),pch=c(1,2,3),cex=0.8)

plot(r,ylim=c(-3,3))
points(rt,pch=2)
legend(5,3,c("R.int.stud","R.ext.stud"),pch=c(1,2),cex=0.8)
```

En la Figura 6.6 tenemos representados todos los residuos calculados, que a continuación utilizaremos en el diagnóstico, el análisis de influencia y la validación del modelo.

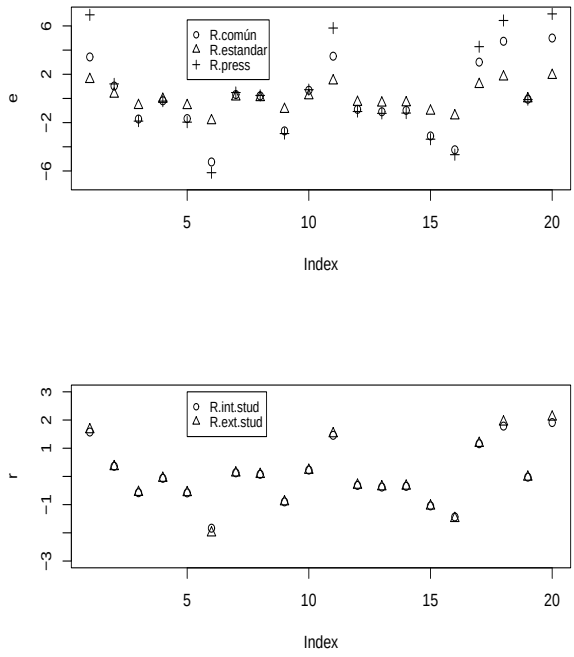


Figura 6.6: Residuos para el ajuste del modelo de regresión múltiple con 'Bosque'.

Una vez definidos los tipos básicos de residuos, procedemos al análisis de los mismos, con objeto de identificar violaciones de las hipótesis del modelo.

6.8.2. Linealidad

Una especificación deficiente del modelo de predicción, esto es, que no incluya como predictoras algunas variables que son útiles para explicar la respuesta, provoca estimaciones sesgadas. Si hay alguna variable explicativa que no ha sido incluida en el ajuste del modelo, representarla versus los residuos ayuda a identificar algún tipo de tendencia que dicha variable pueda explicar. Si se trata de una covariable, utilizaremos un gráfico de dispersión. Si se trata de un factor, diagramas de cajas, una por cada nivel de respuesta del factor. Si no se detecta ninguna tendencia en el gráfico de dispersión, o diferencia en las cajas, en principio no tenemos ninguna evidencia que nos sugiera incorporar dicha variable al modelo para predecir mejor la respuesta.

Los gráficos de residuos parciales sirven para valorar la linealidad entre una

covariable y la variable respuesta, en presencia de los restantes predictores. Estos gráficos fueron sugeridos por Ezekiel y Fox (1959) y Larsen y McCleary (1972). Están basados en los residuos parciales 6.58. Representando estos residuos versus x_{ij} para $i = 1, \ldots, n$, si efectivamente la variable x_j contribuye linealmente en la explicación de la respuesta $\mathbf{y}$, la tendencia que se apreciará en este gráfico será lineal. Estos gráficos son útiles también para detectar outliers y heterocedasticidad.

Ejemplo 6.22 (Verificación de linealidad para el ajuste de 'Bosque')**.** *Para el ajuste obtenido con los datos del bosque, construir los gráficos de regresión parcial y extraer conclusiones sobre la relación lineal entre regresores y respuesta.*

```
# Partimos del ajuste del modelo
fit<-lm(VOL~.,data=bosque)
library(faraway)
opar<-par(mfrow=c(2,2))
# Residuos parciales para DBH
prplot(fit,1)
# Residuos parciales para D16
prplot(fit,2)
# Residuos parciales para HT
prplot(fit,3)
par(opar)
```

Efectivamente, la tendencia que se aprecia en todos los gráficos de residuos parciales (Figura 6.7) es de tipo lineal, por lo que no tenemos indicios de problemas con la hipótesis de linealidad.

6.8.3. Homocedasticidad

La heterocedasticidad, que es como se denomina el problema de varianza no constante, aparece generalmente cuando el modelo está mal especificado, bien en la relación de la respuesta con los predictores, bien en la distribución de la respuesta, bien en ambas cuestiones.

• La violación de la hipótesis de varianza constante, $Var(\epsilon) = \sigma^2 I$, se detecta usualmente a través del análisis gráfico de los residuos:

- Gráficos de residuos versus valores ajustados $\hat{y}_i$.- Cuando aparece alguna tendencia como una forma de embudo o un abombamiento, etc., entonces

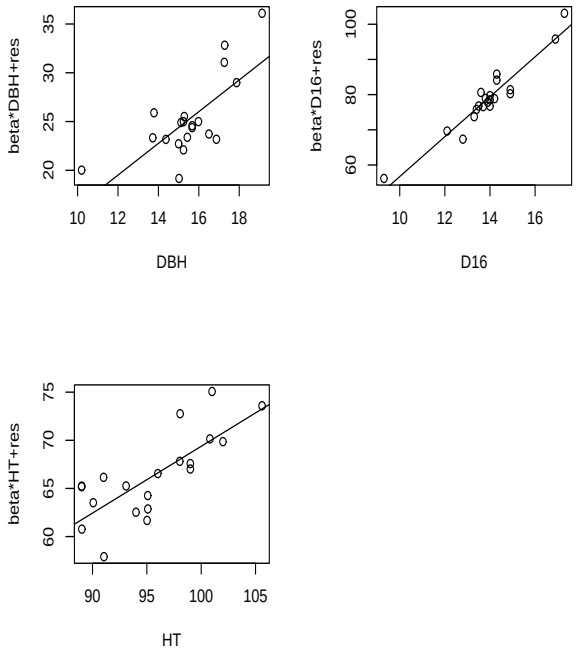


Figura 6.7: Gráficos de regresión parcial para el ajuste de bosque.

decimos que podemos tener algún problema con la violación de la hipótesis de varianza constante para los errores.

- Gráficos de residuos versus predictores x_j.- Básicamente se interpretan como los gráficos de residuos versus valores ajustados $\hat{y}_i$. Es deseable que los residuos aparezcan representados en una banda horizontal sin tendencias alrededor del cero.

En la Figura 6.8 tenemos diversas situaciones para estos tipos de gráficos. El gráfico (a), sin tendencias, es el deseable para corroborar homocedasticidad. En el gráfico (b) la variabilidad aumenta a medida que aumenta la magnitud de la predicción. En el gráfico (c) los valores medios de la predicción proporcionan más variabilidad que los extremos. Por último, el gráfico (d) da indicios de problemas con la linealidad.

Hay numerosos tests en la literatura para reconocer heterocedasticidad. Unos están basados en considerar la variabilidad de los residuos que consiguen explicar las variables explicativas sospechosas de inducir heterocedasticidad. Otros tests

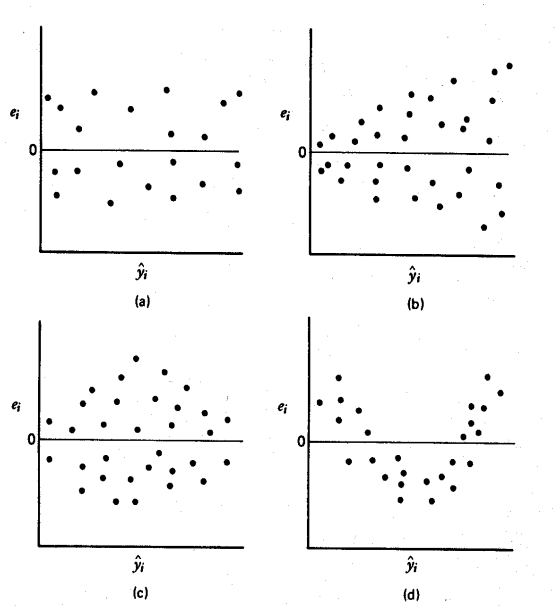


Figura 6.8: Gráficos de residuos versus valores ajustados.

están basados en diferenciar las observaciones en grupos de varianza constante y comparar los ajustes obtenidos respecto a la hipótesis de tener una misma varianza común.

• El test de Breusch-Pagan (Breusch y Pagan, 1979) ajusta un modelo de regresión lineal a los residuos del modelo ajustado, con las variables explicativas adicionales sospechosas de inducir varianza no constante, y rechaza si una buena parte de la varianza es explicada por dichas variables. El estadístico de contraste de Breusch-Pagan (bajo homocedasticidad) sigue una distribución chi-cuadrado con tantos grados de libertad como variables explicativas introducidas para justificar la falta de varianza constante.

• El test de Bartlett sigue la línea de dividir las observaciones en grupos supuestamente homogéneos, y contrasta si las varianzas en todos los grupos considerados son iguales. Para ello, ajusta un modelo por cada grupo y compara las varianzas con un estadístico, el K^2 de Bartlett, que sigue aproximadamente una distribución chi-cuadrado bajo H_0 (varianza constante).

Cuando evidenciamos heterocedasticidad, es importante identificar a qué es debida. En ocasiones, la variabilidad aumenta con el valor medio de la respuesta, en cuyo caso una transformación logarítmica resuelve generalmente el problema; en otras ocasiones está ligada a alguna variable explicativa x_j, de modo que la

heterocedasticidad se corrige dividiendo todas las variables por x_j, o lo que es lo mismo, utilizando una estimación por *mínimos cuadrados ponderados*, como veremos en la Sección 6.9.

Ejemplo 6.23 (Homocedasticidad en el ajuste para 'Insectos')**.** *Consideramos el Ejemplo 3.7, para el que predecimos el número de insectos supervivientes tras la vaporización con el spray, en función del insecticida utilizado. Investigar si se evidencia distinta variabilidad en los grupos definidos por el tipo de insecticida utilizado.*

```
# Cargamos los datos
data(InsectSprays)

# Ajustamos un modelo de ANOVA para predecir el número de insectos
# que quedan tras la vaporización con el insecticida:
fit<-lm(count~spray,data=InsectSprays)

# Investigamos primeramente con un análisis gráfico de residuos
# representando los residuos estudentizados:
r<-rstandard(fit)
opar<-par(mfrow=c(2,2))
# los residuos solos
plot(r,ylab='Residuos estudentizados')
title(sub="(a)")
# los residuos versus los valores ajustados
plot(fitted(fit),r,xlab='Valores ajustados',
ylab='Residuos estudentizados')
title(sub="(b)")
# y los residuos versus la variable predictora 'spray'
plot(r~spray,data=InsectSprays,xlab='Insecticida',
ylab='Residuos estudentizados')
title(sub="(c)")
par(opar)
```

En todos los gráficos de la Figura 6.9 se aprecia un problema de heterocedasticidad. Los residuos tienen más variabilidad para valores predichos más altos (gráfico (b)). Y en particular, con el gráfico (c), se identifica que la variabilidad parece ir asociada al tipo de insecticida utilizado: mayor variabilidad en los insecticidas A,B y F, y menor para los restantes, como ya se apreciaba en el gráfico (a), en el que las observaciones iban ordenadas por el tipo de insecticida pero no se distinguía el mismo .

Procedamos pues, a hacer la prueba estadística para homocedasticidad y veamos si conseguimos evidencias estadísticas para rechazarla:

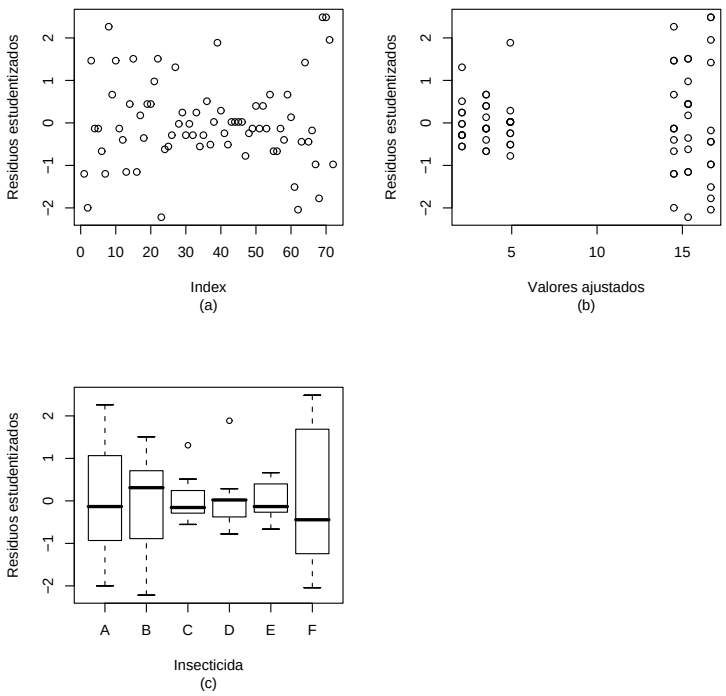


Figura 6.9: Diferente variabilidad ocasionada por el insecticida administrado. Datos InsectSprays.

```
library(lmtest)
# El test de Breusch-Pagan  se obtiene con
bptest(fit)

# El test de Bartlett se consigue con
bartlett.test(r ~ spray, data = InsectSprays)
```

Tanto el test de Breusch-Pagan, con un p-valor de 0.0009223, como el de Bartlett, con un p-valor de $9{,}085e-05$*, rechazamos homocedasticidad a favor de heterocedasticidad, esto es, varianza desigual. El gráfico nos proporcionaba evidencias de que la heterocedasticidad venía asociada a la variable explicativa que identifica el tipo de insecticida.*

6.8.4. Normalidad

La hipótesis de normalidad de los errores ϵ_i en el modelo lineal justifica la utilización de los tests F y t para realizar los contrastes habituales y obtener conclusiones confiables a cierto nivel de confianza $1-\alpha$ dado. En muestras pequeñas, la no normalidad de los errores es muy difícil de diagnosticar a través del análisis de los residuos, pues éstos pueden diferir notablemente de los errores aleatorios ϵ_i.

De hecho, la relación entre los residuos e_i y los errores aleatorios ϵ_i, viene dada por: por:

$$\mathbf{e} = (I-H)\mathbf{y} = (I-H)(X\beta+\epsilon) = (I-H)\epsilon,$$

$$\text{es decir,} \quad e_i = \epsilon_i - \left(\sum_{j=1}^{n} h_{ij}\,\epsilon_j\right). \tag{6.60}$$

En muestras grandes no se esperan demasiadas diferencias entre residuos y errores, y por lo tanto hacer un diagnóstico de normalidad sobre los residuos equivale prácticamente a hacerlo sobre los errores mismos. Esto es debido a que por el teorema central del límite, el término $e_i - \epsilon_i$, al ser una suma converge a una distribución normal, incluso aunque los ϵ_i no sigan tal distribución. Los términos h_{ii} tenderán a cero, y en consecuencia el término ϵ_i en (6.60) tenderá a dominar en la expresión para los residuos e_i.

La forma habitual de diagnosticar no normalidad es a través de los gráficos de normalidad y de tests como el de Shapiro-Wilks, específico para normalidad, o el de bondad de ajuste de Kolmogorov-Smirnov.

• Gráficos de normalidad para los residuos. Se dibujan los residuos ordenados $e_{[i]}$ (o los estandarizados/estudentizados) versus los cuantiles correspondientes de una normal estándar, $\Phi^{-1}[(i-1)/n]$. Si es cierta la normalidad de los residuos, el gráfico resultante debería corresponderse con una recta diagonal. En la Figura 6.10 sólo el gráfico (a) identifica una situación de normalidad de los residuos.

La hipótesis de normalidad se puede chequear también con histogramas de los residuos (estandarizados y estudentizados) cuando el tamaño muestral es grande. La normalidad se reconocerá a través de histogramas simétricos, con forma de campana de Gauss. Si los errores se distribuyen según una normal, entonces aproximadamente el 68 % de los residuos estandarizados (estudentizados) quedarán entre -1 y $+1$, y el 95 % entre -2 y $+2$.

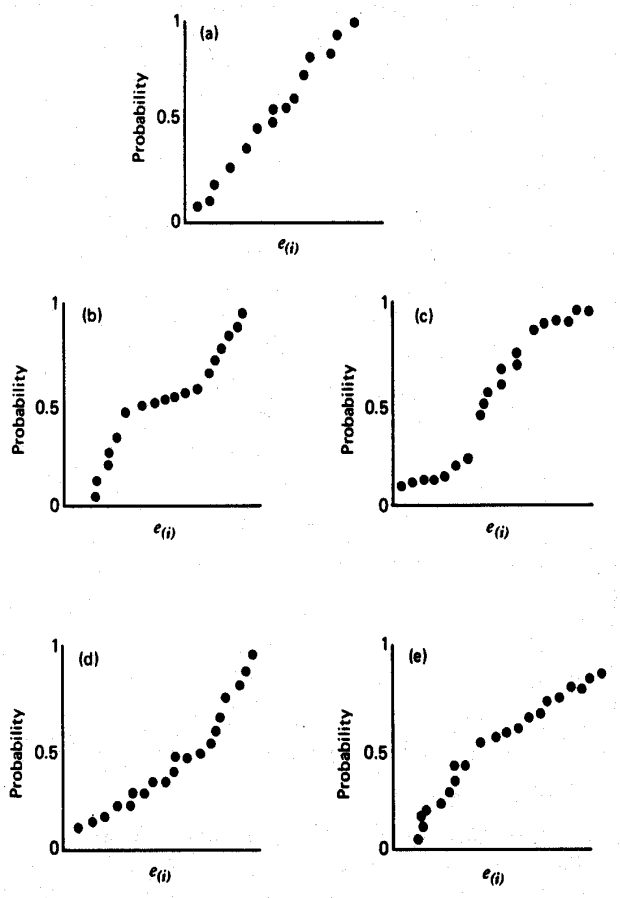


Figura 6.10: Gráficos qq-plot para verificar normalidad.

• El test de Shapiro-Wilks para normalidad se basa, más o menos, en una correlación entre los cuantiles empíricos de los residuos y los teóricos según una distribución normal. Cuanta mayor correlación, más indicios de normalidad para los residuos.

• El test de Kolmogorov-Smirnov es útil para contrastar cualquier distribución sobre un conjunto de datos. Calcula las diferencias entre la función de distribución empírica y la teórica en los puntos observados, toma valor absoluto y da como estadístico el máximo de estos valores. Es decir, valora la máxima distancia entre la distribución propuesta y la distribución empírica obtenida sin ninguna hipótesis paramétrica sólo a partir de los datos.

Por supuesto, la falta de normalidad puede darse cuando no hemos especificado todos los términos significativos para la explicación de la respuesta. Si detectamos una distribución de los residuos con más de una moda, es posible

que hayamos obviado del modelo algún factor de clasificación que diferencia el efecto de las covariables sobre la respuesta en diversos grupos.

Ejemplo 6.24 (Normalidad en el ajuste de 'Bosque'). *Chequear la normalidad de los residuos para el ajuste de regresión lineal múltiple obtenido para los datos del 'Bosque'.*

