FACULTAD DE MATEMÁTICAS

ESTADÍSTICA E INVESTIGACIÓN OPERATIVA

Trabajo Fin de Grado

Técnicas de Selección de Variables en Minería Estadística de Datos

Adrián Guerra de la Corte

Dirigido por:
Dña. Inmaculada Barranco Chamorro

Sevilla, Junio 2016.

Abstract

A common problem in data mining, when statistical regression models are used, is to choose properly the variables to be included in the model. Throughout this work the main statistical techniques for the selection and regularization of variables will be reviewed. Also applications of these techniques will be performed by using R.
The work is divided into four chapters. In Chapter 1, we review the linear regression model, and the different correlation coefficients. In this way we introduce the basic tools to study methods of selection and regularization of variables in linear regression models.
In Chapter 2, we will see the most common criteria used for the selection of variables in classical linear models. So, we will deal with: *Adjusted coefficient of determination*, *Mallow's Coefficient*, *Cross Validation method*, *Akaike Information Criterion (AIC)* and *Bayesian Information Criterion (BIC)*. These criteria will be compared between them. Also, the main problems we may have in practice when using multiple linear regression techniques are studied. An application in R has been included to illustrate the performance of the different methods.
In Chapter 3, we focus on the so-called heuristic methods, which are a first approach to the problem of selection of variables when we have a very large number of regressors. So, selection techniques such as forward, backward and step by step are studied. Their use is again illustrated with an application.
In Chapter 4, we discuss the regularization techniques. We focus on *ridge regression* and *LASSO regression*. In this context, we show that by applying regularization techniques the problem becomes manageable, since a set of restrictions is imposed on the set of admissible solutions. As well, the geometric properties of the estimators are studied. As before, an application is included to illustrate the use of the discussed techniques in the field of medecine.
Finally, the work is completed by an appendix, which contains the R and *Mathematica* codes implemented for the development of the figures, as well as the packages of R used, and the literature consulted.

Resumen

Al utilizar modelos de regresión en Minería Estadística de Datos, un problema común es elegir de forma adecuada las variables a incluir en el modelo. A lo largo de este trabajo se revisarán las técnicas estadísticas que existen para la selección y regularización de variables. Así mismo se realizarán aplicaciones de dichas técnicas, básicamente con el software *R*.
El trabajo se estructura en cuatro capítulos. En el Capítulo 1, revisamos el modelo de regresión lineal, así como los diferentes coeficientes de correlación. De esta forma introducimos las herramientas básicas para abordar el estudio de los métodos de selección y regularización de variables en los modelos de regresión lineal.
En el Capítulo 2, veremos los criterios más usados para la selección de variables en modelos lineales clásicos. Se recogen así: el *coeficiente de determinación corregido o ajustado*, el *coeficiente C_p de Mallows*, el método de *validación cruzada*, el *criterio de información de Akaike (AIC)* y el *criterio de información bayesiana (BIC)*. Se realizan comparaciones entre ellos, y se recogen los principales problemas que se nos pueden presentar en la práctica al utilizar las técnicas de regresión lineal múltiple. Así mismo, cabe destacar que se ha ilustrado el uso de las distintas técnicas expuestas con una aplicación realizada con *R*.
En el Capítulo 3, nos centraremos en los llamados métodos heurísticos, los cuales son una primera aproximación al problema de selección de variables cuando tenemos un número muy grande de variables regresoras. Se recogen las denominadas técnicas de selección hacia adelante, hacia atrás y paso a paso. Su uso se ilustra de nuevo con una aplicación.
En el Capítulo 4, trataremos las técnicas de regularización, principalmente el modelo de regresión contraída (*ridge regression*) y el modelo de regresión *LASSO* (*LASSO regression*). Estas técnicas permiten solventar las dificultades que surgen cuando se presentan problemas de colinealidad o soluciones numéricas inestables. En este contexto, mostramos que regularizar significa, hacer el problema tratable, imponiendo una serie de restricciones al conjunto de soluciones admisibles. Además se estudian las propiedades geométricas de

los estimadores obtenidos. De nuevo se incluye una aplicación, en el campo de la Medicina, que ilustra el uso de las técnicas expuestas.
Finalmente, el trabajo se completa con un anexo, en el que se recogen los códigos R y de $Mathematica$ implementados para la elaboración de las figuras, así como los paquetes de R utilizados, y la bibliografía consultada.

Índice general

Capítulo 1

Conceptos previos

En este capítulo explicaremos los resultados básicos a la hora de introducir y comprender el estudio de métodos para la selección adecuada de las variables a incluir en un modelo de regresión lineal, simple y múltiple, así como las técnicas de regularización de dichas variables. Dividiremos el capítulo en 3 secciones: regresión lineal simple, regresión lineal múltiple y coeficientes de correlación. Esta última sección tendrá especial relevancia en el siguiente capítulo.

1.1. Modelo de regresión lineal simple

Comenzaremos con el modelo de regresión lineal simple, que consiste en expresar la dependencia lineal de la variable objetivo o dependiente, y, respecto a otras dos variables: la variable independiente, explicativa o covariable, x, y el término error o perturbación del modelo, u así

$$y_i = \beta_0 + \beta_1 x_i + u_i, \quad \text{con } (x_i, y_i) \text{ variables numéricas}$$

donde y_i y u_i son variables aleatorias, x_i es una variable conocida una vez observada y_i, y β_0 y β_1 son parámetros desconocidos del modelo.
Las hipótesis del modelo pueden formularse en términos de la variable perturbación, u_i, o de forma equivalente en términos de la variable dependiente, y. Así podemos establecer las siguientes hipótesis:

- La perturbación debe tener esperanza nula, es decir

$$E(u_i) = 0 \Leftrightarrow E(y_i) = \beta_0 + \beta_1 x_i.$$

- La varianza de la perturbación debe ser una constante que no dependa de x, de forma que

$$Var(u_i) = \sigma^2 \Leftrightarrow Var(y_i) = \sigma^2.$$

- La perturbación sigue una distribución normal a consecuencia del teorema central del límite, por tanto la distribución de y para cada x también es una distribución normal.

- Las u_i deben ser independientes entre sí

$$E(u_i u_j) = 0, \forall i \neq j \Rightarrow E(y_i y_j) = E(y_i)E(y_j), \quad \forall i \neq j.$$

Hay que estimar β_0, β_1, y σ^2 para obtener una estimación del modelo

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i.$$

Una vez construido el modelo, para comprobar las hipótesis de normalidad, homocedasticidad, independencia y linealidad, así como para estimar $\hat{\sigma}^2$, usaremos los residuos (e_i) definidos por

$$e_i = y_i - \hat{y}_i$$

siendo $\hat{y}_i$ el valor que predecimos.

1.1.1. Cálculo de los estimadores

Utilizaremos el método de máxima verosimilitud para estimar los parámetros del modelo. Comencemos estimando β_0 y β_1 viendo la función de densidad conjunta de las variables objetivo, donde cada una de las cuales sigue una distribución normal por hipótesis

$$f(y_i) = \frac{1}{\sigma\sqrt{2\pi}} exp\left[\frac{-1}{2\sigma^2}(y_i - \beta_0 - \beta_1 x_i)^2\right].$$

Por tanto la función de verosimilitud es

$$L(\beta_0, \beta_1, \sigma^2) = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n exp\left[\frac{-1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2\right]$$

pues las observaciones son independientes. Tomando logaritmo neperiano calculamos la llamada función *score* de la muestra

$$log(L(\beta_0, \beta_1, \sigma^2)) = -\frac{n}{2}log(\sigma^2) - \frac{n}{2}log(2\pi) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2.$$

Para obtener finalmente los estimadores, derivamos primero respecto a β_0 esta función e igualamos a cero

$$\frac{\partial L}{\partial \beta_0} = 0 = \sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)$$

de lo que se deduce que los residuos de la ecuación verifican

$$y_i - \hat{y}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \Rightarrow \sum_{i=1}^{n} e_i = 0$$

es decir,

$$\sum y_i = n\hat{\beta}_0 + \hat{\beta}_1 \sum x_i$$

dividiendo por n, obtenemos la expresión para calcular β_0

$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x}. \tag{1.1}$$

Y derivando respecto a β_1 e igualando a cero

$$\frac{\partial L}{\partial \beta_1} = 0 = \sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)(-x_i)$$

obtenemos que

$$\sum_{i=1}^{n} e_i x_i = 0.$$

Lo que nos indica que los residuos deben estar incorrelados con las x. Esta expresión puede ser escrita de forma equivalente como

$$\sum y_i x_i = \hat{\beta}_0 \sum x_i + \hat{\beta}_1 \sum x_i^2$$

dividiendo por n y sustituyendo la expresión 1.1

$$\frac{\sum y_i x_i}{n} = (\overline{y} - \hat{\beta}_1 \overline{x}) \underbrace{\frac{\sum x_i}{n}}_{\overline{x}} + \hat{\beta}_1 \frac{\sum x_i^2}{n}$$

$$\frac{\sum y_i x_i}{n} - \overline{x}\,\overline{y} = \hat{\beta}_1 \left(\frac{\sum x_i^2}{n} - \overline{x}^2 \right)$$

el término de la izquierda es la covarianza entre la variable x y la variable y, $Cov(x, y)$, y el término por el que está multiplicado $\hat{\beta}_1$ es la varianza muestral de x, por tanto

$$\hat{\beta}_1 = \frac{Cov(x, y)}{s_x^2}.$$

Por último, para calcular $\hat{\sigma}^2$ por el método de la máxima verosimilitud, se tiene:

$$\hat{\sigma}^2 = \frac{\sum e_i^2}{n}.$$

Observación:

Un estimador importante es la denominada *pendiente estimada como promedio de pendientes*. Sea b_i la pendiente de la recta que une (x_i, y_i) con la media de los datos, es decir:

$$b_i = \frac{y_i - \overline{y}}{x_i - \overline{x}}, \quad i = 1, ..., n.$$

Entonces se tienes que, $\hat{\beta}_1$ se puede expresar como una media ponderada de los b_i:

$$\hat{\beta}_1 = \sum_{i=1}^{n} w_i b_i \tag{1.2}$$

donde los, w_i, coeficientes de ponderación, o pesos de cada pendiente son

$$w_i = \frac{(x_i - \overline{x})^2}{\sum (x_i - \overline{x})^2}.$$

La expresión 1.2 se obtiene observando que

$$\hat{\beta}_1 = \frac{\sum (x_i - \overline{x})(y_i - \overline{y})}{\sum (x_i - \overline{x})^2} = \sum \frac{(x_i - \overline{x})}{\sum (x_i - \overline{x})^2} (y_i - \overline{y})$$

$$= \sum \frac{(x_i - \overline{x})^2}{\sum (x_i - \overline{x})^2} \frac{(y_i - \overline{y})^2}{(x_i - \overline{x})} = \sum w_i b_i.$$

1.2. Modelo de regresión lineal múltiple

El modelo de regresión lineal múltiple es la extensión del modelo de regresión lineal simple cuando consideramos k variables explicativas. En general, la variable objetivo $\boldsymbol{Y}$ depende de muchas otras variables $x_1, ..., x_k$, aunque algunas de éstas pueden no ser observables o desconocidas. El modelo de regresión incluye las que más efecto tienen y las restantes las representa como una variable aleatoria que denominaremos perturbación o error del modelo, por tanto, tenemos

$$y = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k + u,$$

para cada observación, sería:

$$y_i = \beta_0 + \beta_1 x_{1i} + ... + \beta_k x_{ki} + u_i, \quad i = 1, ..., n.$$

Las hipótesis son:

- $E[u] = 0 \Leftrightarrow E[y] = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k$.
- $Var(u) = \sigma^2 \Leftrightarrow Var(y) = \sigma^2$.
- Las perturbaciones, u_i, son independientes entre sí $\Rightarrow$ las variables y_i son independientes entre sí.
- $u \sim N(0, \sigma^2) \Rightarrow y \sim N(\beta_0 + \beta_1 x_1 + ... + \beta_k x_k, \sigma^2)$.