```
# Cargamos los datos del Apéndice y ajustamos el modelo
fit<-lm(VOL~.,data=bosque)

opar<-par(mfrow=c(1,2))
# Los gráficos de normalidad para los residuos son
qqnorm(fit$resid)
qqline(fit$resid)

# Podemos hacer también un histograma y superponer una densidad normal
# para los residuos estandarizados o estudentizados
r<-rstandard(fit)
hist(r,prob=T,xlim=c(-3,3),xlab="Res.estudentizados",main="Histograma")
lines(xseq<-seq(-3,3,length=100), dnorm(xseq,mean(r),sd(r)))
par(opar)

# Y podemos extraer información sobre el reparto de los residuos
# estudentizados por debajo de -2 y encima de 2
summary(rstandard(fit))
quantile(rstandard(fit),prob=c(0.025,0.975))

# Los tests formales de Shapiro-Wilks y Kolmogorov-Smirnov dan:
shapiro.test(r)     #p-valor=0.3922
ks.test(r,pnorm)    #p-valor=0.8606
```

En los gráficos (Figura D.4), observamos cierta desviación de normalidad en la cola de la derecha; los cuantiles empíricos son algo mayores a sus correspondientes teóricos, lo que quiere decir que los residuos tienen una cola derecha algo más pesada que la distribución normal. Esto mismo se puede apreciar en el histograma. Los cuantiles muestrales que encierran el 95 % de probabilidad resultan (-1.65,1.85), no especialmente desviados de los valores esperados para una normal ± 1.96. Los tests de normalidad no son significativos, esto es, la desviación apreciada no es significativa para rechazar normalidad.

6.8.5. Incorrelación

Para el modelo lineal general asumimos que los errores observacionales están incorrelados dos a dos. Si esta hipótesis no es cierta, cabe esperar que un gráfico

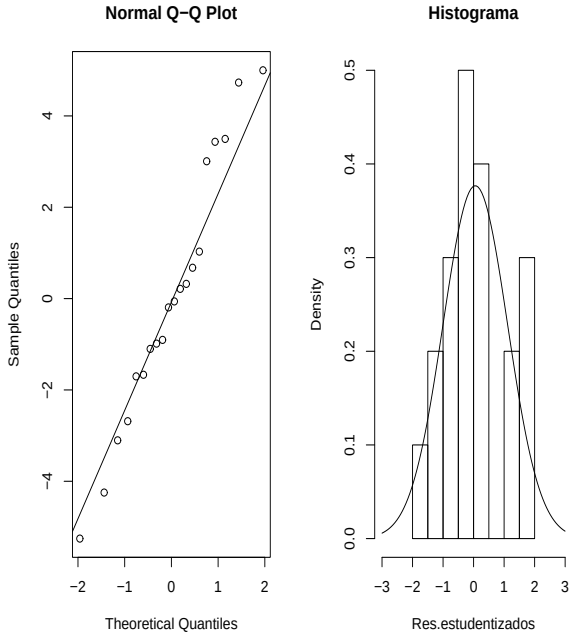


Figura 6.11: Gráficos para chequear normalidad de los residuos en el ajuste de 'Bosque'.

secuencial de los residuos manifieste alguna tendencia. Sin embargo, hay muchas formas en que los errores pueden estar correlados. De hecho, la independencia entre observaciones es una cuestión justificada básicamente por el muestreo realizado.

• Un gráfico de los residuos en función de la secuencia temporal en que se observaron los datos puede ayudar a apreciar un problema de correlación de los residuos (llamada *autocorrelación*), o de inestabilidad de la varianza a lo largo del tiempo. Detectar autocorrelación ha de conducir a considerar otro tipo de modelos distintos. En la Figura 5.8 tenemos dos gráficos de residuos correlacionados positiva y negativamente.

• Un tipo de correlación bastante habitual, es la *correlación serial*, consistente en que las correlaciones entre los errores que distan s posiciones, correlaciones que denotamos en adelante por ρ_s, son siempre las mismas.

$$\rho_s = Cov(r_i, r_{i+s}).$$

Si hay correlación serial positiva, los residuos tienden a ser consecutivos en la secuencia temporal. Si la correlación serial es negativa, un residuo positivo suele ser seguido de uno positivo y viceversa. Los **gráficos lag** ayudan a detectar

este tipo de correlación serial. Dichos gráficos consisten en representar cada residuo (excepto el primero) versus el residuo anterior en la secuencia temporal sospechosa de inducir la correlación.

En la Figura 5.8 se ilustraban las dos situaciones mencionadas de autocorrelación de los residuos, con los gráficos de secuencia temporal (izquierda) y los correspondientes gráficos lag (derecha).

• Un test habitual para detectar cierto tipo de correlación serial es el test de Durbin-Watson. Asumiendo normalidad, todas las correlaciones seriales entre los residuos han de ser cero, $\rho_s = 0$. El test de Durbin-Watson nos permite contrastar esto:

$$H_0 : \rho_s = 0, \quad \text{versus} \quad H_1 : \rho_s = \rho^s. \tag{6.61}$$

En caso de ser cierta H_1, dado que $|\rho| < 1$, los errores estarían relacionados de la forma:

$$\epsilon_i = \rho\epsilon_{i-1} + w_i,$$

donde todos los ϵ_i compartirían la misma media y varianza y $w_i \sim N(0, \sigma^2)$ sería independiente de $\epsilon_{i-1}, \epsilon_{i-2}, \ldots$ y de $w_{i-1}, w_{i-2}, \ldots$.

Aunque el test de Durbin-Watson es apropiado sólo para la alternativa específica (6.61), se utiliza de modo general para contrastar correlación de los errores, sin demasiada atención al tipo de alternativa; el coste de dicho abuso es la pérdida de potencia para otro tipo de alternativas.

Para resolver el test (6.61) se utiliza el *estadístico de Durbin-Watson*, definido por:

$$d = \frac{\sum_{i=2}^{n} (e_i - e_{i-1})^2}{\sum_{i=1}^{n} e_i^2}. \tag{6.62}$$

La distribución de d depende de los datos X, está definida entre 0 y 4 y es simétrica alrededor de 2. Los valores críticos por tanto, han de calcularse para cada problema concreto.

En el caso de detectar correlación serial, cabe estudiar la naturaleza de la correlación de los errores y buscar algún modelo válido para modelizar la dependencia temporal (análisis de series temporales). El planteamiento entonces es utilizar modelos de tendencias temporales para predecir la respuesta, puesto que ésta depende de la secuencia temporal en que se han observado los datos.

Ejemplo 6.25 (Incorrelación en el ajuste de 'Bosque'). *Testar gráfica y formalmente la incorrelación de los residuos en el ajuste de regresión múltiple para los datos de 'Bosque'.*

```
# Partimos del ajuste
fit<-lm(VOL~.,data=bosque)
# y hacemos los test sobre los residuos estudentizados
r<-rstandard(fit)
opar<-par(mfrow=c(1,2))
# El gráfico siguiente nos da el gráfico lag
n<-length(r)
plot(r[2:n],r[1:(n-1)],xlab=expression(r[i-1]),ylab=expression(r[i]))
# y éste permite identificar autocorrelación
plot(r,type="l",ylab="residuals")
par(opar)

# El test de Durbin-Watson:
library(lmtest)
dwtest(fit,alternative="two.sided")
```

En la Figura 6.12 no se aprecian indicios de correlación entre los residuos. De hecho, el test de Durbin Watson da un $p - valor = 0{,}1148$*, por lo que no podemos rechazar incorrelación de los residuos a una significatividad del 5 %, o lo que es lo mismo, no podemos afirmar que los datos evidencien correlación entre los residuos del modelo.*

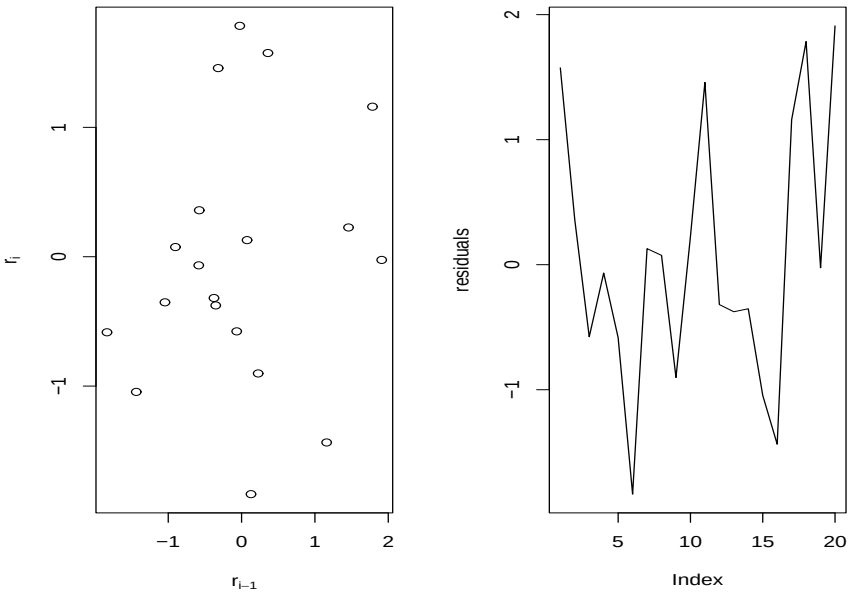


Figura 6.12: Gráficos de correlación de los residuos en el ajuste de los datos de 'Bosque'.

6.9. Soluciones a problemas detectados en el diagnóstico del modelo

6.9.1. Mínimos cuadrados generalizados

Los mínimos cuadrados ponderados son un procedimiento de estimación utilizado para corregir heterocedasticidad, siempre que se ha identificado su origen. Que un modelo padezca de heterocedasticidad significa que la matriz de varianzas-covarianzas para los errores aleatorios es diagonal, con no todos sus elementos iguales:

$$Var(\epsilon) = E(\epsilon\epsilon') = G = diag\{\sigma_1^2, \ldots, \sigma_n^2\}.$$

En ese caso, la log-verosimilitud para los coeficientes del modelo lineal $\mathbf{y} = X\beta + \epsilon$ tendría la forma:

$$L(\beta) = -\frac{1}{2}(\mathbf{y} - X\beta)'G^{-1}(\mathbf{y} - X\beta). \tag{6.63}$$

Así, maximizarla respecto de β equivale a minimizar:

$$(\mathbf{y} - X\beta)'G^{-1}(\mathbf{y} - X\beta), \tag{6.64}$$

esto es,

$$\sum_{i=1}^{n} (y_i - X_i'\beta)^2/\sigma_i^2 = \sum_{i=1}^{n} e_i^2/\sigma_i^2,$$

con X_i la i-ésima fila de la matriz de diseño X.

La estimación $\hat{\beta}$ depende pues, de las cantidades desconocidas σ_i^2, que han de ser estimadas de algún modo para permitir la estimación de β. Para ello es necesario identificar la fuente de heterocedasticidad. Así, podría ocurrir que la varianza fuera constante en diferentes tramos de respuesta, o en diferentes tramos de algún predictor, o que evolucionara según algún funcional de y, $\sigma^2 = f(\mathsf{y})$, o de algún regresor, $\sigma^2 = f(\mathrm{x})$, etc.

Si por ejemplo resulta que las varianzas son constantes por tramos de la respuesta, esto es, que los residuos pueden dividirse en S tramos de varianza constante $\sigma_1^2, \sigma_2^2, \ldots, \sigma_S^2$, con $n_1, n_2, \ldots, n_S$ observaciones cada uno, se pueden utilizar ciertas estimaciones de los $\{\sigma_s^2,\ s = 1, 2, \ldots, S\}$ basadas en los residuos iniciales del modelo ajustado e_i, reagrupados en los tramos considerados, $\{e_{sj},\ j = 1, \ldots, n_s\}$ para $s = 1, 2, \ldots, S$, según:

$$\hat{\sigma}_s^2 = \sum_{j=1}^{n_s} e_{sj}^2/n_s, \qquad s = 1, 2, \ldots, S.$$

Cabría entonces considerar un ajuste por mínimos cuadrados ponderados con los pesos $1/\hat{\sigma}_s^2,\ h = 1, \ldots, S$ correspondientes a cada observación.

Ejemplo 6.26 (Mínimos cuadrados ponderados para el ajuste de 'Insectos' (Ejemplo 3.7))**.** *Sobre los datos 'InsectSprays' de R, obtener estimaciones para las varianzas en cada tipo de insecticida y reajustar un modelo de mínimos cuadrados ponderados.*

Ya en la Figura 3.7 apreciábamos diferente variabilidad en los resultados del número de supervivientes tras aplicar los diversos tipos de insecticida. Cuantificaremos pues dicha variabilidad en cada grupo y la utilizaremos para corregir en el modelo de Anova a ajustar:

```
# Cargamos los datos
data(InsectSprays)
attach(InsectSprays)
# Calculamos los residuos del modelo inicial sin corrección:
res<-residuals(lm(count~spray,data=InsectSprays))
# y verifiquemos con Bartlett los problemas de heterocedasticidad:
bartlett.test(rstandard(lm(count~spray)) ~ spray, data = InsectSprays)
# p-valor=9.085e-05

#  la estimación de la varianza por grupos:
sigma2.s<-unlist(lapply(split(res^2,spray),mean))
# y definimos los pesos:
pesos<-1/rep(sigma2.s,tabulate(spray))

# Ya ajustamos el modelo por mcp:
fit<-lm(count~spray,weights=pesos)
# y dibujamos los residuos para apreciar si hemos corregido
# el problema de heterocedasticidad:
r<-rstandard(fit)

opar<-par(mfrow=c(2,2))
# los residuos solos
plot(r,ylab='Res.estudentizados')
# los residuos versus los valores ajustados
plot(fitted(fit),r,xlab='Valores ajustados',
ylab='Res.estudentizados')
# y los residuos versus la variable predictora 'spray'
plot(r~spray,xlab='Insecticida',ylab='Res.estudentizados')
par(opar)
detach(InsectSprays)
```

Como se aprecia en la Figura 6.13, las diferencias de variabilidad para los residuos entre los tipos de insecticida se han reducido substancialmente. Verifiquemos con el test de Bartlett si hay evidencias en contra de la homocedasticidad:

```
bartlett.test(r ~ spray, data = InsectSprays)
```

que con un p-valor de 1 no permite rechazar homocedasticidad. Luego el problema de heterocedasticidad en el ajuste inicial (patente con el p-valor$= 9{,}085e-05$ *en el test de Bartlett) queda corregido con los mínimos cuadrados ponderados.*

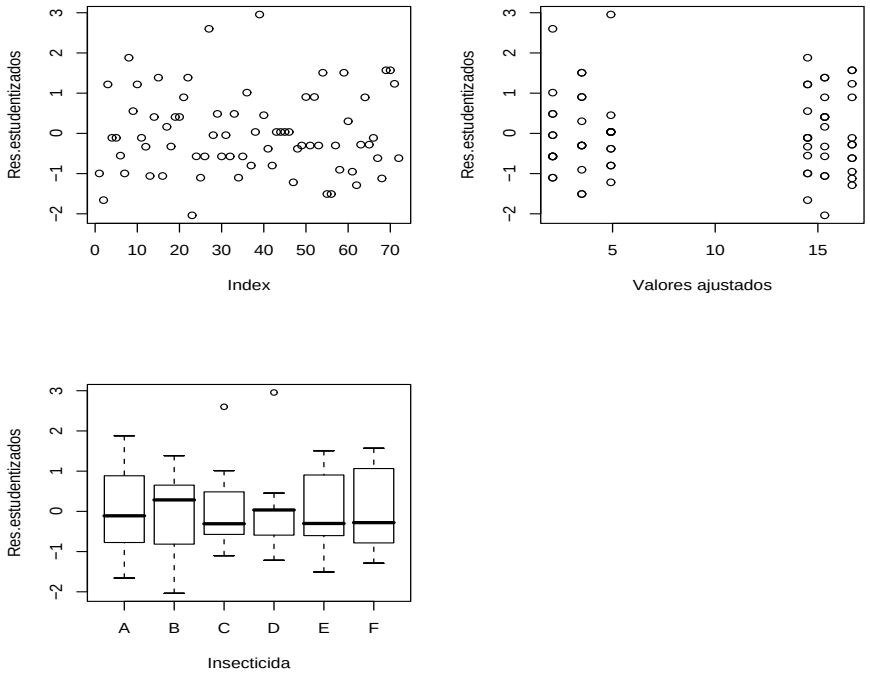


Figura 6.13: Residuos para la estimación por mínimos cuadrados ponderados. Datos InsectSprays.

En general, ante determinados problemas con la violación de hipótesis en el modelo lineal general, se pueden plantear soluciones en términos de:

1. Propuesta de otros modelos adecuados a la distribución de la respuesta y su relación con los predictores. Como una prolongación del modelo lineal general tenemos el *modelo lineal generalizado* (ver Mayoral y Morales, 2001).

2. Transformar la variable respuesta.

3. Transformar las covariables continuas.

6.9.2. Transformaciones de la variable respuesta

Puede ser un remedio para resolver problemas de no linealidad, no normalidad, o de varianza no constante. En ocasiones, incluso alguna transformación de la respuesta permite simplificar el modelo de predicción de un conjunto de datos.

Las desventajas de transformar la respuesta son que no siempre la misma transformación resuelve todos los problemas detectados, y por regla general, si resuelve alguno, ocasiona otros. Sin embargo, a veces es posible llegar a cierto compromiso con una transformación aceptable, sugerida generalmente por los gráficos del ajuste, el análisis de los residuos, y la propia experiencia del analista. En todo caso, siempre hay que tener en cuenta, a la hora de proponer una transformación, la naturaleza de los errores, es decir, cómo se producen en relación a la respuesta y a los predictores, pues toda transformación sobre la respuesta es una transformación sobre ellos.

Si bien el problema de encontrar una transformación aceptable para la respuesta puede llegar a ser tedioso, existe algún método que permite obtener, de una forma rápida, una transformación que proporciona ciertos beneficios, sin más que asumir que dicha transformación pertenece a la *familia de transformaciones potenciales*. Las tranformaciones de Box-Cox (Box y Cox, 1964) pertenecen a dicha familia y consiste en considerar:

$$z_{\lambda,i} = \begin{cases} \dfrac{y_i^{\lambda} - 1}{\lambda\, \dot{\mathbf{y}}^{\lambda-1}} & \text{si } \lambda \neq 0, \\ \dot{\mathbf{y}}\, ln(y_i) & \text{si } \lambda = 0, \end{cases} \tag{6.65}$$

donde $\dot{\mathbf{y}} = (y_1 y_2 \dots y_n)^{1/n}$ es la media geométrica de las observaciones.

Surge así un nuevo parámetro λ a estimar para obtener una transformación adecuada. La idea básica de esta transformación es encontrar un λ tal que dé lugar a un modelo lineal aditivo de la forma:

$$\mathbf{z}_{\lambda} = X\beta + \epsilon, \tag{6.66}$$

con ϵ errores normales, independientes y con varianza constante, estimable por máxima verosimilitud.

Para estimar el mejor valor de λ, básicamente se utiliza un grid de valores (en general es suficiente un grid entre -2 y 2), para el que se calcula la suma de cuadrados de la regresión SSR, o lo que es lo mismo, la log-verosimilitud en la estimación máximo-verosímil cuando se ajusta el modelo siguiendo la transformación de Box-Cox. El valor de λ óptimo es aquel para el que el modelo explica

el máximo de variabilidad de la respuesta, esto es, proporciona el valor máximo para SSR. Suele ser aconsejable utilizar un valor de λ redondeado (por ejemplo, si $\hat{\lambda} = 0{,}1$, conviene usar $\hat{\lambda} = 0$ en su lugar), por facilitar la interpretación de la transformación.

La transformación de Box-Cox es aplicable sólo cuando la respuesta es estrictamente positiva. Si hay valores cero o negativos, la forma habitual de proceder es añadiendo una constante a la respuesta antes de aplicar el método, con el fin de hacer todos los valores positivos; sin embargo, suele haber poca información en los datos para elegir dicha constante.

Ejemplo 6.27 (Box-Cox para el ajuste de Bosque2)**.** *Utilizando los datos presentados en el Ejemplo 3.3, ajusta un modelo de regresión, realiza el diagnóstico gráfico, identifica los problemas que padece y propón una transformación de Box-Cox para corregirlos. Verifica si efectivamente tales problemas se resuelven. Recuerda que el objetivo en este problema era el de predecir el volumen de madera de un árbol en función de su altura y de su circunferencia a cierta altura.*