Introduciendo notación matricial, llamemos $\boldsymbol{U'}$ al vector de errores, $\boldsymbol{U'} = (u_1, ..., u_n)$. $\boldsymbol{U'}$ sigue una distribución normal multivariante con un vector de medias nulo, y matriz de varianzas y covarianzas $\sigma^2 \mathbf{I}_n$, siendo $\mathbf{I}_n$ la matriz identidad $n \times n$, $\mathbf{U} \sim N(0, \sigma^2 \mathbf{I}_n)$. Estableceremos dos hipótesis más para estimar los parámetros del modelo:

- El número de datos que tenemos para el estudio es mayor que el número de parámetros del modelo.
- Las variables explicativas son linealmente independientes.

Nota 1.2.1. *Si el número de observaciones fuera menor o igual que el número de parámetros, haciendo todas las $u_i = 0$ tendríamos un sistema con $k+1$ incógnitas y n ecuaciones. Por tanto, si n=k+1 habría solución única y si $n < k+1$ habría infinitas soluciones, en ninguno de los dos casos podríamos estimar la variabilidad de la perturbación, pues todos los residuos (o perturbaciones estimadas) serían cero.*

Nota 1.2.2. *Si alguna variable explicativa no fuera linealmente independiente, no tendríamos k variables sino $k-1$, por ejemplo $x_2 = a + bx_1$, luego tendremos:*

$$y = \beta_0 + \beta_1 x_1 + \beta_2(a + bx_1) + ... + \beta_k x_k + u$$

$$y = \underbrace{(\beta_0 + a\beta_2)}_{\beta_0'} + \underbrace{(\beta_1 + b\beta_2)}_{\beta_1'} x_1 + \beta_3 x_3 + ... + \beta_k x_k + u$$

y este modelo equivale a uno con $k-1$ variables, por lo que debe tenerse en cuenta al modelizar unos datos.

1.2.1. Cálculo de los estimadores

Como en la regresión lineal simple, usaremos la estimación por el método de máxima verosimilitud.

Por hipótesis tenemos que la distribución de la variable objetivo, y, es normal, por tanto, la función de densidad de una observación es:

$$f(y_i) = \frac{1}{\sigma\sqrt{2\pi}} exp\left[\frac{-1}{2\sigma^2}(y_i - \beta_0 - \beta_1 x_{1i} - ... - \beta_k x_{ki})^2\right]$$

de la cual obtenemos la función de verosimilitud:

$$L(\beta_0, \beta_1, ..., \beta_k, \sigma^2) = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n exp\left[\frac{-1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_{1i} - ... - \beta_k x_{ki})^2\right]$$

pues las observaciones son independientes. Tomando logaritmo calculamos la función *score* de la muestra:

$$log(L(\beta_0, \beta_1, ..., \beta_k, \sigma^2)) = -\frac{n}{2}log(\sigma^2) - \frac{n}{2}log(2\pi) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_{1i} - ... - \beta_k x_{ki})^2.$$

Para obtener finalmente los estimadores, derivamos primero respecto a β_0 ésta función e igualamos a cero:

$$\frac{\partial L}{\partial \beta_0} = 0 = \sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{1i} - ... - \hat{\beta}_k x_{ki})$$

de lo que se deduce que los residuos de la ecuación son los siguientes y verifican las siguientes relaciones:

$$y_i - \hat{y}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{1i} - ... - \hat{\beta}_k x_{ki} \Rightarrow \sum_{i=1}^{n} e_i = 0.$$

Y derivando respecto a los demás parámetros β_j e igualando a cero se obtiene

$$\sum_{i=1}^{n} e_i x_{ji} = 0, \quad j = 1, ..., k.$$

Las ecuaciones obtenidas se conocen como *ecuaciones normales de la regresión.*
De la primera ecuación se deduce que la media de los residuos debe ser cero, y junto con las restantes ($j = 1, ..., k$), que las covarianzas entre los residuos y las variables explicativas deben ser también cero (no tienen relación lineal). Queda entonces asegurado que toda la información sobre la relación entre la variable objetivo y las variables explicativas, está en $\hat{y}_i$, y la que no está relacionada con las variables explicativas incluidas en el modelo, está en e_i.
Para simplificar la notación, podemos escribir el sistema de forma matricial

de manera que, definiremos una matriz $\boldsymbol{X}$, con dimensión $n \times (k+1)$, que tiene por columnas a las variables explicativas y la columna que corresponde al término β_0. La variable objetivo irá en el vector $\boldsymbol{Y}$, de dimensión $n \times 1$, y los parámetros en el vector $\boldsymbol{\beta}$, de dimensión $k+1$, por tanto tenemos:

$$\boldsymbol{X'Y} = \boldsymbol{X'X}\widehat{\boldsymbol{\beta}}.$$

Por las hipótesis antes expuestas, la matriz $\boldsymbol{X'X}$ es no singular, entonces podemos estimar los parámetros por el método de mínimos cuadrados ordinarios (MCO):

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X'X})^{-1}\boldsymbol{X'Y}$$

luego

$$\widehat{\boldsymbol{Y}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}} = \boldsymbol{X}(\boldsymbol{X'X})^{-1}\boldsymbol{X'Y}.$$

Como al final de la sección 1.1.1, daremos la expresión del estimador de $\hat{\sigma}^2$ basado en los residuos del modelo, que están dados por:

$$\boldsymbol{e} = \boldsymbol{Y} - \widehat{\boldsymbol{Y}} = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y}$$

donde $\boldsymbol{H}$ es la *matriz de predicción o hat matrix*, la cual es simétrica e idempotente y está definida como sigue:

$$\boldsymbol{H}_M = \boldsymbol{X}_M(\boldsymbol{X'}_M\boldsymbol{X}_M)^{-1}\boldsymbol{X'}_M$$

donde el subíndice M hace referencia al modelo.
Por tanto, el estimador por el método de máxima verosimilitud de σ^2, teniendo en cuenta la hipótesis de normalidad, es:

$$\hat{\sigma}^2 = \frac{1}{n}\boldsymbol{e'e} = \frac{1}{n}\sum e_i^2 = \frac{1}{n}\boldsymbol{Y'}(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y}.$$

Proposición 1.2.1. *La matriz de predicción o hat matrix, es simétrica e idempotente.*

Demostración. Veamos primero que es simétrica:

$$\boldsymbol{H'} = (\boldsymbol{X'})'[(\boldsymbol{X'X})^{-1}]'\boldsymbol{X'} = \boldsymbol{X}(\boldsymbol{X'X})^{-1}\boldsymbol{X'} = \boldsymbol{H}.$$

Y ahora veamos que es idempotente, es decir, $\boldsymbol{H} = \boldsymbol{H^2}$:

$$\boldsymbol{H}^2 = [\boldsymbol{X}(\boldsymbol{X'X})^{-1}\boldsymbol{X'}][\boldsymbol{X}(\boldsymbol{X'X})^{-1}\boldsymbol{X'}] =$$

$$= \boldsymbol{X}(\boldsymbol{X'X})^{-1}\boldsymbol{X'X}(\boldsymbol{X'X})^{-1}\boldsymbol{X'} = \boldsymbol{X}(\boldsymbol{X'X})^{-1}\boldsymbol{X'} = \boldsymbol{H}.$$

□

1.2.2. Intepretación de los coeficientes en un modelo de regresión lineal múltiple

El coeficiente de regresión estimado para una variable x_i en el modelo de regresión lineal múltiple, $\hat{\beta}_i$, representa el efecto sobre la variable objetivo cuando la variable x_i aumenta en una unidad y las demás variables explicativas premanecen constantes. Puede interpretarse como el efecto diferencial de esta variable cuando eliminamos o controlamos los efectos de las demás.
Debemos distinguir dos situaciones a la hora de interpretar los coeficientes, cuando las variables explicativas están incorreladas y cuando no lo están:

- Cuando todas las variables explicativas están incorreladas se calcula de la misma manera que en la regresión simple. Pues en este caso, el efecto diferencial de la variable, medido por la regresión múltiple, es igual al efecto total medido por la regresión simple.

- Cuando las variables están correladas, el coeficiente de regresión de x_i se puede expresar también como el cociente entre una covarianza y una varianza. Con la salvedad de que en la regresión simple se utiliza la covarianza entre la variable objetivo y x_i, y en la múltiple se utiliza la covarianza entre la variable objetivo y la parte diferencial de x_i o no correlada con el resto de variables explicativas. La parte diferencial de x_i está definida por los residuos de una regresión entre la variable x_i y el resto de variables explicativas en la ecuación de regresión. Estos residuos serán $e_{i,R}$. Luego, $\hat{\beta}_i = Cov(y, e_{i,R})/Var(e_{i,R})$, donde no se usa la variable x_i como en la regresión simple, sino la parte diferencial de ella, $e_{i,R}$.
  En el caso de que esta variable sí esté incorrelada, $x_i = e_{i,R}$ y el coeficiente de la regresión múltiple es igual al de la regresión simple.

1.2.3. Contrastes

A la hora de calcular un modelo de regresión lineal, los contrastes son una herramienta importante.
En esta sección hablaremos de los dos contrastes que usaremos en este trabajo, el *contraste global de regresión* y el *contraste individual de la t*.

Contraste global de regresión

El contraste es el siguiente:

$$H_0 : \beta_1 = ... = \beta_k = 0.$$

$$H_1 : \text{algún } \beta_i \neq 0, \quad i = 1, ..., k.$$

Y el estadístico resultante se denota por F y se calcula

$$F = \frac{SCE/k}{SCR/n-k-1} = \frac{\hat{s}^2_{expl}}{\hat{s}^2_r}.$$

Bajo la hipótesis H_0, $F \sim F_{k,n-k-1}$.
Dicho contraste se traduce en

- Si acepto $H_0 \Rightarrow$ **ninguna** de las variables explicativas consideradas influyen linealmente en la variable respuesta.
- Si rechazo $H_0 \Rightarrow$ **alguna o todas** las variables explicativas consideradas influyen linealmente en la variable respuesta.

Contraste individual de la t

Para cada variable se plantea el siguiente contraste:

$$H_0 : \beta_i = 0.$$

$$H_1 : \beta_i \neq 0.$$

El estadístico que resulta del contraste está basado en el estadístico de $Wald$ y se define como sigue

$$t_i = \frac{\hat{\beta}_i}{\widehat{s}_r \sqrt{q_{ii}}}$$

siendo $\hat{\beta}_i$ el estadístico de $Wald$ y q_{ii} el término (ii) de la matriz $(\boldsymbol{X'X})^{-1}$. Éste, bajo H_0, sigue una distribución t_{n-k-1}.
Como es usual, en contrastes de hipótesis, se rechaza H_0, si el *p-valor* obtenido es menor o igual que el nivel de significación del contraste, α. En este caso podremos suponer que $\beta_i \neq 0$. Si no se rechaza H_0, podremos suponer que $\beta_i = 0$.

1.3. Coeficientes de correlación

En esta sección se definen y estudian las propiedades de los *coeficientes de correlación lineal simple, coeficiente de determinación, coeficiente de correlación múltiple* y *parcial*, así como las relaciones existentes entre ellos.
Una medida de la relación lineal entre dos variables cualesquiera es el *coeficiente de correlación lineal simple.*

Definición 1.3.1. *Dadas dos variables x e y, se denomina* ***coeficiente de correlación lineal simple****, r_{xy}, a*

$$r_{xy} = \frac{Cov(x,y)}{s_x s_y}$$

donde s_x y s_y son las desviaciones típicas muestrales de las variables x e y, respectivamente.