```
# Cargamos los datos en R:
data(trees)
# y ajustamos el modelo de regresión múltiple
fit<-lm(Volume ~ Height + Girth, data = trees)
summary(fit)
# R2=0.948 y p-valor< 2.2e-16

# Hacemos el diagnóstico gráfico para detectar problemas
opar<-par(mfrow=c(2,2))
plot(fit)
par(opar)
# y un test de normalidad
shapiro.test(rstandard(fit))
# p-valor=0.6389
```

El ajuste resultante es significativo (p-valor$< 2{,}2\,10^{-16}$*) y el coeficiente de determinación habla de un ajuste razonablemente bueno* $R^2 = 0{,}95$*. Sin embargo, en el diagnóstico gráfico del modelo (Figura 6.14), apreciamos cierta tendencia curvilínea en el gráfico de residuos versus valores ajustados (valores extremos producen residuos positivos y valores centrales, residuos negativos). Para corregirla, planteamos utilizar una transformación de Box-Cox.*

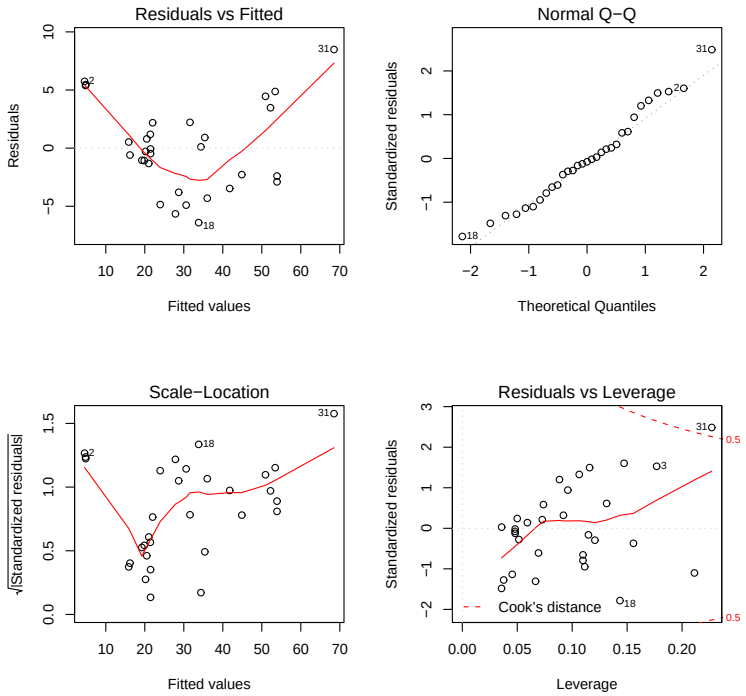


Figura 6.14: Diagnóstico gráfico del modelo ajustado para Bosque2 ('trees' en $\mathcal{R}$).

```
# Buscamos el valor de lambda para la transformación:
library(MASS)
par(mfrow=c(1,1))
boxcox(fit)
# que pinta el valor óptimo para lambda y un
# intervalo de confianza al 95%:
bc<-boxcox(fit,plotit=F)
# el valor de lambda que maximiza SSR es
lambda<-bc$x[which.max(bc$y)]; lambda #=0.3

# Luego reajustamos el modelo con la variable Volume transformada:
library(labstatR)  # librería que calcula la media geométrica
attach(trees)
# la variable transformada por Box-Cox es:
z<-(Volume^lambda-1)/(lambda*mean.g(Volume)^(lambda-1))
# y el nuevo ajuste con dicha variable
fit.bc<-lm(z~Height+Girth,data=trees)
# cuyo diagnóstico gráfico resultante es:
opar<-par(mfrow=c(2,2))
plot(fit.bc)
par(opar)
```

En el gráfico de residuos versus valores ajustados (arriba izquierda en la Figura 6.15) apreciamos que ahora los residuos ya están repartidos uniformemente alrededor del cero, lo que implica que el problema detectado anteriormente ha desaparecido. La transformación de Box-Cox con $\lambda = 0{,}3$ ha dado buenos resultados, sin afectar (aparentemente) la violación de otras hipótesis del modelo.

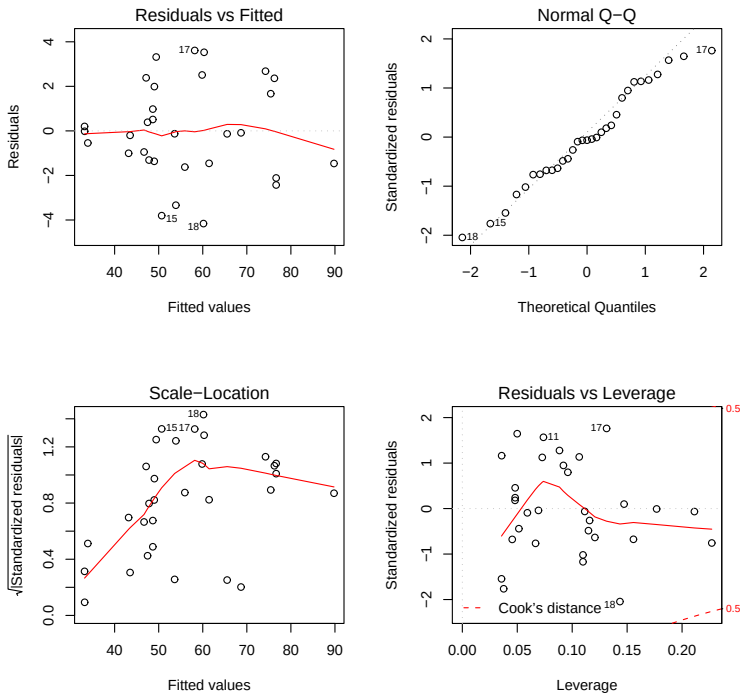


Figura 6.15: Diagnóstico gráfico del modelo ajustado para Bosque2 ('trees' en $\mathcal{R}$) con la transformación de Box-Cox.

6.9.3. Transformaciones de las covariables

Si bien las transformaciones sobre la respuesta afectan a la estructura del error aleatorio (lo que puede acarrear desajustes respecto de otras hipótesis ya verificadas en el modelo para la variable original), no ocurre así con las transformaciones de las covariables, aunque éstas tienen una utilidad notablemente inferior a las primeras. La transformación de los predictores tiene sentido para capturar tendencias no lineales entre la respuesta y éstos. Los gráficos de los residuos ayudan a decidir qué tipo de transformaciones conviene considerar.

Hay ciertos procedimientos analíticos, del tipo de las transformaciones potenciales de Box-Cox, para encontrar transformaciones adecuada sobre un predictor. Con todo, suele ser la práctica la que guía sobre propuestas que puedan mejorar el ajuste. En ocasiones, de hecho, basta con la inclusión de términos polinómicos para capturar la tendencia no explicada.

Weisberg (1985), pag.142, proporciona una tabla resumen que incorporamos aquí (Tabla 6.4), y que ilustra muy bien distintas transformaciones útiles para linealizar la relación entre la respuesta y los predictores en regresión simple y múltiple. A veces, al corregir linealidad se corrige simultaneamente la falta de normalidad, aunque no siempre es así. De hecho, la mayoría de las veces, al corregir un problema del modelo con una determinada transformación, queda al descubierto otro que conllevaría otra transformación. La mejor opción entonces suele ser la utilización de los modelos lineales generalizados, que permiten especificar separadamente la distribución de los datos, la relación entre la respuesta media y la varianza, y la relación de linealidad con los predictores (ver Mayoral y Morales, 2001).

Transformación	Regresión Simple	Regresión Múltiple
$log(y) \sim log(X)$	$y = \alpha \mathbf{x}^{\beta}$	$y = \alpha \mathbf{x}_1^{\beta_1} \dots \mathbf{x}_k^{\beta_k}$
$log(y) \sim X$	$y = \alpha\, exp(\mathbf{x}\beta)$	$y = \alpha\, exp\{\sum \mathbf{x}_j \beta_j\}$
$y \sim log(X)$	$y = \alpha + \beta log(\mathbf{x})$	$y = \alpha + \sum \beta_j log(\mathbf{x}_j)$
$1/y \sim 1/X$	$y = \frac{\mathbf{x}}{\alpha \mathbf{x} + \beta}$	$y = \frac{1}{\alpha + \sum \beta_j / \mathbf{x}_j}$
$1/y \sim X$	$y = \frac{1}{\alpha + \beta \mathbf{x}}$	$y = \frac{1}{\alpha + \sum \beta_j \mathbf{x}_j}$
$y \sim 1/X$	$y = \alpha + \beta \frac{1}{\mathbf{x}}$	$y = \alpha + \sum \beta_j \frac{1}{\mathbf{x}_j}$

Tabla 6.4: Transformaciones de las covariables para linealizar la relación respuesta-predictores.

Ejemplo 6.28 (Transformación de covariables en el ajuste de 'Tractores' (Ejemplo 3.1))**.** *En la Figura 5.5 teníamos dibujados a la izquierda el ajuste transformando la covariable 'costes' con el logaritmo, y a la derecha el obtenido con los datos en su escala original. Claramente el logaritmo corregía linealidad y mejoraba el ajuste.*

6.10. Análisis de influencia

En ocasiones hay algún subconjunto de los datos que influencia desproporcionadamente el ajuste del modelo propuesto, con lo cual las estimaciones y

predicciones dependen mucho de él. Es interesante siempre, localizar este tipo de datos, si existen, y evaluar su impacto en el modelo. Si estos datos influyentes son "malos" (provienen de errores en la medición, o de condiciones de experimentación diferentes, etc.) habrían de ser excluidos del ajuste; si son "buenos", esto es, efectivamente proceden de buenas mediciones aunque raras, contendrán información sobre ciertas características relevantes a considerar en el ajuste. En todo caso, es importante localizarlos, y para ello existen una serie de procedimientos basados en diversos estadísticos que presentamos a continuación.

Hay diversos criterios para valorar la influencia de las observaciones en el ajuste, y en base a los cuales se proponen diversos estadísticos. Vamos a considerar tres de ellos: i) contribución a la estimación de los coeficientes; ii) influencia en la predicción y iii) influencia sobre la precisión de las estimaciones.

6.10.1. Influencia sobre los coeficientes del modelo

Se han construido diversas medidas para valorar la influencia de las observaciones en la estimación de los coeficientes del modelo. Entre ellas, las más habituales son:

Distancia de Cook. Cook (1977, 1979) sugirió la utilización de una medida de influencia para una observación y_i, basada en la distancia entre la estimación de mínimos cuadrados obtenida con las n observaciones, $\hat{\mathbf{y}} = X\hat{\beta}$, y la obtenida eliminando dicha observación, $\hat{\mathbf{y}}^{(i)}$. Una formulación habitual del *estadístico de Cook* es:

$$D_i = \frac{(\hat{\mathbf{y}} - \hat{\mathbf{y}}^{(i)})'(\hat{\mathbf{y}} - \hat{\mathbf{y}}^{(i)})}{ps^2} = \frac{(\hat{\beta}^{(i)} - \hat{\beta})'X'X(\hat{\beta}^{(i)} - \hat{\beta})}{ps^2}, \quad i = 1, \ldots, n, \tag{6.67}$$

donde $\hat{\beta}^{(i)}$ es el vector de parámetros estimados en la regresión $\hat{\mathbf{y}}^{(i)}$.

Los puntos con un valor grande del estadístico D_i identifican observaciones tales que el hecho de incluirlas o no en el ajuste dan lugar a diferencias considerables en las estimaciones de los coeficientes. La magnitud de D_i se puede evaluar comparándola con el cuantil $F_{p,n-p,\alpha}$. Si $D_i \approx F_{p,n-p,0,5}$, eliminar la observación i del ajuste movería $\hat{\beta}$ a la frontera de una región de confianza al 50 % construida con todos los datos (idealmente, se espera que $\hat{\beta}^{(i)}$ quede dentro de una región con confianza entre el 10 y el 20 %). Como aproximadamente $F_{p,n-p,0,5} \approx 1$, generalmente se consideran como influyentes aquellas observaciones con un valor del estadístico $D_i > 1$.

El estadístico de Cook se relaciona con los residuos mediante:

$$D_i = \frac{e_i^2\, h_{ii}}{ps^2(1-h_{ii})^2} = \frac{r_i^2\, h_{ii}}{p(1-h_{ii})}, \quad i = 1, \ldots, n. \tag{6.68}$$

El cociente $h_{ii}/(1-h_{ii})$ representa la distancia del vector $(1, x_{i1}, \ldots, x_{i,p-1})$ al centroide del resto de los datos, es decir, de algún modo mide la distancia entre un punto y el resto de los datos. La otra parte de D_i, r_i^2/p, mide lo bien que el modelo ajusta la i-ésima observación.

DFBETAS. Belsley, Kuh y Welsch (1980) sugieren un estadístico que indica cuánto cambia el coeficiente $\hat{\beta}_j$ en desviaciones estándar, si se excluye la i-ésima observación:

$$DFBETAS_{j,i} = \frac{\hat{\beta}_j - \hat{\beta}_j^{(i)}}{s_{(i)}^2 C_{jj}^X}, \quad j = 0, 1, \ldots, p;\ i = 1, \ldots, n \tag{6.69}$$

donde $\hat{\beta}_j^{(i)}$ es la j-ésima componente del vector $\hat{\beta}_{(i)}$, y C_{jj}^X es el elemento j de la diagonal de $(X'X)^{-1}$.

Para interpretar $DFBETAS_{j,i}$ conviene considerar su expresión en términos de la matriz $W = (X'X)^{-1}X'$, con elementos w_{ji} y cuyas filas $\mathbf{w}_j'$ contienen la influencia que las n observaciones tienen en la estimación $\hat{\beta}_j$,

$$DFBETAS_{j,i} = \frac{w_{ji}}{\sqrt{\mathbf{w}_j'\mathbf{w}_j}}\, \frac{e_i}{s_{(i)}(1-h_{ii})} = \frac{w_{ji}}{\sqrt{\mathbf{w}_j'\mathbf{w}_j}}\, \frac{r_i}{\sqrt{1-h_{ii}}}, \tag{6.70}$$

donde r_i es el residuo estudentizado (6.55). Así, $DFBETAS$ mide, tanto el impacto de la i-ésima observación sobre la estimación $\hat{\beta}_j$, con $w_{ji}/\sqrt{\mathbf{w}_j'\mathbf{w}_j}$, como el efecto de un residuo grande, dado por $r_i/\sqrt{1-h_{ii}}$. Los autores sugieren como punto de corte para las $DFBETAS$, $2/\sqrt{n}$, esto es, si $|DFBETAS_{j,i}| > 2/\sqrt{n}$, entonces la observación i-ésima se considera potencialmente influyente.

6.10.2. Influencia sobre las predicciones

Para investigar la influencia de la i-ésima observación sobre los valores predichos por el modelo disponemos de los estadísticos DFFITS y PRESS:

DFFITS. Se define el estadístico DFFITS para la observación i-ésima como:

$$DFFITS_i = \frac{\hat{y}_i - \hat{y}_i^{(i)}}{\sqrt{s_{(i)}^2 h_{ii}}}, \quad i = 1, \ldots, n, \tag{6.71}$$

donde $\hat{y}_i^{(i)}$ es el valor predicho para y_i por el modelo sin utilizar en la estimación la observación i. Así, $DFFITS_i$ se puede interpretar como el número de desviaciones estándar que cambia la predicción de la i-ésima respuesta cuando dicha observación es excluida del ajuste.

$DFFITS_i$ se puede calcular a través de los residuos habituales e_i (6.52), o bien de los residuos externamente estudentizados rt_i (6.56), según:

$$DFFITS_i = \sqrt{\frac{h_{ii}}{1-h_{ii}}}\,\frac{e_i}{s_{(i)}\sqrt{h_{ii}}} = \sqrt{\frac{h_{ii}}{1-h_{ii}}}\,rt_i.$$

Ya comentamos que, utilizando la distribución de los residuos estudentizados (tanto internos como externos), podemos identificar observaciones "raras", esto es, con residuos especialmente grandes en valor absoluto. En términos del estadístico $DFFITS$, la identificación de estos puntos depende, tanto de la influencia de una observación, medida por h_{ii}, como del error de estimación, cuantificado por el residuo. Generalmente, una observación para la que $|DFFITS_i| > 2\sqrt{p/n}$ merece ser tratada con atención.

PRESS. El error de predicción PRESS está basado en los residuos de predicción PRESS (6.59). Con ellos, es posible estimar un error de predicción del modelo a través del estadístico $PRESS$ (Allen, 1971, 1974) con:

$$PRESS = \sum_{i=1}^{n} e_{(i)}^2 = \sum_{i=1}^{n} \left(\frac{e_i}{1-h_{ii}}\right)^2. \tag{6.72}$$

Así, las observaciones para las que h_{ii} sea grande, tendrán unos residuos de predicción grandes y contribuirán más al error de predicción evaluado según el $PRESS$. Utilizando exclusivamente los residuos de predicción (6.59), un dato se cataloga como influyente si existe una diferencia grande entre el residuo ordinario e_i y el residuo de predicción $e_{(i)}$, pues se entiende entonces que la predicción que da el modelo sin ese punto resulta "pobre".

6.10.3. Influencia sobre la precisión de las estimaciones

Los diagnósticos vistos hasta ahora cuantifican de algún modo el efecto de las observaciones en las estimaciones. Sin embargo, no proporcionan información alguna sobre la precisión conjunta del ajuste. En ese contexto, es común utilizar el determinante de la matriz de covarianzas como una medida (escalar) de precisión. Este determinante se denomina varianza generalizada. Se define la varianza

generalizada para la estimación $\hat{\beta}$, como:

$$VG(\hat{\beta}) = |Var(\hat{\beta})| = \sigma^2 \, |(X'X)^{-1}|.$$

Así, la influencia de la i-ésima observación sobre la precisión de la estimación de $\hat{\beta}$ se puede medir en función del estadístico COVRATIO:

$$COVRATIO_i = \frac{s^2_{(i)}|(X'_{(i)}X_{(i)})^{-1}|}{s^2_{(i)}|(X'X)^{-1}|}, \quad i = 1, \ldots, n. \tag{6.73}$$

Si $COVRATIO_i < 1$, excluir la i-ésima observación proporciona un ajuste más preciso; si $COVRATIO_i > 1$, la i-ésima observación mejora la precisión de la estimación. Este estadístico se puede calcular también de la forma:

$$COVRATIO_i = \left(\frac{s^2_{(i)}}{s^2}\right)^p \left(\frac{1}{1-h_{ii}}\right),$$

con $s^2_{(i)}$ definido en (6.57). Así, un punto muy influyente respecto de h_{ii} hará el $COVRATIO$ grande, lo cual es lógico. De hecho, un punto influyente mejora siempre la precisión del ajuste, a no ser que se trate de un outlier (una respuesta y_i remota). Sólo si dicha observación es un outlier, el cociente $s^2_{(i)}/s^2$ será mucho menor que 1.

Los puntos de corte para detectar observaciones influyentes a través del $COVRATIO$ no son triviales de obtener, y se consideran más apropiados para muestras grandes. Belsley, Kuh y Welsch (1980) sugieren considerar como influyente un punto para el que $COVRATIO_i > 1+3p/n$ o $COVRATIO_i < 1-3p/n$ (este último, apropiado sólo cuando $n > 3p$).

Todas estas herramientas presentadas para la detección de observaciones influyentes resultarán más o menos útiles dependiendo del análisis concreto que se esté resolviendo. Los puntos de corte dados para los estadísticos de influencia propuestos son meramente indicativos. Finalmente es el analista el que ha de decidir, para cada problema, cuándo considera influyente una observación en base a lo que pretende y a lo que obtiene.

Ejemplo 6.29 (Análisis de influencia en el ajuste de 'Bosque')**.** *Sobre los datos de Bosque (Ejemplo 3.2) y para el modelo de regresión múltiple obtenido, realizar un análisis completo de influencia.*