Dicho coeficiente se puede expresar en función de la varianza residual

$$r_{xy}^2 = \frac{SCE(x,y)}{SCT(y)} = 1 - \frac{SCR(x,y)}{SCT(y)}.$$

Definición 1.3.2. *Definimos el* ***coeficiente de determinación****, R^2, de un modelo, para evaluar la bondad de ajuste de una recta de regresión (simple o mútiple) con una proporción de la variación explicada con la siguiente expresión*

$$R^2 = \frac{SCE}{SCT} = \frac{\sum(\hat{y}_i - \overline{y})^2}{\sum(y_i - \overline{y})^2}, \quad 0 \leq R^2 \leq 1.$$

Siendo $\sqrt{R^2}$ el ***coeficiente de correlación múltiple****.*

Definición 1.3.3. *Dado un conjunto de variables explicativas $(x_1, ..., x_k)$, el* ***coeficiente de correlación parcial****, denotado por $r_{ij,1,...,i-1,i+1,...,j-1,j+1,...,k}$, entre dos cualesquiera de ellas, x_i y x_j, mide la relación lineal entre x_i y x_j una vez eliminados los efectos de las demás sobre ellas. Se denotará por $r_{12,34...k}$.*

La manera de calcularlo es útilizando la siguiente expresión:

$$e_{1,34..k} = \hat{\beta} e_{2,34..k} + u$$

donde $e_{1,34..k}$ y $e_{2,34..k}$ son los residuos de la regresión múltiple de x_1 y x_2 respecto a las demás variables $(x_3, ..., x_k)$, de esta manera obtendríamos $r_{12,34..k}$.

Cálculo del coeficiente de correlación parcial

A continuación se deduce la fórmula de la correlación parcial de dos variables (x, y) cuando se mantiene constante una tercera variable z.

Proposición 1.3.1. *El coeficiente de correlación parcial entre x y y manteniendo constante z, viene dado por*

$$r_{xy.z} = \frac{r_{xy} - r_{xz}r_{yz}}{\sqrt{(1-r_{xz}^2)(1-r_{yz}^2)}}. \tag{1.3}$$

donde $r_{xy}, ..., r_{yz}$ son los coeficientes de correlación lineal simple entre las variables implicadas.

Demostración. Supondremos que las tres variables tienen media cero para simplicar la exposición, esto no altera el resultado.
Sean las rectas $\hat{x} = az$, $\hat{y} = bz$ cuyos coeficientes son, por definición:

$$a = \frac{\sum x_i z_i}{\sum z_i^2}$$

$$b = \frac{\sum y_i z_i}{\sum z_i^2}.$$

Puesto que la correlación parcial entre x e y, fijada z, es la correlación entre los residuos de estas regresiones, tenemos que:

$$\text{Correlación}[(x-\hat{x})(y-\hat{y})] = \frac{Cov(x,y)}{\sqrt{Var(x-\hat{x})Var(y-\hat{y})}} = r_{xy.z}.$$

Pasemos a calcular el numerador:

$$n\ Cov(x,y) = \sum (x_i - \hat{x}_i)(y_i - \hat{y}_i) = \sum (x_i - az_i)(y_i - bz_i)$$

$$= \sum x_i y_i - a\sum z_i y_i - b\sum z_i x_i + ab\sum z_i^2.$$

Sustituyendo a y b en la expresión anterior, llegamos a:

$$n\ Cov(x,y) = \sum x_i y_i - \frac{(\sum x_i z_i)(\sum z_i y_i)}{\sum z_i^2} - \frac{(\sum y_i z_i)(\sum x_i z_i)}{\sum z_i^2}$$

$$+\frac{(\sum x_i z_i)(\sum y_i z_i)}{\sum z_i^2} = \sum x_i y_i - \frac{(\sum x_i z_i)(\sum z_i y_i)}{\sum z_i^2}.$$

Si introducimos ahora los coeficientes de correlación simples, nos queda:

$$n\ Cov[(x-\hat{x})(y-\hat{y})] = r_{xy}\sqrt{\sum x_i^2 \sum y_i^2} - r_{xz}r_{yz}\sqrt{\sum x_i^2 \sum y_i^2}$$

$$= (r_{xy} - r_{xz}r_{yz})\sqrt{\sum x_i^2 \sum y_i^2}.$$

Con eso tenemos calculado el numerador, pasemos a hallar denominador:

$$n\ Var(x - \hat{x}) = \sum (x_i - az_i)^2 = \sum x_i^2 - \frac{(\sum x_i z_i)^2}{\sum z_i^2}$$

$$n\ Var(y - \hat{y}) = \sum (y_i - bz_i)^2 = \sum y_i^2 - \frac{(\sum y_i z_i)^2}{\sum z_i^2}.$$

Sustituyendo en las varianzas de los residuos los coeficientes de correlación simples, se tiene:

$$n\ Var(x - \hat{x}) = \sum x_i^2 (1 - r_{xz}^2)$$

$$n\ Var(y - \hat{y}) = \sum y_i^2 (1 - r_{yz}^2).$$

Por lo que finalmente llegamos a la expresión 1.3.

□

Este coeficiente al cuadrado tiene la misma interpretación que el coeficiente de correlación simple. Es decir, $r_{xy.z}^2$ representa la proporción de variación explicada respecto a la variación no explicada por otra regresión previa.

1.3.1. Relación entre las correlaciones parciales y la múltiple

Supongamos, por simplificar, que tenemos nada más dos variables explicativas, x_1 y x_2. Sea r_{yx_1} el coeficiente de correlación simple entre la variable objetivo y x_1. Que por lo visto anteriormente es:

$$r_{yx_1}^2 = \frac{SCE(y, x_1)}{SCT(y)} = 1 - \frac{SCR(y, x_1)}{SCT(y)}$$

donde $SCE(y, x_1)$ es la variación explicada en la regresión de y respecto a x_1. Luego:

$$SCR(y, x_1) = SCT(y)(1 - r_{yx_1}^2).$$

Ahora debemos determinar la parte diferencial de la segunda variable, $e_{2,1}$, los cuales calculamos haciendo la regresión $x_2 = \hat{b}x_1 + e_{2,1}$. Una vez calculados, los relacionamos con la parte de la variable objetivo que no está explicada por x_1, que serán los residuos $e_{y.x_1}$ de la regresión simple de y respecto a x_1. La relación de ambos residuos está dada por el coeficiente de correlación parcial, el cual nos proporciona los residuos de la regresión múltiple.
La estimación será $e_{y.x_1} = \hat{\beta} e_{2,1} + e_{y,12}$, por tanto:

$$r_{y2,1}^2 = 1 - \frac{SCR(e_{y,12})}{SCT(e_{y.x_1})} = 1 - \frac{SCR(e_{y,12})}{SCR(y, x_1)}.$$

Y como $e_{y,12}$ son los residuos de la regresión múltiple con ambas variables:

$$R^2 = 1 - \frac{SCR(e_{y,12})}{SCT(y)}$$

luego

$$1 - R^2 = (1 - r^2_{yx_1})(1 - r^2_{y2,1}).$$

Que se interpreta como la proporción de la variabilidad no explicada en la regresión múltiple es el producto de:

- la proporción no explicada en la regresión simple de la variable objetivo y x_1.
- la proporción no explicada en la regresión de la variable objetivo y x_2 con x_1 fija.

Dicho resultado se puede extender para k regresores:

$$1 - R^2 = (1 - r^2_{y1})(1 - r^2_{y2,1})(1 - r^2_{y3,12})...(1 - r^2_{yk,12...k-1}).$$

Con esta idea también podemos relacionar los coeficientes de correlación múltiple con k y $k-1$ variables, en función del coeficiente de correlación parcial de la variable no incluida, x_h. Llamando R^2_k y R^2_{k-1} a los coeficientes de correlación múltiple con k y $k-1$ variables, respectivamente, llegamos a

$$1 - R^2_k = (1 - R^2_{k-1})(1 - r^2_{yh,12..k}).$$

lo que es lo mismo que:

$$R^2_k - R^2_{k-1} = r^2_{yh,12..k}(1 - R^2_{k-1}).$$

El término de la izquierda representa el incremento de variación explicada entre la regresión que incluye a la variable y la que no la incluye. El término de la derecha es el producto del porcentaje de variación explicada por x_h respecto a la variación no explicada por las restantes variables y del porcentaje de variación no explicada por las restantes variables, $x_1, ..., x_k$ respecto al total. Esta expresión nos permite calcular los coeficientes de correlación parcial de cada variable a partir de un programa que nos calcule la regresión.
En general, la fórmula anterior puede escribirse de tal manera que, notando $(\bar{h})$, por el modelo que no incluye a x_h, y por (h) al modelo que sí la incluye:

$$\frac{\Delta SCE(h)}{SCT} = \frac{\Delta SCE(h)}{SCR(\bar{h})} \cdot \frac{SCR(\bar{h})}{SCT}$$

donde $\Delta SCE(h) = SCE(todas) - SCE(\bar{h})$.
Por último, utilizando el estadístico t mencionado en la Sección 1.2.3 para contrastar la hipótesis de $\beta_h = 0$, llegamos a:

$$r^2_{yh,12\ldots k} = \frac{t_h^2}{t_h^2 + n - (k+1)}$$

lo cual nos permite calcular el *coeficiente de correlación parcial* si sabemos el estadístico t para ese coeficiente.

Capítulo 2

Técnicas de selección de variables en modelos lineales clásicos

A la hora de construir un modelo tenemos diferentes posibilidades, las cuales se ajustan mejor o peor a la realidad. En este capítulo nos centraremos en los criterios más usados para la selección de variables en modelos lineales, que son:

- Coeficiente de determinación corregido o ajustado: Es un coeficiente que mide la intensidad de la relación lineal entre la variable objetivo y las predictoras.
- Coeficiente C_p *de Mallows*: Criterio que recibe el nombre del estadístico británico *Colin Lingwood Mallows*. Este criterio selecciona el modelo que tiene mayor capacidad de predicción en vez del que está mejor ajustado. La capacidad de predicción se mide con el error cuadrático medio (ECM).
- Validación cruzada: Evolución del llamado *holdout method* que se basa en la partición del conjunto de datos en dos, uno nos permite estimar los parámetros del modelo, y el otro evaluar la capacidad predictiva de éste. De esta forma se selecciona el modelo valorando su bondad de ajuste y capacidad de predicción.
- Criterio de Informacion de Akaike (AIC): Criterio propuesto por el estadístico japonés *Hirotugu Akaike* y que está basado en la teoría de la información. Está definido de forma que bonifica la bondad de ajuste y penaliza la inclusión de parámetros a estimar, lo que ayuda a evitar el fenómeno del sobreajuste.

- Criterio de Información Bayesiana (BIC): El profesor *Gideon E. Schwarz* propuso este criterio bajo un enfoque bayesiano que se basa en las probabilidades a *posteriori* de los modelos. Es, junto al AIC, el más usado.

2.1. Coeficiente de determinación corregido o ajustado

El coeficiente de determinación R^2 es una medida de bondad de ajuste de un modelo a unos datos. Recordemos que R^2 nos da la proporción de la variabilidad de Y aplicada por el modelo, es decir,

$$R^2 = \frac{SCExplicada}{SCT} = \frac{\sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2} = 1 - \frac{SCResidual}{SCT}$$

y además $0 \leqslant R^2 \leqslant 1$. Cuanto más cercano esté a 1, mejor ajustado está el modelo.
R^2 no nos sirve para comparar modelos diferentes, puesto que siempre aumenta cuando se añaden nuevas variables explicativas al modelo, lo que nos llevaría a tomar modelos con innumerables variables superfluas. Por eso surge el coeficiente de determinación corregido o ajustado, que sirve para solventar este problema puesto que incluye un término de corrección por el número de parámetros en el modelo.

$$\widehat{R}^2_{aj,k} = 1 - \frac{n-1}{n-k}(1 - R^2_k).$$

$\widehat{R}^2_{aj}$ es muy popular y viene incorporado en los programas estadísticos y resulta de especial interés en situaciones en las que el número de variables explicativas está cercano al número de observaciones de la muestra.