```
# Nos aseguramos de tener los datos cargados
# A partir del ajuste obtenido prescindiendo de DBH:
fit<-lm(VOL~D16+HT,data=bosque)
# calculamos las medidas de influencia
im<-influence.measures(fit); im
# que proporciona los DFBETAS,DFFITS, COVRATIO y Distancia de Cook,
# indicando si cada punto es influyente (columna 'inf').

# Además, identifica los puntos influyentes con un asterisco:
im$is.inf
summary(im)

# Los valores de los estadísticos los podemos representar para apreciar
# mejor los puntos influyentes:
n<-length(VOL)
opar<-par(mfrow=c(2,3))
estadistico<-dimnames(im$infmat)[[2]]
nstat<-dim(im$infmat)[2]

# Los DFBETA:
for(i in 1:3){
dotchart(im$infmat[,i],main=estadistico[i],xlim=c(-1,1))
# con una línea para el valor crítico 2/sqrt(n):
abline(v=2/sqrt(n),lty=2)
abline(v=-2/sqrt(n),lty=2)}

# El DFFIT:
i<-4
dotchart(im$infmat[,i],main=estadistico[i])
# con valores influyentes en rojo:
puntos.i<-which(im$is.inf[,i]==T)
vpuntos.i<-im$infmat[puntos.i,i]
points(vpuntos.i,puntos.i,pch=21,bg="red",col="red")
# y valores críticos:
p<-length(fit$coef)     # número de coeficientes del modelo
abline(v=2*sqrt(p/n),lty=2)
abline(v=-2*sqrt(p/n),lty=2)

# El COVRATIO:
i<-5
dotchart(im$infmat[,i],main=estadistico[i])
# con valores influyentes en rojo:
puntos.i<-which(im$is.inf[,i]==T)
vpuntos.i<-im$infmat[puntos.i,i]
points(vpuntos.i,puntos.i,pch=21,bg="red",col="red")
# y valores críticos:
abline(v=1+3*p/n,lty=2)
abline(v=1-3*p/n,lty=2)
```

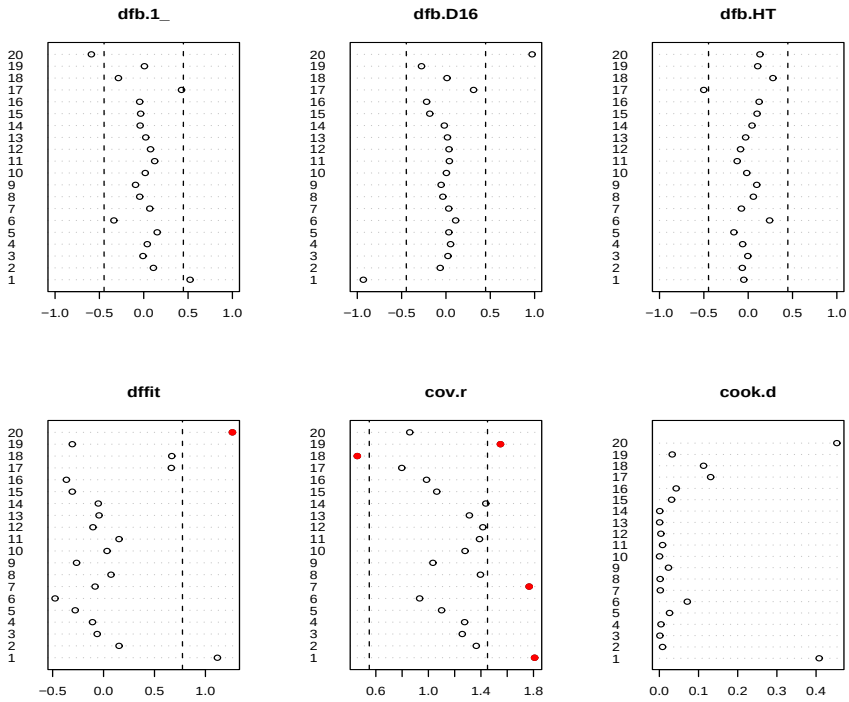


Figura 6.16: Gráficos de influencia para el ajuste de Bosque.

```
# La distancia de Cook
i<-6
dotchart(im$infmat[,i],main=estadistico[i])
# con valores críticos:
abline(v=qf(0.5,p,n-p),lty=2)
par(opar)
```

En la Figura 6.16 hemos representado los valores para los estadísticos de influencia presentados, con líneas verticales indicando los valores críticos apuntados, que difieren respecto de los que utiliza $\mathcal{R}$ para identificar puntos influyentes (pintados con un punto relleno en los gráficos). Como apreciamos en dicho gráfico, así como en la Tabla 6.5, respecto de DFFIT sólo destaca como influyente la observación 20; respecto del error PRESS, sólo destaca la observación 1, con un valor grande de h_{11}; respecto del COVRATIO, destacan como influyentes las observaciones 1,7,18 y 19; por último, no detectamos puntos influyentes a través de los DFBETAS.

	dfb.1	dfb.D16	dfb.HT	dffit	cov.r	cook.d	hat
1	0.52	−0.93	−0.05	1.12	1.81 ∗	0.41	0.48 ∗
7	0.07	0.03	−0.08	−0.08	1.77 ∗	0.00	0.32
18	−0.29	0.01	0.28	0.67	0.46 ∗	0.11	0.06
19	0.01	−0.28	0.11	−0.31	1.55 ∗	0.03	0.26
20	−0.59	0.97	0.13	1.26 ∗	0.86	0.45	0.29

Tabla 6.5: Resumen de puntos influyentes detectados por $\mathcal{R}$ en el ajuste de Bosque.

6.11. Validación del modelo: validación cruzada

Una vez ajustado un modelo cabe comprobar su validez predictiva, esto es, si es o no, especialmente sensible a la aleatoriedad de la muestra con la que se realiza el ajuste. Esto es lo que se denomina *validación del modelo*. La validación cruzada consiste en realizar el ajuste con una parte de los datos y compararla con la que se obtiene con otra parte de los datos. Si se obtienen diferencias severas, concluiremos con que el ajuste no es robusto y por lo tanto su validez predictiva queda en cuestión. Para realizar la validación cruzada disponemos de dos procedimientos básicos: i) separar la muestra en dos trozos, ajustar el modelo con uno de ellos y utilizar el otro para validar; y ii) realizar todos los ajustes quitando cada una de las observaciones y calcular el error cuadrático de validación para concluir sobre la validez del modelo. Veámoslos.

• Un procedimiento riguroso de validación del modelo es verificar sus resultados con una muestra independiente de la utilizada para construirlo. Cuando no es posible muestrear más, se puede considerar el ajuste con una parte de los datos y dejar los restantes para la validación del mismo. Por supuesto, este procedimiento resta fiabilidad a las estimaciones, ya que al obtenerse con menos datos producen errores estándar mayores.

Si por ejemplo se ha ajustado el modelo con una muestra de n_1 observaciones denotada por M_1, $\mathbf{y}_1 = X_1\beta_1 + \epsilon_1$, cabe plantear el ajuste con una muestra independiente M_2 con n_2 observaciones, $\mathbf{y}_2 = X_2\beta_2 + \epsilon_2$ y a continuación resolver el contraste $H_0 : \beta_1 = \beta_2$ frente a $H_1 : \beta_1 \neq \beta_2$.

Dicho contraste se resuelve, cuando $n_2 > p$, con p el número de coeficientes del modelo, con el estadístico F basado en las sumas de cuadrados residuales del modelo completo, SSE_T, y de los parciales para el conjunto de observaciones $\mathbf{y}_1$,

SSE_1, y para $\mathbf{y}_2$, SSE_2, definido por:

$$\frac{(SSE_T - SSE_1 - SSE_2)/p}{(SSE_1 + SSE_2)/(n_1 + n_2 - 2p)} \sim F_{(p,n_1+n_2-2p)}, \tag{6.74}$$

y cuando $n_2 < p$ (en cuyo caso $SSE_2 = 0$), con:

$$\frac{(SSE_T - SSE_1)/n_2}{SSE_1/(n_1 - p)} \sim F_{(n_2,n_1-p)}, \tag{6.75}$$

El contraste (6.74) es más potente para detectar errores de especificación asociados a sesgos en los parámetros β debidos a la omisión de alguna variable, o a errores en los datos. El contraste (6.75) es más potente para detectar cambios fundamentales en la especificación del modelo, es decir que la muestra M_2 requiere de un modelo distinto más complejo (con otras variables a las utilizadas con M_1).

Ejemplo 6.30 (Validación con dos muestras para el ajuste de 'Bosque')**.** *Llevar a cabo la validación del modelo propuesto con predictores D16 y HT partiendo la muestra en dos trozos y aplicando el test F correspondiente.*

```
# Cargamos los datos del Apéndice
# Queremos particionar el banco de datos en dos trozos
# de tamaños similares. Seleccionamos pues, dos muestras
# al azar:
n<-length(VOL)
n1<-round(n/2)
n2<-n-n1
muestra1<-sample(1:20,n1)
muestra2<-(1:n)[-muestra1]

s1<-bosque[muestra1,]
s2<-bosque[muestra2,]

# Y ajustamos los modelos correspondientes:
formula<-VOL~D16+HT
fit1<-lm(formula,data=s1)
fit2<-lm(formula,data=s2)
# para compararlos con el ajuste global
fit<-lm(formula, data=bosque)
```

```
# y seleccionamos las sumas de cuadrados del error:
p<-length(fit1$coefficients)   # número de coeficientes estim
sse1<-anova(fit1)[[2]][p]
sse2<-anova(fit2)[[2]][p]
sset<-anova(fit)[[2]][p]

# para calcular el estadístico F:
f<-((sset-sse1-sse2)/p)/((sse1+sse2)/(n-2*p))
# y el p-valor correspondiente
1-pf(f,p,n-2*p)

# Y pintamos los ajustes
d16<-seq(min(D16),max(D16),length=20)
ht<-seq(min(HT),max(HT),length=20)
newdata<-data.frame(D16=d16,HT=ht)

par(mfrow=c(1,2))
plot(d16,predict(fit,newdata),type="l",lwd=2,xlab="D16",ylab="VOL")
# superponemos el ajuste f1
lines(d16,predict(fit1,newdata),lty=2,lwd=2,col="red")
# y el ajuste f2
lines(d16,predict(fit2,newdata),lty=3,lwd=2,col="blue")

plot(ht,predict(fit,newdata),type="l",xlab="HT",ylab="VOL")
lines(ht,predict(fit1,newdata),lty=2,lwd=2,col="red")
lines(ht,predict(fit2,newdata),lty=3,lwd=2,col="blue")

legend(90,95,c("Ajuste Global","Ajuste M1","Ajuste M2"),lty=1:3,
lwd=2,col=c("black","red","blue"))
```

Hemos particionado los datos en dos trozos aleatoriamente. El test F en la partición que nos ha surgido dio un valor para el estadístico $F = 1{,}07$ con un p-valor= 0,393, lo cual no nos permite rechazar la igualdad de coeficientes en los dos ajustes obtenidos de las particiones. En principio no tenemos evidencias en contra de la robustez del modelo. En la Figura 6.17 se aprecia la variación en los ajustes obtenidos para las 2 particiones consideradas.

• Otro procedimiento interesante para juzgar la robustez de un modelo es la validación cruzada una a una, y consiste en, para cada observación i, ajustar el modelo con las $n-1$ observaciones restantes y con él predecir la respuesta i, $\hat{y}_i^{(i)}$. Se define el error cuadrático de validación como:

$$EC_v = \sum_{i=1}^{n} (y_i - \hat{y}_i^{(i)})^2. \tag{6.76}$$

Algunos autores (Allen, 1971, Stone, 1974) propusieron elegir los estimadores de β que minimizaran la expresión (6.76), procedimiento que requiere de los

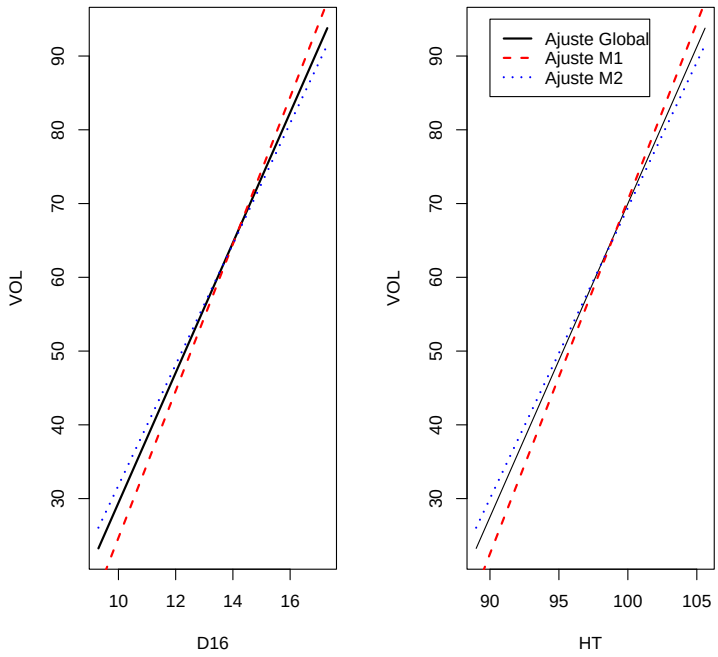


Figura 6.17: Validación del modelo de regresión múltiple sobre 'Bosque' particionando los datos en dos grupos.

mínimos cuadrados ponderados y que da lugar a estimadores poco robustos, pues las observaciones más influyentes (alejadas del resto) son las que más peso tienen en la estimación de los parámetros. No hay un criterio unificado; de hecho, Peña (1993) desaconseja el estadístico EC como criterio para estimar el modelo.

Otra medida de robustez la proporciona el coeficiente de robustez:

$$B^2 = \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \hat{y}_i^{(i)})^2}, \tag{6.77}$$

que da un valor entre 0 y 1 y cuantifica la robustez del modelo. Cuando las predicciones $\hat{y}_i^{(i)}$ sean próximas a las del ajuste con todos los datos, $\hat{y}_i$, el valor de B^2 será próximo a 1, y si hay mucha diferencia, será casi cero.

Ejemplo 6.31 (Validación cruzada uno a uno en el ajuste de 'Bosque')**.** *Calcular el error cuadrático de validación y el coeficiente de robustez para el ajuste de 'Bosque' con los predictores D16 y HT, y concluir sobre la validez del ajuste.*