Teorema 2.1.1. *El estadístico $\widehat{R}^2_{aj,k}$ aumenta al introducir un nuevo parámetro, β_{k+1}, en la ecuación de regresión si el estadístico Q_h asociado al contraste de significación de dicho parámetro es mayor que 1. Q_h se define como:*

$$Q_h = \frac{SCR_k - SCR_{k+1}}{SCR_{k+1}} \times \frac{n-k-1}{1}$$

donde SCR_k es la suma de cuadrados de los residuos en el modelo con k covariables.

Demostración. Para hacer el contraste del (k+1)-ésimo parámetro emplearemos el estadístico Q_h, que se definió como:

$$Q_h = \frac{SCR_k - SCR_{k+1}}{SCR_{k+1}} \times \frac{n-k-1}{1}$$

$$= \frac{R^2_{k+1} - R^2_k}{1 - R^2_{k+1}} \times \frac{n-k-1}{1}$$

por tanto:

$$(1 - R^2_{k+1})Q_h = (R^2_{k+1} - R^2_p)(n-k-1)$$

$$Q_h - Q_h R^2_{k+1} = (n-k-1)R^2_{k+1} - (n-k-1)R^2_k$$

$$Q_h + (n-k-1)R^2_k = R^2_{k+1}[(n-k-1) + Q_h]$$

despejando R^2_{k+1}:

$$R^2_{k+1} = \frac{Q_h + (n-k-1)R^2_k}{(n-k-1) + Q_h}$$

$$= \frac{\frac{1}{n-k-1}Q_h + R^2_k}{1 + \frac{1}{n-k-1}Q_h}.$$

Sustituyendo esta expresión en la definición de $\widehat{R}^2_{aj,k+1}$, tenemos:

$$\widehat{R}^2_{aj,k+1} = 1 - (1 - R^2_{k+1})\frac{n-1}{n-k-1} = 1 - \frac{1 - R^2_k}{\frac{n-k-1+Q_h}{n-k-1}} \times \frac{n-1}{n-k-1}$$

$$= 1 - (1 - R^2_k)\frac{n-1}{n-k-1+Q_h} = \underbrace{1 - (1 - R^2_k)\frac{n-1}{n-k}}_{\widehat{R}^2_{aj,k}} \times \underbrace{\frac{n-k}{n-k-1+Q_h}}_{t}$$

de lo que se deduce que $\widehat{R}^2_{aj,k+1} \geq \widehat{R}^2_k$ si $Q_h > 1$.

□

2.1.1. Aplicación

A continuación se recoge un ejemplo que ilustra el uso del *coeficiente de determinación corregido o ajustado*. Utilizaremos el conjunto de datos que nos proporciona *Fahrmeir, L. et al.* [1].
En primer lugar se cargará el conjunto de datos.

```
> golf <- read.table("golffull.txt", header=TRUE)
> attach(golf)
```

Dicho conjunto posee 5 variables explicativas:

- Age: edad (en años).
- Kilometer: kilómetros (en miles).
- TIA: número de meses hasta la próxima Inspección Técnica de Vehículos o ITV.
- Extras1: Si tiene ABS o no.
- Extras2: Si tiene techo solar o no.

Las 5 primeras variables son cuantitativas y las 2 últimas son factores, por lo que habrá que tenerlo en cuenta.
Estudiamos gráficamente qué relación hay entre la variable objetivo y las variables explicativas, para ello realizamos gráficos de nube de puntos de la variable $\mathbf{Y}$ frente a cada una de ellas [Figura 2.1].

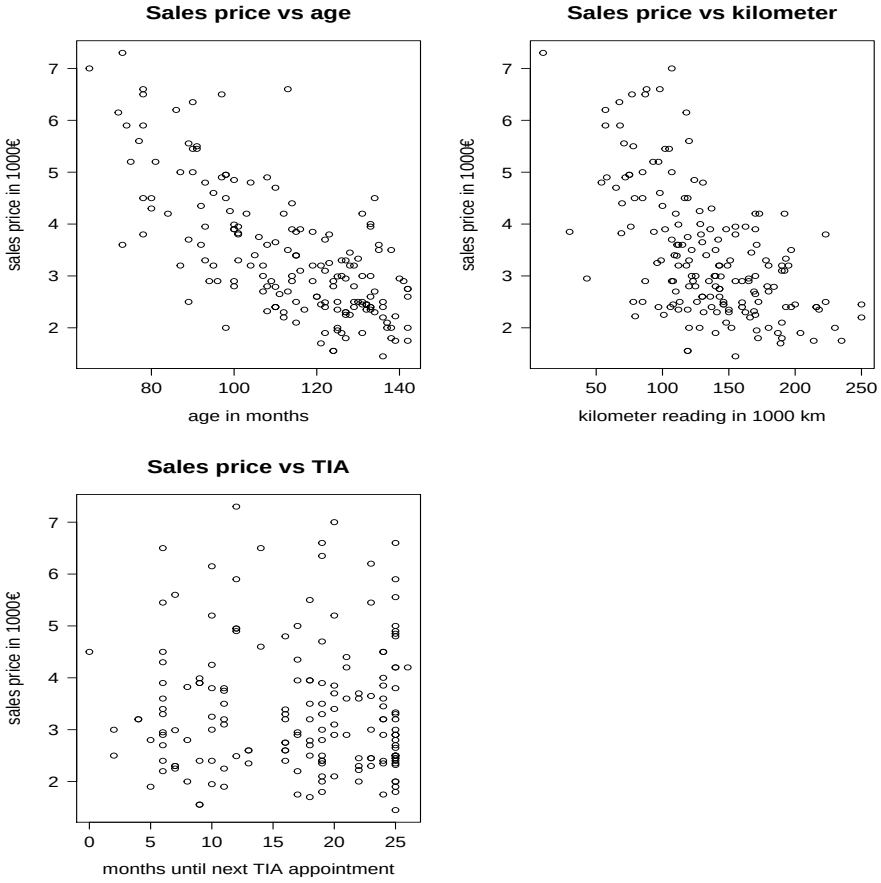


Figura 2.1: Podemos ver que tanto *age* como *kilometer* tienen una relación lineal inversa con la variable objetivo. Y que la variable explicativa TIA, no tiene una relación lineal con $\mathbf{Y}$.