```
# Partimos del ajuste inicial para acceder a la matriz
# de diseño X:
fit<-lm(VOL~D16+HT,data=bosque,x=T)
x<-fit$x
# y calculamos las predicciones yi(i) sin la observación i:
n<-length(VOL)  # número de datos
y.i<-vector(length=n)
for(i in 1:n)
y.i[i]<-x[i,]%*%coef(lm(VOL~D16+HT,data=bosque[-i,]))
# para calcular el error cuadrático de validación
ecv<-sum((VOL-y.i)^2)   #=251

# y el coeficiente de robustez
b2<-sum(residuals(fit)^2)/ecv;b2 #= 0.707
```

Es más fácil interpretar el valor obtenido para el coeficiente de robustez, muy próximo a 1 (0.707), que valorar la magnitud del error cuadrático de validación (ecv=251). La conclusión es similar a la que obteníamos con el test F obtenido de dividir la muestra en dos trozos: el ajuste lo podemos catalogar como robusto de cara a que no se evidencian variaciones relevantes entre los ajustes obtenidos con todos los datos y otros conseguidos con sólo una parte de ellos. En principio, el modelo ajustado tiene una validez predictiva respetable.

6.12. Ejercicios

Descripción de bancos de datos y problemas a resolver

1. Parece ser que el dinero gastado en la manutención de tractores es mayor a medida que aumenta la edad del tractor. Se pretende ratificar esta hipótesis.

2. Para estimar la cantidad de madera de un bosque se suele realizar un muestreo en el que se toman una serie de medidas no destructivas con las que predecir el volumen de madera generado. Disponemos de las correspondientes mediciones para 20 árboles, así como los volúmenes de madera que originan una vez cortados. Las variables consideradas son:

   HT o altura en pies
   DBH el diámetro del tronco a 4 piés de altura (en pulgadas)
   D16 el diámetro del tronco a 16 pies de altura (en pulgadas)
   VOL el volumen de madera conseguida (en pies cúbicos).

3. Queremos estudiar la relación existente entre la concentración de madera contenida en la pulpa con la que se elabora papel y la resistencia (en términos de tensión que soporta) del papel resultante.

4. Se lleva a cabo un experimento clínico en el que un grupo aleatorizado de pacientes, identificado como 'grupo T', prueba un nuevo fármaco contra la hipertensión, y otro grupo también aleatorizado, llamado 'grupo P', toma un placebo. A todos los pacientes se les mide la tensión arterial antes de administrar el tratamiento (fármaco o placebo) y después de finalizar el período de observación. El objetivo del análisis es investigar si el fármaco tiene un efecto sobre la reducción de la tensión arterial y cuantificar dicho efecto si existe.

5. Se desea estudiar el tiempo de vida de piezas cortadoras de dos tipos, A y B, en función de la velocidad del torno en el que está integrada (en revoluciones por segundo). El objetivo del análisis es describir la relación entre el tiempo de vida de la pieza y la velocidad del torno, teniendo en cuenta de qué tipo es la pieza.

6. Se ha llevado a cabo un experimento para estudiar la concentración presente de un fármaco en el hígado después de sufrir un tratamiento. Se piensa que las variables que pueden influir en la concentración son el peso del cuerpo, el peso del hígado y la dosis de fármaco administrada.

7. Durante los últimos años se ha venido realizando un experimento para investigar qué factores pueden afectar al desarrollo intelectual de los individuos. Para realizar este experimento se han venido siguiendo pares de gemelos que han sido separados y criados por separado: uno con padres adoptivos y el otro en una casa de acogida. Se dispone además de información sobre la clase social de los padres adoptivos. Para valorar el desarrollo intelectual de los individuos se les pasa un test de inteligencia a ambos gemelos y se anota la puntuación obtenida.

8. Se trata de determinar el precio de las viviendas (en miles de dólares) en función de las características de dichas viviendas. La información recogida es: impuestos (en miles de dólares), número de baños, tamaño de la parcela (en miles de pies cuadrados), tamaño de la vivienda (en miles de pies cuadrados), número de plazas de garaje, número de habitaciones, edad de la vivienda (en años), número de bocas de incendios.

9. Las elecciones presidenciales en Francia se realizan mediante un proceso de dos vueltas. En la primera vuelta se presentan todo los candidatos. La segunda vuelta vuelta se realiza con los dos candidatos que obtuvieron más votos en la primera vuelta. El interés de muchos analistas políticos radica en estudiar la transferencia de votos entre los candidatos eliminados en la primera vuelta y los resultados definitivos de la segunda vuelta. Los datos disponibles corresponden a las elecciones de 1981, en las que inicialmente había 10 candidatos. Se dispone de los datos para los diferentes distritos electorales y se presentan como:

   ditrito: distrito electoral
   EI: electores del distrito (en miles)
   A: votantes de Mitterand en la primera vuelta (en miles)
   B: votantes de Giscard en la primera vuelta (en miles)
   C: votantes de Chirac en la primera vuelta (en miles)
   D: votantes del partido comunista en la primera vuelta (en miles)
   E: votantes del partido verde en la primera vuelta (en miles)
   F: votantes del partido F en la primera vuelta (en miles)
   G: votantes del partido G en la primera vuelta (en miles)
   H: votantes del partido H en la primera vuelta (en miles)
   J: votantes del partido J en la primera vuelta (en miles)
   K: votantes del partido K en la primera vuelta (en miles)
   A2: votantes de Mitterand en la segunda vuelta (en miles)
   B2: votantes de Giscard en la segunda vuelta (en miles)

10. Se ha realizado un estudio para determinar las características que influyen en la calidad de los vinos de la variedad Pino Noir. Para ello se ha recogido la siguiente información para las diferentes clases bajo estudio: claridad, aroma, cuerpo, olor, matiz, calidad y región de procedencia.

11. Se trata de estudiar el calor desarrollado en calorías por gramo (y) en la obtención de cemento a partir de los compuestos que forman la mezcla. Dichos compuestos son tricalcium aluminate (x_1), triclacium silicate (x_2), tetracalcium alumino ferrite (x_3) y dicalcium silicate (x_4).

12. Se realiza un experimento, en concreto un test, para cuantificar la cantidad de energía termal (en kw) emitida por el sol (total.heat.flux) como función del grado de insolación (en w/m^2) (insolation), posición del punto focal situado en la dirección este (focal.pt.east), posición del punto focal situado en la dirección sur (focal.pt.south), posición del punto focal situado en la dirección norte (focal.pt.north), y la hora del día a la que se realiza el test (time.of.day).

13. Un periodista deportivo americano dedicado a la NFL trata de obtener un modelo (lineal) que le permita determinar cuáles son la características que determinan que un equipo gane más partidos que otro. Para ello se recoge la siguiente información: partidos ganados (y), yardas corridas (x_1), yardas de pase (x_2), promedio de pateo (x_3), porcentaje de puntos conseguidos (x_4), diferencia entre balones recuperados y perdidos (x_5), yardas de penalización (x_6), porcentaje de tiempo que los corredores están en el campo (x_7), yardas corridas por los oponentes (x_8) y yardas de pase de los oponentes (x_9).

14. Se esta realizando un estudio para determinar que marca de automóviles tiene un menor consumo por kilómetro recorrido en función de sus características. La información recogida es consumo (y), desplazamiento (x_1), potencia (x_2), torque (x_3), ratio de compresión (x_4), ratio del eje trasero (x_5), carburadores (x_6), número de marchas (x_7), longitud (x_8), anchura (x_9), peso (x_{10}) y tipo de transmisión (x_{11}, A=automática, M=Manual)

15. Se pretende estudiar el rendimiento de un proceso químico basado en la fabricación de CO_2 (y) como función de diferentes variables del proceso bajo control. Dichas variables son: tiempo en minutos del proceso (x_1), temperatura a la que se realiza el proceso (x_2), porcentaje de solubilidad (x_3), cantidad de aceite usada (x_4), total de carbón usado (x_5), cantidad de disolvente usada (x_6), y consumo de hidrógeno (x_7).

16. Se desea estudiar la concentración de $NbOCL3$ en el tubo de flujo de un reactor como función de diferentes características controlables. Dichas ca-

racterísticas son la concentración de $COCL2$ (x_1), el tiempo transcurrido (x_2), la densidad molar (x_3) y la fracción molar de CO_2 (x_4).

17. En un estudio medio ambiental sobre la diversidad de especies de tortuga en las Islas Galápagos se recoge información sobre el número de especies de tortuga (*Species*) encontradas en cada isla, así como del número de especies endémicas (*Endemics*), el área de la isla (*Area*), la altura del pico más alto de la isla (*Elevation*), la distancia a la isla más cercana ($Nearest$), la distancia a la isla de Santa Cruz ($Scruz$) y el área de la isla más próxima (*Adjacent*).

18. Se trata de estudiar el índice de mortalidad infantil (*mortality*) de diferentes países en función de su renta per cápita (*income*), el continente al que pertenecen (*region*), y el hecho de ser o no un país exportador de petróleo (*oil*).

Apéndice A

Algebra Matricial

A.1. Introducción

Este apéndice ofrece un resumen de las definiciones y propiedades básicas del álgebra matricial aparecidas en el desarrollo de algunas demostraciones del Tema 6.

Definición A.1. *Una matriz* $\boldsymbol{A}$ *es un conjunto rectangular de números. Si* $\boldsymbol{A}$ *tiene n filas y p columnas diremos que es de orden* $n \times p$*. Por ejemplo, n observaciones correspondientes a p variables aleatorias son representadas matemáticamente a través de una matriz.*

Notaremos a la matriz $\mathbf{A}$ de orden $n \times p$ como

$$\mathbf{A} = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1p} \\ a_{21} & a_{22} & \dots & a_{2p} \\ . & . & . & . \\ a_{n1} & a_{n2} & \dots & a_{np} \end{pmatrix}$$

donde a_{ij} es el elemento en la fila i-ésima y la columna j-ésima de la matriz $\mathbf{A}$, $i = 1, \dots, n$; $j = 1, \dots, p$.

Definición A.2. *La traspuesta de una matriz A se calcula intercambiando las filas por las columnas:*

$$\boldsymbol{A}' = \begin{pmatrix} a_{11} & a_{21} & \dots & a_{n1} \\ a_{12} & a_{22} & \dots & a_{n2} \\ . & . & . & . \\ a_{1p} & a_{2p} & \dots & a_{np} \end{pmatrix}$$

Definición A.3. *A una matriz columna de orden 1 la denominaremos vector columna. En consecuencia,*

$$\boldsymbol{a} = \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{pmatrix}$$

es un vector columna con n componentes.

En general, un vector fila aparecerá en este libro como un vector columna traspuesto, es decir,

$$\mathbf{a}' = (a_1, \dots, a_n).$$

Notaremos las columnas de la matriz $\mathbf{A}$ como $\mathbf{a}_{(1)}, \mathbf{a}_{(2)}, \dots, \mathbf{a}_{(p)}$ y las filas como $\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_n$ de forma que

$$\mathbf{A} = (\mathbf{a}_{(1)}, \dots, \mathbf{a}_{(p)}) = \begin{pmatrix} \mathbf{a}_1' \\ \mathbf{a}_2' \\ \vdots \\ \mathbf{a}_n' \end{pmatrix}$$

Definición A.4. *A una matriz escrita en términos de sus sub-matrices la llamaremos matriz particionada.*

Sean, entonces, $\mathbf{A}_{11}, \mathbf{A}_{12}, \mathbf{A}_{21}$ y $\mathbf{A}_{22}$ sub-matrices tales que

$$\mathbf{A} = \begin{pmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{pmatrix}$$

A.2. Operaciones Matriciales

La siguiente tabla nos ofrece un resumen de varios tipos de operaciones que pueden ser aplicadas sobre matrices.

Operación	Restricciones	Definición	Observación
Suma	$\mathbf{A}$, $\mathbf{B}$ del mismo orden	$\mathbf{A}+\mathbf{B} = (a_{ij} + b_{ij})$	
Resta	$\mathbf{A}$, $\mathbf{B}$ del mismo orden	$\mathbf{A}$-$\mathbf{B} = (a_{ij} - b_{ij})$	
Producto por escalar		$\mathrm{c}\mathbf{A} = (ca_{ij})$	
Producto escalar	$\mathbf{a}$, $\mathbf{b}$ del mismo orden	$\mathbf{a}'\mathbf{b} = \sum a_i b_i$	
Producto	$\mathbf{A}(n \times p)$, $\mathbf{B}(p \times s)$	$\mathbf{AB}=(\mathbf{a}_i'\mathbf{b}_{(j)})$	$\mathbf{AB} \neq \mathbf{BA}$

A.2.1. Traspuesta

La traspuesta satisface una serie de propiedades:

- $(\mathbf{A}')' = \mathbf{A}$.
- $(\mathbf{A} + \mathbf{B})' = \mathbf{A}' + \mathbf{B}'$.
- $(\mathbf{AB})' = \mathbf{B}'\mathbf{A}'$.
- Si $\mathbf{A}$ es una matriz simétrica entonces $\mathbf{A}' = \mathbf{A}$.

A.2.2. Determinante

Definición A.5. *El determinante de una matriz cuadrada $\boldsymbol{A}$ se define como*

$$|\boldsymbol{A}| = \sum (-1)^{|\tau|} a_{1\tau(1)} \dots a_{p\tau(p)},$$

donde el sumatorio es tomado sobre todas las permutaciones τ de $(1, 2, \dots, p)$, y $|\tau|$ es igual a $+1$ o -1, dependiendo de la paridad de la suma de los índices de filas y columnas que aparecen en el producto.

Para $p = 2$,

$$|\mathbf{A}| = a_{11}a_{22} - a_{12}a_{21}.$$

Definición A.6. *Una matriz cuadrada es no-singular si* $|\boldsymbol{A}| \neq 0$*; en otro caso diremos que* $\boldsymbol{A}$ *es singular.*

A.2.3. Inversa

Antes de introducir el concepto de operación inversa de una matriz cuadrada, definiremos lo que se conoce como *rango* de una matriz. Ambos conceptos se encuentran estrechamente relacionado, como mostraremos posteriormente.

Definición A.7. *El rango de una matriz* $\boldsymbol{A}(n \times p)$ *se define como el número máximo de filas (columnas) de* $\boldsymbol{A}$ *linealmente independientes.*

Denotamos por $rg(\mathbf{A})$ al rango de la matriz $\mathbf{A}$.

Se cumplen las siguientes propiedades:

1. $0 \leq rg(\mathbf{A}) \leq min(n, p)$.
2. $rg(\mathbf{A}) = rg(\mathbf{A}')$.
3. $rg(\mathbf{A} + \mathbf{B}) \leq rg(\mathbf{A}) + rg(\mathbf{B})$.
4. $rg(\mathbf{AB}) \leq min\{rg(\mathbf{A}), rg(\mathbf{B})\}$.
5. $rg(\mathbf{A}'\mathbf{A}) = rg(\mathbf{AA}') = rg(\mathbf{A})$.

Definición A.8. *La inversa de una matriz cuadrada* $\boldsymbol{A}$*,* $\boldsymbol{A}^{-1}$*,es la única matriz que satisface*

$$\boldsymbol{A}\boldsymbol{A}^{-1} = \boldsymbol{A}^{-1}\boldsymbol{A} = \boldsymbol{I}.$$

La inversa existe si y sólo si $\mathbf{A}$ es no-singular, es decir, si y sólo si $|\mathbf{A}| \neq 0$.

Las siguientes propiedades se verifican con la operación inversa:

- $(c\mathbf{A})^{-1} = c^{-1}\mathbf{A}^{-1}$.
- $(\mathbf{AB})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1}$.
- La única solución de $\mathbf{Ax} = \mathbf{b}$ es $\mathbf{x} = \mathbf{A}^{-1}\mathbf{b}$.

Además, si $n = p$ entonces $rg(\mathbf{A}) = p$ si y sólo si existe $\mathbf{A}^{-1}$ ($\mathbf{A}$ no-singular).

A.2.4. Inversa Generalizada

Mostraremos en esta sección un método que nos permitirá definir una inversa para cualquier matriz.

Definición A.9. *Para una matriz $\boldsymbol{A}(n \times p)$, $\boldsymbol{A}^-$ representará a la matriz g-inversa (inversa generalizada) de $\boldsymbol{A}$ si*

$$\boldsymbol{A}\boldsymbol{A}^-\boldsymbol{A} = \boldsymbol{A}.$$

La inversa generalizada siempre existe, aunque en general no es única.

A.3. Independencia, Producto Escalar, Norma y Ecuaciones Lineales

A.3.1. Independencia lineal entre vectores, Producto Escalar y Norma

Definición A.10. *Diremos que los vectores $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k$ son linealmente dependientes si existen números $\lambda_1, \ldots, \lambda_k$ no todos cero, tales que*

$$\lambda_1\boldsymbol{x}_1 + \cdots + \lambda_k\boldsymbol{x}_k = \boldsymbol{0}.$$

En cualquier otro caso los k vectores son linealmente independientes.

Definición A.11. *El producto escalar entre dos vectores $\boldsymbol{x}, \boldsymbol{y} \in \Re^n$ se define como*

$$\boldsymbol{x} \cdot \boldsymbol{y} = \boldsymbol{x}'\boldsymbol{y} = \sum_{i=1}^{n} x_i y_i.$$

Diremos que los vectores $\mathbf{x}$ e $\mathbf{y}$ son ortogonales si $\mathbf{x} \cdot \mathbf{y} = 0$.

Definición A.12. *La norma de un vector* $\boldsymbol{x} \in \Re^n$ *viene dada por*

$$||\boldsymbol{x}|| = (\boldsymbol{x} \cdot \boldsymbol{x}^{1/2}) = (\sum x_i^2)^{1/2}.$$

Entonces la distancia entre dos vectores $\mathbf{x}$ e $\mathbf{y}$ viene dada por

$$||\mathbf{x} - \mathbf{y}||.$$

A.3.2. Ecuaciones Lineales

Para las n ecuaciones lineales

$$x_1\mathbf{a}_{(1)} + \cdots + x_p\mathbf{a}_{(p)} = \mathbf{b}$$

o, equivalentemente,

$$\mathbf{A}\mathbf{x} = \mathbf{b}$$

con la matriz de coeficientes $\mathbf{A}(n \times p)$, tenemos los siguientes resultados:

1. Si $n = p$ y $\mathbf{A}$ no-singular, la única solución al sistema de ecuaciones lineales es

$$\mathbf{x} = \mathbf{A}^{-1}\mathbf{b}.$$

2. El sistema es consistente (es decir, admite al menos una solución) si y sólo si

$$rg(\mathbf{A}) = rg[(\mathbf{A}\ \mathbf{b})],$$

donde la matriz $(\mathbf{A}\ \mathbf{b})$ representa a la matriz $\mathbf{A}$ a la que se le añade el vector de términos independientes del sistema de ecuaciones, $\mathbf{b}$.

3. Si $\mathbf{b} = \mathbf{0}$, existe una solución no trivial (es decir, $\mathbf{x} \neq \mathbf{0}$) si y sólo si $rg(\mathbf{A}) < p$.

4. La ecuación $(\mathbf{A}'\mathbf{A})\mathbf{x} = \mathbf{A}'\mathbf{b}$ es siempre consistente.

A.4. Valores y Vectores Propios

Si $\mathbf{A}(p \times p)$ es cualquier matriz entonces

$$q(\lambda) = |\mathbf{A} - \lambda\mathbf{I}| \tag{A.1}$$

es un polinomio en λ de orden p. Las p raíces de $q(\lambda)$, $\lambda_1, \ldots, \lambda_p$, son los denominados valores propios de $\mathbf{A}$.

Para cada $i = 1, \ldots, p$, $|\mathbf{A} - \lambda_i\mathbf{I}| = 0$, así $\mathbf{A} - \lambda_i\mathbf{I}$ es una matriz singular. En consecuencia, existe un vector no nulo γ satisfaciendo

$$\mathbf{A}\gamma = \lambda_i\gamma. \tag{A.2}$$

Llamaremos vector propio de la matriz $\mathbf{A}$ asociado al valor propio λ_i a cualquier vector γ que satisface (A.2).

Una propiedad importante nos dice que el determinante de la matriz $\mathbf{A}$ coincide con el producto de los valores propios de $\mathbf{A}$, es decir,

$$|\mathbf{A}| = \lambda_1\lambda_2 \ldots \lambda_p.$$

A.5. Diferenciación Matricial y Problemas de Minimización

Definimos la derivada de la función $f(\mathbf{X})$ con respecto a $\mathbf{X}(n \times p)$ como la matriz

$$\frac{\partial f(\mathbf{X})}{\partial \mathbf{X}} = \left(\frac{\partial f(\mathbf{X})}{\partial x_{ij}}\right).$$

Presentamos a continuación una serie de resultados interesantes:

1. $\frac{\partial \mathbf{a}'\mathbf{x}}{\partial \mathbf{x}} = \mathbf{a}$.
2. $\frac{\partial \mathbf{x}'\mathbf{x}}{\partial \mathbf{x}} = 2\mathbf{x}$, $\frac{\partial \mathbf{x}'\mathbf{A}\mathbf{x}}{\partial \mathbf{x}} = (\mathbf{A} + \mathbf{A}')\mathbf{x}$, $\frac{\partial \mathbf{x}'\mathbf{A}\mathbf{y}}{\partial \mathbf{x}} = \mathbf{A}\mathbf{y}$.

En cuanto a la aplicación de los anteriores resultados sobre problemas de optimización, presentamos lo siguiente:

Teorema A.1. *El vector* $\boldsymbol{x}$ *que minimiza*

$$f(\boldsymbol{x}) = (\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x})'(\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x})$$

debe cumplir

$$\boldsymbol{A}'\boldsymbol{A}\boldsymbol{x} = \boldsymbol{A}'\boldsymbol{y}.$$

Observar de hecho que la segunda derivada de la función $f(\mathbf{x})$ es $2\mathbf{A}'\mathbf{A} \geq 0$.

A.6. Ideas Geométricas: Proyecciones

Consideremos un punto $\mathbf{a}$, en n dimensiones (ver Figura A.1). Su proyección sobre un plano P a través del origen es el punto $\mathbf{a}^*$ en el pie de la perpendicular de $\mathbf{a}$ a P. El vector $\mathbf{a}^*$ se conoce como la proyección ortogonal del vector $\mathbf{a}$ sobre el plano.

Supongamos entonces que el plano P pasa por los puntos $0, \mathbf{b}_1, \ldots, \mathbf{b}_k$, lo que implica que la ecuación del plano pueda expresarse como

$$\mathbf{x} = \sum \lambda_i \mathbf{b}_i, \quad \mathbf{B} = (\mathbf{b}_1, \ldots, \mathbf{b}_k).$$

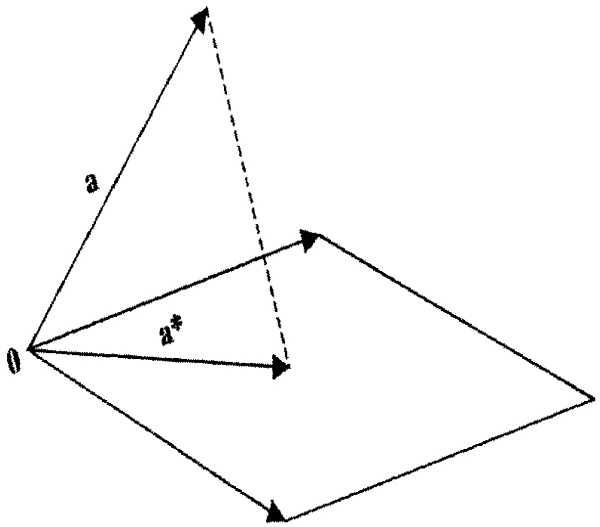


Figura A.1: Proyección del vector **a** sobre el plano P

Si el rango de la matriz **B** es k, entonces el plano P es un subespacio k-dimensional. El punto $\mathbf{a}^*$ viene definido por $\mathbf{x} = \sum \lambda_i^* \mathbf{b}_i$, donde $\lambda_1^*, \ldots, \lambda_k^*$ minimiza la distancia entre el punto **a** y el plano P.

Apéndice B

Datos

Datos del Tema 3

Ejemplo 3.1

```
edad<-c(4.5,4.5,4.5,4.0,4.,4.,5,5,5.5,5,0.5,0.5,6,6,1,1,1)
costes<-c(619,1049,1033,495,723,681,890,1522,987,1194,
163,182,764,1373,978,466,549)
plot(edad,costes,xlab="edad (en años)",ylab="Costes en 6 meses")
fit<-lm(costes~edad)
edad.s<-seq(min(edad),max(edad),length=100)
lines(edad.s,predict(fit,data.frame(edad=edad.s)))
```

Ejemplo 3.2

```
DBH <- c(10.2,13.72,15.43,14.37,15,15.02,15.12,15.24,15.24,15.28,
13.78,15.67,15.67,15.98,16.5,16.87,17.26,17.28,17.87,19.13)
D16 <-c(9.3,12.1,13.3,13.4,14.2,12.8,14,13.5,14,13.8,13.6,14,
13.7,13.9,14.9,14.9,14.3,14.3,16.9,17.3)
HT <-c(89,90.07,95.08,98.03,99,91.05,105.6,100.8,94,93.09,89,
102,99,89.02,95.09,95.02,91.02,98.06,96.01,101)
VOL <-c(25.93,45.87,56.2,58.6,63.36,46.35,68.99,62.91,58.13,
59.79,56.2,66.16,62.18,57.01,65.62,65.03,66.74,73.38,82.87,95.71)
bosque<-data.frame(VOL=VOL,DBH=DBH,D16=D16,HT=HT)
plot(bosque)
```

Ejemplo 3.4

```
# Los datos del Pizza Shack son
sales<-c(43.6,38,30.1,35.3,46.4,34.2,30.2,40.7,38.5,22.6,37.6,35.2)
ads<-c(12,11,9,7,12,8,6,13,8,6,8,10)
cost<-c(13.9,12,9.3,9.7,12.3,11.4,9.3,14.3,10.2,8.4,11.2,11.1)
pizza<-data.frame(sales,ads,cost)
```

Ejemplo 3.5

```
madera<-c(1,1.5,2,3,4,4.5,5,5.5,6,6.5,7,8,9,10,11,12,13,14,15)
tension<-c(6.3,11.1,20.0,24,26.1,30,33.8,34,38.1,39.9,42,46.1,
53.1,52,52.5,48,42.8,27.8,21.9)
plot(madera,tension,xlab="Concentración de madera (en %)",
ylab="Resistencia a la tensión")
fit<-lm(tension~madera+I(madera^2))
madera.sec<-seq(min(madera),max(madera),length=100)
lines(madera.sec,predict(fit,data.frame(madera=madera.sec)))
```

Ejemplo 3.6

```
xt<-c(2.26,0.4,1.26,2.18,4.81,2.18,5.1,0.86,-0.44,1.79,2.73,
 -5.27,1.98,1.88,1.17,-0.96,2.4,2,1.63,3.32,1.64)
xc<-c(1.23,0.42,1.01,-0.11,-0.29,-1.73,1.63,-1.04,0.12,-0.69,
0.86,0.68,-1.26,0.56,0.79,-0.49,1.5,4,-0.73,-0.93,1.38)
datos<-c(xt,xc)
tratamiento<-gl(2,21,42,labels=c("Tratamiento","Placebo"))
boxplot(datos~tratamiento)
```

Ejemplo 3.7

```
data(InsectSprays)
boxplot(count~spray,data=InsectSprays,xlab="spray",
ylab="N° supervivientes")
```

Ejemplo 3.8

```
velocidad<-c(610,950,720,840,980,530,680,540,980,730,
670,770,880,1000,760,590,910,650,810,500)
vida<-c(18.73,14.52,17.43,14.54,13.44,25.39,13.34,22.71,12.68,
19.32,30.16,27.09,25.40,26.05,33.49,35.62,26.07,36.78,34.95,43.67)
herramienta<-gl(2,10,20,labels=c("A","B"))
plot(velocidad~vida,pch=as.character(herramienta),
col=as.numeric(herramienta))

# Ajustamos pues, dos modelos diferentes para el ejemplo de ANCOVA.
# Uno con interacción (rectas distintas) y otro sin (rectas paralelas):
fit1<-lm(vida~velocidad+herramienta,x=T); summary(fit1)
fit2<-lm(vida~velocidad*herramienta,x=T); summary(fit2)

x<-seq(min(velocidad),max(velocidad),length=50)
pred.A<-data.frame(velocidad=x,herramienta=factor(rep("A",50)))
pred.B<-data.frame(velocidad=x,herramienta=factor(rep("B",50)))
y1A<-predict(fit1,pred.A)
y1B<-predict(fit1,pred.B)
y2A<-predict(fit2,pred.A)
y2B<-predict(fit2,pred.B)

etiquetas<-as.character(herramienta)
par(mfrow=c(1,2))
plot(velocidad,vida,pch=etiquetas,main="Rectas Paralelas")
lines(x,y1A,lwd=2)
lines(x,y1B,lwd=2)

plot(velocidad,vida,pch=etiquetas,main="Rectas Distintas")
lines(x,y2A,lwd=2)
lines(x,y2B,lwd=2)
```

Sección 4.5

Datos

```
x<-c(4.94,5.9,3.05,6.61,6.36,4.92,5.83,4.21,3.75,4.8,3.95,5.3,
7.15,4.38,4.42,4.39,4.55,6.68,6.01,5.65,5.54,5.19,6.39,2.89,5,
5.31,6.18,3.74,4.35,4.32)

# Respuesta Banco de datos B1
y1<-c(1762.66,41885.33,883.73,10703.56,5294,2929.36,2049.4,
665.38,1017.73,5643.21,1062.86,64464.37,12246.12, 943.32,
16123.23,4674.16,4314.54, 35907.26,27171.17,17012.14,5833.36,
783.38,5330.68,168.06, 2823.49,3167.86,33221.36,5139.98,
982.27,1481.25)

# Respuesta Banco de datos B2

y2<-c(166,364,23,751,613,147,325,59,38,106,50,200,1253,84,82,
91,89,772,421,276,298,164,637,22,137,218,509,46,66,84)

# Respuesta Banco de datos B3

y3<-c(15.47,17.5,10.95,19.12,19.24,15.76,16.83,14.7,12.64,15.48,
13.96,17.1,20.11,14.09,15.84,14.61,15.48,19.09,17.97,17.5,17.62,
16.6,18.28,10.45,16.13,16.65,18.3,12.88,14.45,14.66)

# Respuesta Banco de datos B4

y4<-c(1.91,1.63,3.48,1.56,1.57,1.99,1.63,2.45,2.43,1.96,2.51,
1.8,1.41,2.25,2.35,2.33,2.06,1.64,1.7,1.77,1.91,1.8,1.62,3.36,
1.99,1.86,1.61,2.63,2.14,2.27)
```

Apéndice C

Sintaxis en $\mathcal{R}$

Ejercicio 4.1

Simulamos 4 bancos de datos con correlaciones 0.2, 0.8, -0.5 y -0.99.

```
library(mvtnorm)
rho<-c(0.2,0.8,-0.5,-0.99)
datos<-list()
par(mfrow=c(2,2))
for(i in 1:length(rho)){
datos[[i]]<-rmvnorm(50,c(0,0),matrix(c(1,rho[i],rho[i],1),ncol=2))
plot(datos[[i]][,1],datos[[i]][,2],
main=paste("rho=",rho[i]),xlab="X",ylab="Y")
}
```

Sección 4.6

Función para calcular el Coeficiente de correlacion multiple.

```
rho.mult<-function(datos)
# datos: matriz con las variables del problema.
# La primera columna debe ser la variable respuesta.
# Las restantes p-1 columnas son las variables explicativas.
{
matriz<-var(datos)
# calculo
n<-nrow(datos)
p<-ncol(matriz)
sxx<-matriz[2:p,2:p]
syx<-matrix(matriz[1,2:p],nrow=1)
sxy<-t(syx)
#coeficiente
rho.mult<-sqrt(syx%*%solve(sxx)%*%sxy)/sqrt(matriz[1,1])
cat("\n Coeficiente de correlación multile:  ",rho.mult,"\n")
#estadistico
if(abs(rho.mult)==1)
stop("Imposible resolver contraste.
Coeficiente de correlación múltiple igual a 1", call. = FALSE)
else{
est<-((n-(p-1)-1)*rho.mult)/((p-1)*(1-rho.mult^2))
#grafico
par(mfrow=c(1,1))
x<-seq(0,qf(0.999,p-1,n-(p-1)-1),length=500)
plot(x,df(x,p-1,n-(p-1)-1),type="l",ylab="densidad",
main="Contraste Correlación Múltiple")
abline(v=qf(0.975,p-1,n-(p-1)-1),col="red")
abline(v=qf(0.025,p-1,n-(p-1)-1),col="red")
abline(v=est,col="blue")
abline(h=0)
legend(qf(0.98,p-1,n-(p-1)-1),pf((p+3)/(n-p),p-1,n-(p-1)-1),
c("Estadístico","Región Crítica"),lty=rep(1,2),
col=c("blue","red"),bty="n")
cat("\n Estadístico de contraste:  ",round(est,3),"\n")
cat("\n p-valor:  ",round(2*(1-pf(est,p-1,n-(p-1)-1)),3),"\n\n")}
return(invisible())
}
```

Ejercicio 4.3

Los gráficos de la Figura 4.3 se han obtenido de la siguiente forma:

```
n<-30
x<-rnorm(n,5,1)

y1<-(x^4)*exp(rnorm(n,2))
plot(x,y)
plot(log(x),log(y))

y2<-rpois(n,exp(x))
plot(x,y)
plot(x,log(y))

y3<-rnorm(n,10*log(x),0.4)
plot(x,y)
plot(log(x),y)

y4<-rnorm(n,10/x,0.1)
plot(x,y)
plot(1/x,y)
```

Apéndice D

Resolución de un análisis de regresión lineal

En esta práctica abordaremos la estimación de una función de producción mediante los datos procedentes de distintos sectores productivos de la economía aragonesa, obtenidos a partir de las tablas input-output de Aragón elaboradas por Ibercaja en 1990 para el año 1985 (ver Trívez, 2004).

Una función de producción de una industria establece las relaciones técnicas existentes entre factores productivos (inputs) y cantidad de producto (output) bajo cierta tecnología.

Una de las formas funcionales más utilizada en la literatura de producción en economía es la *función Cobb-Douglas*,

$$Y = AL^{\alpha}K^{\beta},$$

donde Y es la producción (output), L el factor trabajo (input 1) y K el factor capital (input 2). A, α y β son constantes positivas, donde A es un factor de escala y α y β representan las elasticidades del output a variaciones en las cantidades de los factores productivos, L y K, respectivamente.

La *elasticidad total de producción* (o elasticidad de escala), ε, mide el cambio proporcional en el output resultante de aumentar equiproporcionalmente todos los inputs. La elasticidad total de producción puede calcularse como la suma de las elasticidades parciales, es decir, con respecto a cada input. En el caso de la función de Cobb-Douglas: $\varepsilon = \alpha + \beta$.

El valor de ε se encuentra relacionado con el retorno de escala de la tecnología de la siguiente manera:

Retorno de Escala	**Elasticidad Total de Producción** (ε)
Constante	$= 1$
Creciente	> 1
Decreciente	< 1

Podemos expresar la función de producción de Cobb-Douglas de forma econométrica como

$$Y_i = AL_i^{\alpha} K_i^{\beta} e^{u_i},$$

donde u representa la perturbación aleatoria.

Para poder aplicar las herramientas estudiadas hasta el momento necesitamos transformar la expresión de arriba en un modelo lineal. Para ello tomaremos logaritmos en ambos miembros obteniendo:

$$Ln(Y_i) = Ln(A) + \alpha Ln(L_i) + \beta Ln(K_i) + u_i.$$

En definitiva, necesitaremos trabajar no sobre las variables originales sino sobre las variables transformadas a través del logaritmo.

Los datos que utilizaremos para la estimación de los parámetros de nuestro modelo econométrico estarán formados por 33 sectores productivos, el Valor Añadido Bruto (VAB) a costes de factores de cada uno de dichos sectores (output), y como inputs: la población ocupada (empleo) y el stock de capital sectorial. Disponemos de dicha información en el CD que acompaña al libro (fichero aragon.dat).

El objetivo del análisis es obtener la función de producción de la economía aragonesa haciendo uso de los datos sobre los 33 sectores productivos. Así seremos capaces de predecir niveles medios de output a partir de unos niveles de empleo y capital preestablecidos.

En primer lugar, leemos los datos en R suponiendo que el fichero aragon.dat se encuentra en el directorio de trabajo. Recuerda que puedes modificar dicho directorio a través del menú 'File' y 'Change Dir...'.

```
datos<-read.table("aragon.dat",header=T)
names(datos)
aragon<-datos[,-1]
```

Ahora tomamos logaritmos sobre las tres variables que componen la función de producción para así trabajar sobre un modelo lineal.

```
laragon<-log(aragon)
dimnames(laragon)[[2]]<-c("lvab","lempleo","lcapital")
attach(laragon)
```

Dado que estamos trabajando con una variable respuesta continua y un par de variables explicativas también continuas, es habitual comenzar el análisis mediante una inspección gráfica de la asociación existente entre las distintas variables.

```
plot(laragon)
```

Las relaciones lineales entre las variables explicativas y la variable respuesta en escala logarítmica es evidente en la Figura D.1.

Intentamos cuantificar ahora la relación lineal detectada a través del cálculo de correlaciones.

```
#Correlación simple
cor.test(lvab,lempleo)
cor.test(lvab,lcapital)
```

La relación lineal entre cada covariable y la variable respuesta es significativamente distinta cero y positiva (cor(lvab,lempleo)=0.89 y cor(lvab,lcapital)=0.85).

Sin embargo, hemos de considerar el coeficiente de correlación múltiple y las correlaciones parciales para cuantificar mejor dicha relación lineal cuando la describamos a partir de un modelo de regresión.

```
# Correlación múltiple (una vez cargada la función
# rho.mult del CD)
rho.mult(laragon)

# Correlaciones parciales
library(ggm)
parcor(cov(laragon))
```

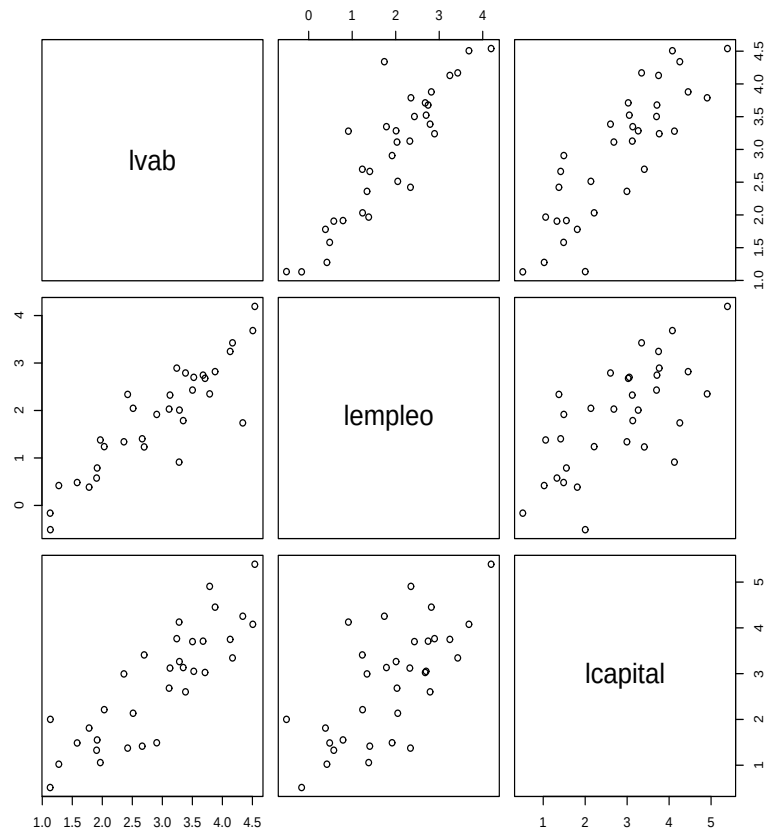


Figura D.1: Relación entre las variables. Banco de datos ARAGON.

Un resumen de las correlaciones parciales aparece en la tabla siguiente:

	lempleo	**lcapital**
lvab	0.7993322	0.7321581

La correlación múltiple es muy alta (0.95), lo que implica una relación lineal muy fuerte entre la recta de regresión y la variable del logaritmo del valor añadido, 'lvab'. A partir de los resultados de correlaciones parciales, tenemos que 'lempleo' es la covariable más relacionada linealmente con 'lvab' en presencia de la otra covariable 'lcapital'. La conclusión que se extraía en este caso a partir de las correlaciones simples era similar. Además, cada covariable parece aportar información adicional respecto de la otra a la hora de explicar la variable respuesta, como se puede concluir de la magnitud alta de sus correlaciones parciales.

Abordamos a continuación el análisis de regresión, ajustando el modelo y procediendo a su diagnóstico.

```
#Ajustamos el modelo lineal con variable respuesta 'lvab',
# y covariables 'lempleo' y 'lcapital'.
fit<-lm(lvab~lempleo+lcapital, x=T)

# Resumen del ajuste
sfit<-summary(fit)
sfit

# Error estándar residual
sfit$sigma
```

El modelo económico que deseamos estimar es el siguiente:

$$\text{LVAB} = \beta_0 + \beta_1 * \text{LEMPLEO} + \beta_2 * \text{LCAPITAL}.$$

Y estimamos $\beta_0 = 0{,}96$, $\beta_1 = 0{,}50$ y $\beta_2 = 0{,}37$.

Si ahora lo reescribimos como una función de producción Cobb-Douglas, tomando exponencial, tenemos:

$$\text{VAB} = 2.62 * \text{EMPLEO}^{0{,}50} * \text{CAPITAL}^{0{,}37}.$$

Las correlaciones parciales altas ya nos hacían intuir que los contrastes $H_0 : \beta_1 = 0$ y $H_0 : \beta_2 = 0$ iban a resultar significativos a favor del rechazo de no contribución lineal de las covariables en la explicación de la variable respuesta. La interpretación del coeficiente de determinación (raíz cuadrada del coeficiente de correlación múltiple), con un valor de $R^2 = 90\,\%$, es que el 90 % de la variabilidad de los datos es explicada por el modelo ajustado. Respecto al contrate de bondad del ajuste $H_0 : \beta_1 = \beta_2 = 0$, obtenemos un valor para el estadístico F de 138, con un p-valor asociado prácticamente de cero. Por tanto, rechazamos H_0 a favor de la hipótesis de que alguna de las covariables explica de un modo lineal la respuesta. El error residual es considerablemente pequeño, de 0.3201, lo cual da peso a la conclusión sobre la calidad del ajuste.

Procedemos ahora a construir los intervalos de confianza para los coeficientes del modelo:

```
# Intervalos de confianza para los coeficientes al 95%
confint(fit)

# y con corrección de Bonferroni
alpha<-0.05
estim<-sfit$coefficients[,1]
error<-sfit$coefficients[,2]
p<-3 #número de parámetros (b0, b1 y b2)
t.alpha<-qt(1-alpha/(2*p),df.residual(fit))
cbind("2.5 %"=estim-t.alpha*error,"97.5 %"=estim+t.alpha*error)

# o, equivalentemente,
p.adjust(sfit$coefficients[,4],method="bonferroni")
```

Obtenemos, entonces, los siguientes intervalos de confianza individuales:

IC(β_0,95 %)=(0.68, 1.25), IC(β_1,95 %)=(0.36, 0.65) y IC(β_2,95 %)=(0.24, 0.50).

Y usando la corrección propuesta por Bonferroni tenemos:

IC(β_0,95 %)=(0.61, 1.32), IC(β_1,95 %)=(0.33, 0.68) y IC(β_2,95 %)=(0.21, 0.53).

Como era de esperar, cada uno de estos últimos intervalos es más amplio que los anteriores.

Obtengamos ahora una serie de estimaciones y contrastes para el output, Valor Añadido, basadas en el modelo ajustado:

- Queremos estimar el valor del VAB para los valores medios observados de

las variables explicativas empleo y capital.

```
lempleo.0<-log(mean(aragon$empleo))
lcapital.0<-log(mean(aragon$capital))
predict(fit,data.frame(lempleo=lempleo.0,lcapital=lcapital.0),interval=
"confidence")
```

Así pues, la estimación puntual del promedio de 'VAB' en los valores medios de los factores de producción la obtenemos deshaciendo la transformación logaritmo con la función exponencial:

$$\overline{VAB} = e^{3,47} = 32,25.$$

- Queremos estimar el VAB de un nuevo sector productivo cuyo nivel de empleo es 65 y capital 200.

```
lempleo.0<-log(65)
lcapital.0<-log(200)
predict(fit,data.frame(lempleo=lempleo.0,lcapital=lcapital.0),interval=
"prediction")
```

Por tanto, tomando exponenciales, la estimación puntual para el valor añadido bruto resulta ser:

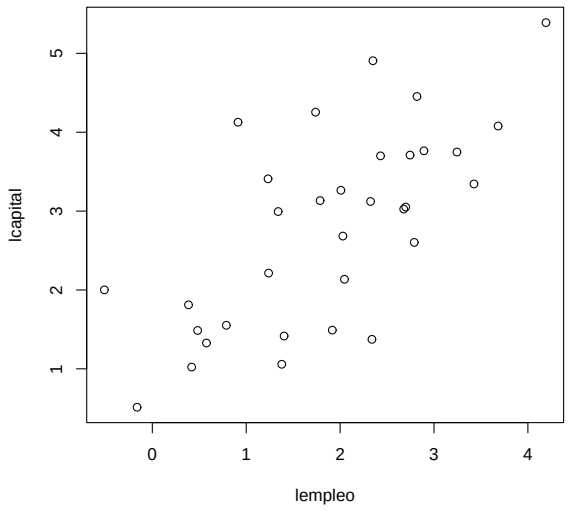
$$VAB = e^{5,02} = 151,79.$$

- Contrastemos ahora la hipótesis de si el retorno de escala de la tecnología es constante, esto es,

$$H_0 : \beta_1 + \beta_2 = 1$$

```
# Especificamos la matriz de contraste del modelo
matriz.cont<-matrix(c(0,1,1),nrow=1)
# Nos descargamos primero el package gregmisc
# y llamamos a la librería gmodels.
library(gmodels)
fit.c<-glh.test(fit,matriz.cont,d=1)
fit.c
```

El p-valor asociado al contraste sale $< 0,05$, por lo que rechazamos la hipótesis propuesta H_0. Es decir, podemos afirmar con seguridad que el retorno de escala de la tecnología no es constante. Ahora bien, la estimación puntual de la elasticidad total de producción es igual a $\beta_1 + \beta_2 = 0,50 + 0,37 = 0,87 < 1$, luego el retorno de escala de la tecnología parece ser decreciente. Es decir, si aumentamos los factores de producción en el doble, por ejemplo, el valor añadido bruto aumentaría pero en una cantidad inferior al doble.



Análisis de Multicolinealidad

Realizamos un gráfico de dispersión entre las dos covariables del modelo para ver si existe relación lineal entre ambas, y calculamos su correlación para cuantificar tal asociación.

```
# Gráfico de dispersión:
plot(laragon[,-1])
# Correlación lineal
cor(laragon[,-1])
```

El gráfico de dispersión (Figura D) muestra la existencia de una relación lineal positiva entre las dos covariables, cuantificada con un coeficiente de correlación moderadamente alto (0.68).

Pasamos a evaluar la existencia o no de multicolinealidad a través de una serie de criterios algo más sofisticados que los utilizados hasta este momento.

```
# Calculamos el FIV
library(DAAG)
vif(fit)
```

Obtenemos un valor de 1.88 que no es superior a 10. En este sentido parece que no se detectan problemas de multicolinealidad.

Una manera de relativizar FIV es através del coeficiente de determinación múltiple.

```
R2<-sfit$r.squared

# y comparamos el FIV con
1/(1-R2)
```

FIV no es mayor que 10,2. Por tanto, siguiendo este criterio no parece que existan problemas graves de multicolinealidad.

Otra forma de chequear nuestros datos es haciendo uso de los valores propios de ma matriz $X'X$. En R:

```
# Calcularemos los valores propios de la matriz X'X:
xx<-t(fit$x)%*%fit$x
lambdas<-eigen(xx)$values
lambdas
```

En nuestro caso, los valores propios son 474,91133, 15,08721 y 4,83718. Ninguno de ellos es excesivamente pequeño. Es decir, parece que mediante este criterio tampoco se concluye multicolinealidad.

Calculemos ahora el número de condición.

```
max(lambdas)/min(lambdas)
```

Con un valor de 98,18 no hay problemas de multicolinealidad o si los hay son problemas muy moderados.

Para terminar, calcularemos los índice de condición.

```
max(lambdas)/lambdas
```

Obtenemos: 1,00, 31,48 y 98,18. Como no aparece ningún índice mayor que 1000 concluimos que no parece haber problemas de multicolinealidad en nuestros datos.

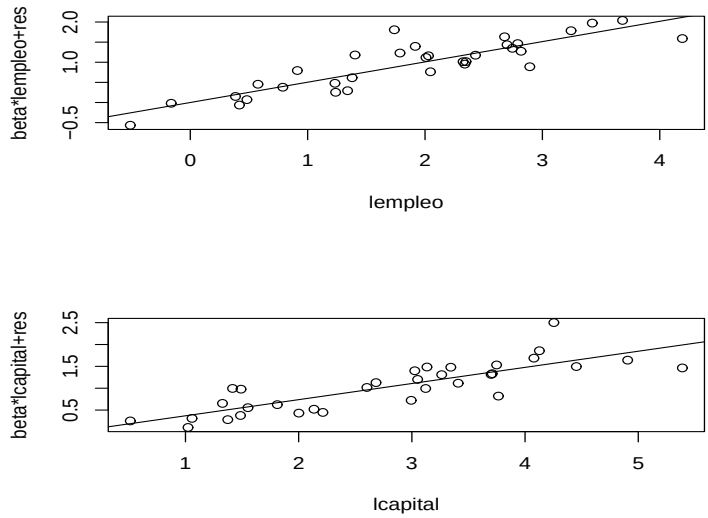


Figura D.2: Gráficos parciales de regresión.

Diagnóstico del modelo

Procedamos ahora con la diagnosis del modelo a través del análisis de los residuos. Estudiaremos punto por punto cada una de las hipótesis básicas de nuestro modelo de tegresión lineal.

- Linealidad.

```
#Calculamos  los residuos estudentizados
r<-rstudent(fit)

# Hipótesis de Linealidad:  gráficos parciales
library(faraway)
opar<-par(mfrow=c(2,1))
prplot(fit,1)
prplot(fit,2)
par(opar)
```

  En la Figura D.2 se aprecia una tendencia lineal en ambos gráficos parciales, por lo que no tenemos motivos para desconfiar de la hipótesis de linealidad.

- Homocedasticidad.

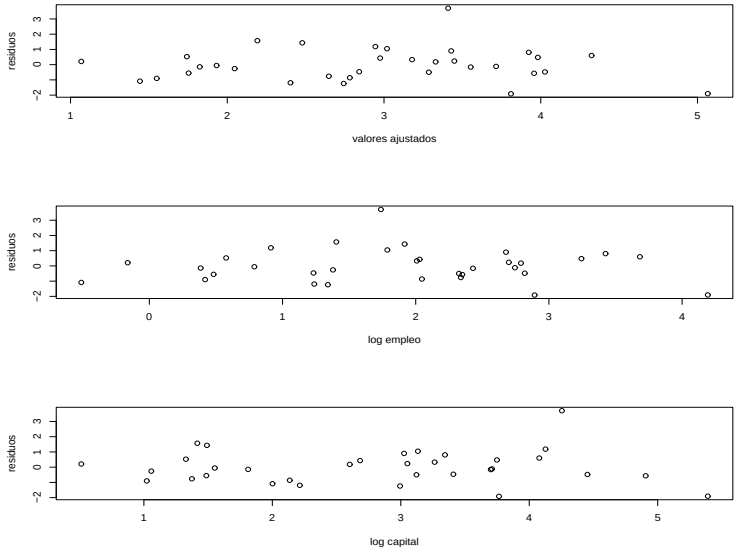


Figura D.3: Gráficos para detectar heterocedasticidad.

```
# Gráficos de residuos versus predictores y valores ajustados.

opar<-par(mfrow=c(3,1))
plot(fitted(fit),r, xlab="valores ajustados",ylab="residuos")
plot(lempleo,r, xlab="log empleo",ylab="residuos")
plot(lcapital,r, xlab="log capital",ylab="residuos")
par(opar)
```

Como podemos apreciar en la Figura D.3, en ninguno de los gráficos se detecta tendencia o curvatura. Más o menos los puntos que se observan se encuentran dentro de una banda horizontal centrada en el cero. Sólo un sector, el número 29 (Servicios a las empresas y alq. de inmuebles) presenta un residuo muy elevado (más de 3), por lo que posiblemente podamos identificarlo como 'punto influyente' tras el correspondiente análisis de influencia del modelo.

Una comprobación más rigurosa de la hipótesis de homocedasticidad la obtenemos a través del test de Breusch-Pagan:

```
library(lmtest)
bptest(fit)
```

El resultado (p-valor = 0.2086 $>$ 0,05) indica que no podemos rechazar la hipótesis nula de varianza similar en todas las observaciones.

- Normalidad.

```
# Los gráficos qqplot e histograma:
opar<-par(mfrow=c(1,2))
qqnorm(r)
qqline(r)
hist(r,probability=T)
par(opar)

# Utilizamos el test de normalidad de Shapiro-Wilks:
shapiro.test(r)

# Y el test de Kolmogorov-Smirnov para normalidad:
ks.test(r)
```

  Observando los gráficos de normalidad en la Figura D.4, tenemos que la mayoría de puntos se encuentran cercanos a la línea recta (todos salvo algunos extremos). Esto nos está indicando que la distribución de los residuos se parece mucho a una normal. En cuanto al histograma, podemos apreciar cierta asimetría. Sin embargo, es probable que este hecho se deba a la inclusión en el análisis de la observación 'rara' (sector 29), y a la partición que estamos considerando para la construcción del histograma. El test de normalidad de Shapiro-Wilks, así como el de Kolmogorov-Smirnov para normalidad, no dan evidencias para rechazar la normalidad (p-valor=0.07566 y p-valor=0.9946, respectivamente).

- Incorrelación.

```
# Gráfico de las autocorrelaciones de los residuos:
plot(acf(r))

# Resolvemos estadísticamente con el test de Durbin-Watson:
library(lmtest)
dwtest(fit,alternative="two.sided")
```

  El gráfico de autocorrelación (Figura D.5) no da indicios de correlación serial de ningún orden entre los residuos. El test de Durbin-Watson nos permite concluir que no podemos rechazar la hipótesis de incorrelación serial (p-valor=0.5836).

Como conclusión al diagnóstico del modelo, tenemos que todas las hipótesis se verifican; tan sólo se aprecia cierta desviación (no significativa) respecto de la normalidad de los residuos. Ya se ha detectado una observación (sector 29) 'rara'; con un análisis de influencia la identificaremos como influyente en el ajuste.

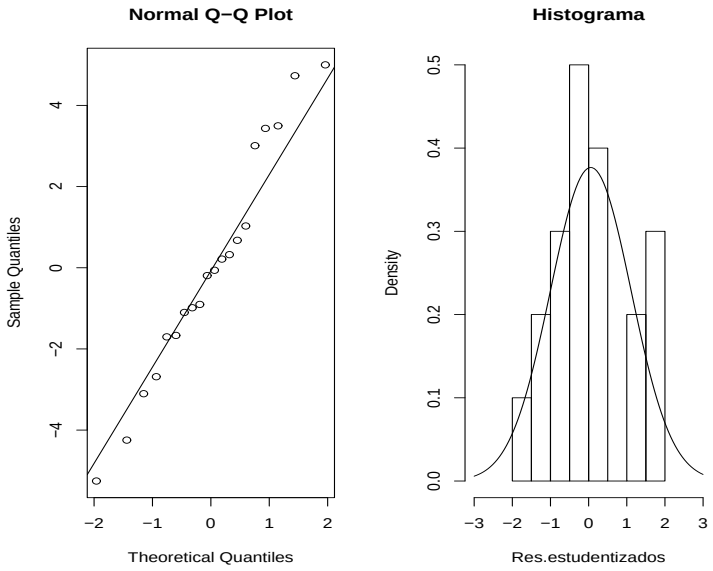


Figura D.4: Gráficos de normalidad de los residuos.

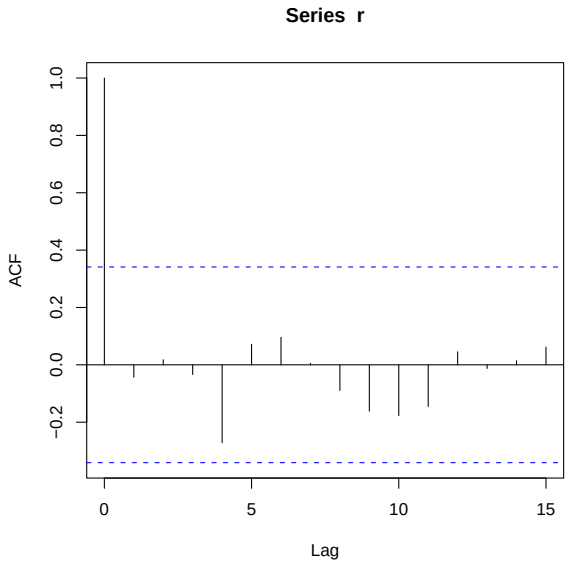


Figura D.5: Gráfico de autocorrelación de los residuos.

Análisis de Influencia

Es habitual que exista un subconjunto de los datos que influyan desproporcionadamente sobre el ajuste del modelo propuesto, con lo cual las estimaciones y predicciones dependerán mucho de él. Es interesante siempre localizar este tipo de datos. En R:

```
# Detección de puntos influyentes
im<-influence.measures(fit)
im$infmat
im$is.inf
```

El sector 21 es influyente según el criterio COVRATIO, mientras que no nos sorprende obtener el sector 29 también como influyente. En este último caso, la influencia se detecta gracias a los criterios: DFBETA (de lcapital), DFFITS y COVRATIO.

Por último, dedicaremos nuestra atención al análisis de multicolinealidad y a la validación del modelo.

Validación del modelo

Dividiremos los datos en dos grupos y realizamos el test F de validación correspondiente.

```
#Seleccionamos dos muestras al azar:

n<-length(lvab)
n1<-round(n/2)
n2<-n-n1
muestra1<-sample(1:20,n1)
muestra2<-(1:n)[-muestra1]

s1<-laragon[muestra1,]
s2<-laragon[muestra2,]

# Y ajustamos los modelos correspondientes:
formula<-lvab~lempleo+lcapital
fit1<-lm(formula,data=s1)
fit2<-lm(formula,data=s2)
# para compararlos con el ajuste global
fit<-lm(formula, data=laragon)
```

```
# y seleccionamos las sumas de cuadrados del error:
p<-length(fit1$coefficients)
sse1<-anova(fit1)[[2]][p]
sse2<-anova(fit2)[[2]][p]
sset<-anova(fit)[[2]][p]

# para calcular el estadístico F:
f<-((sset-sse1-sse2)/p)/((sse1+sse2)/(n-2*p))
# y el p-valor correspondiente
1-pf(f,p,n-2*p)

# Y pintamos los ajustes
lempleo.seq<-seq(min(lempleo),max(lempleo),length=20)
lcapital.seq<-seq(min(lcapital),max(lcapital),length=20)
newdata<-data.frame(lempleo=lempleo.seq,lcapital=lcapital.seq)

par(mfrow=c(1,2))
plot(lempleo.seq,predict(fit,newdata),type="l",lwd=2,xlab="lempleo",
ylab="lvab")
# superponemos el ajuste f1
lines(lempleo.seq,predict(fit1,newdata),lty=2,lwd=2,col="red")
# y el ajuste f2
lines(lempleo.seq,predict(fit2,newdata),lty=3,lwd=2,col="blue")

plot(lcapital.seq,predict(fit,newdata),type="l",xlab="lcapital",
ylab="lvab")
lines(lcapital.seq,predict(fit1,newdata),lty=2,lwd=2,col="red")
lines(lcapital.seq,predict(fit2,newdata),lty=3,lwd=2,col="blue")

legend(0.5,5,c("Ajuste Global","Ajuste M1","Ajuste M2"),lty=1:3,
lwd=2,col=c("black","red","blue"))
```

Así obtenemos un p-valor asociado al contraste de aproximadamente 0,80. Por lo cual, no podemos rechazar la hipótesis nula, es decir, que la función de producción ajustada da resultados similares sobre las dos muestras tomadas aleatoriamente. En consecuencia, las estimaciones de los parámetros parecen no depender demasiado de la muestra observada.

Ahora, nos gustaría calcular el error cuadrático de validación y el coeficiente de robustez, como otra forma alternativa pero a la vez complementaria, de evaluar la validez del modelo.

```
# Accedemos a la matriz de diseño X:
x<-fit$x
# y calculamos las predicciones yi(i) sin la observación i:
n<-length(lvab)  # número de datos
y.i<-vector(length=n)
for(i in 1:n)
y.i[i]<-x[i,]%*%coef(lm(lvab~lempleo+lcapital,data=laragon[-i,]))
# para calcular el error cuadrático de validación
ecv<-sum((lvab-y.i)^2);ecv

# y, finalmente, el coeficiente de robustez
b2<-sum(residuals(fit)^2)/ecv;b2
```

Por tanto, ECV= 3,89 y $B^2 = 0{,}79$. Este valor para el coeficiente de robustez, prácticamente de 0.8 (recordemos que los valores posibles estaban entre 0 y 1, y valores próximos a 1 implican robustez), nos llevaría a la conclusión de que los ajustes obtenidos quitando cada una de las observaciones no difieren demasiado del obtenido con todos los datos, esto es, que el modelo es bastante robusto.

En la Figura D.6 no se observan diferencias entre los ajustes obtenidos con las dos particiones aleatorias de datos consideradas. El resultado gráfico es coherente con el resultado estadístico obtenido.

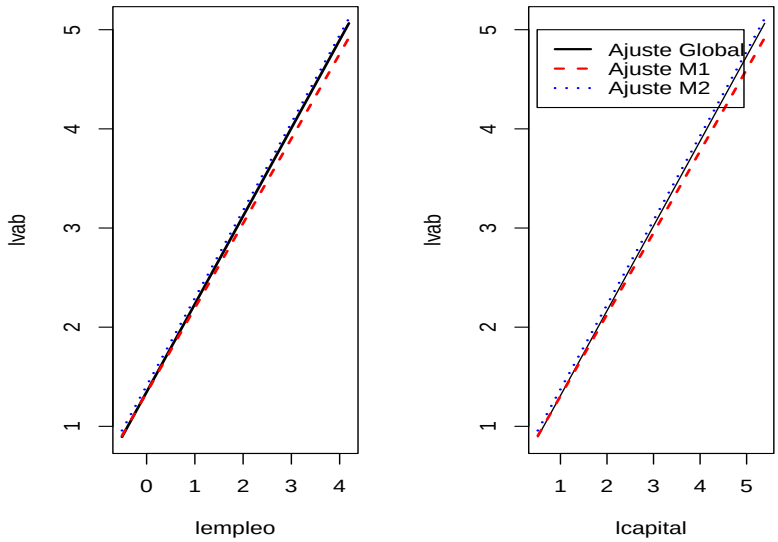


Figura D.6: Validación grupal: ajustes con dos particiones aleatorias de los datos.

Bibliografía

Aitkin, M.A. (1974). Simultaneous inference and the choice of variable subsets. *Technometrics*, **16**, 221-227.

Akaike, H. (1973). Information theory and extension of the maximum likelihood principle. In B.N.Petrov and (eds.), F., editors, *Proceedings of the 2nd International Symposium of Information Theory*, pages 267-281, Akademiai Kiado, Budapest.

Allen, D.M. (1971). Mean square error of prediction as a criterion for selecting variables. *Technometrics*, **13**, 469-475.

Allen, D.M. (1974). The relationship between variable selection and data augmentation and a method for prediction. *Technometrics*, **163**, 125-127.

Anscombe, F.J. (1973). Graphs in statistical analysis. *American Statistician*, **27**, 17-21.

Atkinson, A.C. (1985). *Plots, Transformations and Regression*. Clarendon Press, Oxford.

Barnett, V.D. and Lewis, T. (1994). *Outliers in Statistical Data, 3rd. ed.* Wiley, N.Y.

Belsley, D.A. (1991). *Conditioning Diagnostics, Collinearity and Weak Data in Regression*. (Revision of Belsley, Kuh and Welsch, 1980).

Belsley, D.A., Kuh, E. and Welsch, R.E. (1980). *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*. Willey, N.Y.

Bickel, P. and Doksum, K. (1981). An analysis of transformations revisited. *J. Am.Statist.Assoc.*, **76**, 296-311.

Box, G.E.P. and Cox, D.R. (1964). An analysis of transformations (with discussion). *J.Roy.Statist.Soc.Ser.B*, **26**, 211-246.

Box, G.E.P. and Tidwell, P.W. (1962). Transformations of the independent variables. *Technometrics*, **4**, 531-550.

Breusch, T.S. and Pagan, A.R. (1979). A Simple Test for Heteroscedasticity and Random Coefficient Variation. *Econometrica*, **47**, 1287–1294.

Carroll, R. (1980). A robust method for testing transformations to achieve approximate normality. *J.Roy.Statist. Soc., Ser.B*, **42**, 71-18.

Cook, R.D (1977). Detection of influential observation in linear regression. *Technometrics*, **19**, 15-18.

Cook, R.D (1979). Influential observations in linear regression. *J.Am.Statist.Assoc.*, **74**, 169-174.

Cook,R.D. (1998). *Regression Graphics*. New York: Wiley.

Cook, R.D and Wang, P.C. (1983). Transformations and influential cases in regression. *Technometrics*, **25**, 337-344.

Cook, R.D. and Weisberg, S. (1982). *Residuals and Influence in Regression*. Chapman and Hall, London.

Cook, R.D. and Weisberg, S. (1999). *Applied Regression Including Computing and Graphics*. New York: Wiley.

Chambers, J.M. and Hastie, T.J. (1993). *Statistical Models in S*. Chapman and Hall.

Chatterjee, S. and Hadi, A.S. (1988). *Sensitivity Analysis in Linear Regression*. Wiley N.Y.

Christensen,R. and Bedrick, E.J. (1997). Testing the independence assumption in linear models. *Joour.Amer.Statist.Assoc.*, **92**, 1006-1016.

Dobson, A.J. (2001). *An Introduction to Generalized Linear Models, 2nd ed.*. London: Chapman and Hall.

Draper, N.R. and Smith, H. (1998). *Applied Regression Analysis, 3rd ed*. John Wiley and Sons, Inc.

Ezekiel, M. and Fox, K.A. (1959). *Methods of Correlation and Regression Analysis*. Wiley, N.Y.

Ezekiel, M. (1930). *Methods of Correlation Analysis*. Wiley.

Ferrándiz Ferragud, J.R. (1994). *Apunts de bioestadística. Curs bàsic per a biòlegs.* Universitat de Valéncia.

Fisher, R.A. (1922). On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society*, **222**, 309-368.

Fisher, R.A. (1934). Two new properties of mathematical likelihood. *Proceedings of the Royal Society*, **A 144**, 285-307.

Fox, J. (1997). *Applied Regression, Linear Models, and Related Methods.* Sage.

Freund, R.J. and Wilson, W.J. (1998) *Regression Analysis. Statistical Modeling of a Response Variable.* Academic Press.

Fuller, W.A. (1996). *Introduction to statistical time series, 2nd ed.*. Wiley, New York.

Galton, F. (1894). *Natural Inheritance, 5th ed.* New York: McMillan and Company.

Garthwaite, P.H., Jolliffe, I.T. and Jones, B. (1997) *Statistical Inference.* Prentice Hall.

Gnanadesikan, R. (1977). *Methods for Statistical Analysis of Multivariate Data.* Wiley, NY.

Goldfeld,S.M. and Quandt, R.E. (1965). Some Tests for Homoskedasticity. *J. Amer. Statist. Assoc.*, **60**, 539–547.

Hahn, G.J. (1973) The coefficient of determination exposed!. *Chem. Technol.*, **3**, 609-614.

Harrison, M.J. and McCabe, B.P.M. (1979). A Test for Heteroscedasticity based on Ordinary Least Squares Residuals. *J. Amer. Statist.Assoc.*, **74**, 494–499.

Hawkins, D.M. (1980). *Identification of outliers.* Chapman and Hall, London.

Hinkley, D.V. and Runger, G. (1984). The analysis of transformed data (with discussion). *J.Am.Statist.Assoc.,* **79**, 302-319.

Hoaglin, D.C. and Welsch, R.E. (1978). The hat matrix in regression and Anova. *Am. Statist.*,, **32** N.1, 17-22.

Hocking, R.R. (1996). *Methods and Applications of Linear Models. Regression and the Analysis of Variance.* John Wiley and Sons, Inc.

Hoerl, A.E. and Kennard, R.W. (1970a). Ridge Regression: biased estimation for nonorthogonal problems. *Technometrics*, **12**, 55-67.

Hoerl, A.E. and Kennard, R.W. (1970b). Ridge Regression: Applications to nonorthogonal problems. *Technometrics*, **12**, 69-82.

Holland, P.W. (1986). Statistics and causal inference (with discussions). *Journal of the American Statistical Association*, **81**, 945.

Kennard, R.W. (1971) A note on the C_p statistic. *Technometrics*, **13**, 899-900.

Kraemer, W. and Sonnberger, H. (1986). *The Linear Regression Model under Test*. Heidelberg: Physica.

Krzanowski, W.J. (1998) *An Introduction to Statistical Modelling*. Arnold.

Larsen, W.A. and McCleary, S.J. (1972). The use of partial residual plots in regression analysis. *Technometrics*, **14**, 781-790.

Levin, R.I. and Rubin, D.S. (1998). *Statistics for Management, 7th ed.* Prentice Hall, Upper Saddle River, New Jersey.

Mallows, C.L. (1973). Some comments on C_p. *Technometrics*, **15**, 661-675.

Mayoral, A.M. y Morales, J. (2001). *Modelos Lineales Generalizados*. Universidad Miguel Hernández de Elche.

McCullagh, P. and Nelder, J.A. (1989) *Generalized Linear Models, 2nd ed.*. London: Chapman and Hall.

Mirkin, B. (2001). Eleven ways to look at the Chi-Squared coefficient for contingency tables. *The American Statistician*, **55**, 111-120.

Montgomery, D.C. and Peck, E.A. (1992). *Introduction to Linear Regression Analysis*, Second edition. John Wiley and Sons, Inc.

Myers, R.H. (1990). *Classical and Modern Regression with Applications, 2nd ed.* PWS-Kent Publishers, Boston.

Nelder, J.A. (1966). Inverse polynomials, a useful group of multi-factor response functions. *Biometrics*, **22**, 128-141.

Osborne, C. (1991). Statistical calibration: a review. *International Statistical Review*, **59**, 309-336.

Pearson, K. (1896). Mathematical contributions to the theory of evolution III. Regression, Heredity and panmixia. *Philosophical Transactions of the Royal Society of London*, **187**, 253-318.

Peña, D. (1993). *Modelos y métodos 2. Modelos lineales y series temporales.* Alianza Universidad Textos.

Rao, C.R. and Toutenburg, H. (1995). *Linear Models: Least Squares and Alternatives.* Springer-Verlag New York, Inc.

Rasch, G. (1960). *Probabilistic models for some intelligence an attainment tests.* Copenhagen: Danmarks Paedogogiske Institut.

Ryan, T. A., Joiner, B. L. and Ryan, B. F. (1976). *The Minitab Student Handbook*. Duxbury Press.

Stanton, J.M. (2001). Galton, Pearson, and the Peas: A brief history of Linear Regression for Statistics instructors. *Journal of Statistics Education*, **9**, N.3.(*On line: http://www.amstat.org/publications/jse/v9n3/stanton.html*)

Stuart, A., Ord, K. and Arnorl, S. (1999). *Classical Inference and the Linear Model, 6th ed.* Kendall's advanced theory of Statistics, volume 2A. Arnold.

Schwarz, G. (1978). Estimating the Dimension of a Model, *Annals of Statistics* **6**, 461-464.

Tibshirani, R. (1987). Estimating optimal transformations for regression. *Journal of the American Statistical Association*, **83**, 394ff.

Trívez Bielsa, F.J. (2004). *Introducción a la Econometría.* Editorial Pirámide.

Venables, W.N. and Ripley, B.D. (1997) *Modern Applied Statistics with Splus, 2nd ed.* Springer Verlag.

Weisberg, S. (1985). *Applied Linear Regression, 2nd ed.* John Wiley and Sons.

Wood, F.S. (1973). The use of individual effects and residuals in fitting equations to data. *Technometrics*, **15**, 677-695.