Para tratar con los 2 factores haremos un diagrama de *boxplot* y así poder comparar las medianas de llevar o no llevar $extras1$ y $extras2$ para visualizar si puede haber diferencias significativas entre los grupos [Figura 2.2].

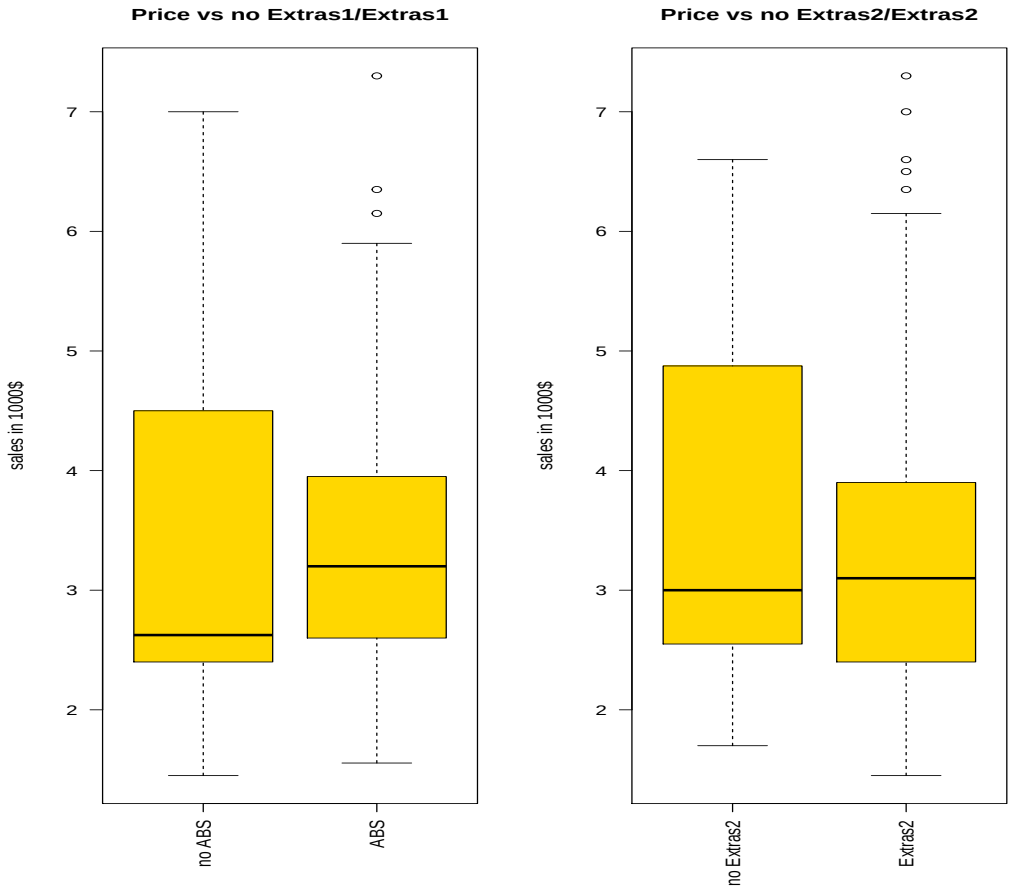


Figura 2.2: Vemos que los coches con ABS son un poco más caros que los que no lo tienen, pero sin una diferencia significativa; podemos apreciar también que los coches con y sin techo solar tienen medianas similares, por lo que tampoco hay una diferencia significativa.

Por tanto, después de esta inspección visual, decidimos que las variables *age* y *kilometer* deben estar en nuestro modelo de regresión lineal. Ahora haremos un estudio de los distintos modelos combinando las restantes variables explicativas.
Este estudio nos va a servir también para los ejemplos con los demás criterios.
Nos quedaremos con el modelo que mayor *coeficiente de determinación corregido o ajustado* tenga.
Consideramos los 8 modelos posibles:

- mod1: $price = kilometer + age$.

- mod2: $price = kilometer + age + extras1$.
- mod3: $price = kilometer + age + extras2$.
- mod4: $price = kilometer + age + TIA$.
- mod5: $price = kilometer + age + extras1 + extras2$.
- mod6: $price = kilometer + age + extras1 + TIA$.
- mod7: $price = kilometer + age + extras2 + TIA$.
- mod8: $price = kilometer + age + extras1 + extras2 + TIA$.

Para ver los *coeficiente de determinación corregido o ajustado* de manera ordenada y así seleccionar el mejor según este criterio, los colocamos en una matriz de la siguiente forma:

```
> nf <- 8
> nc <- 1
> resRaj <- matrix(nrow=nf, ncol=nc, byrow=TRUE)
> rownames(resRaj)<-c("mod1","mod2","mod3", "mod4", "mod5", "mod6",
+ "mod7", "mod8")
> colnames(resRaj)<-c("adj.r.squared")
> resRaj[,1] <- c(summary(mod1)$adj.r.squared,
+ summary(mod2)$adj.r.squared,
+ summary(mod3)$adj.r.squared,
+ summary(mod4)$adj.r.squared,
+ summary(mod5)$adj.r.squared,
+ summary(mod6)$adj.r.squared,
+ summary(mod7)$adj.r.squared,
+ summary(mod8)$adj.r.squared)
> resRaj
     adj.r.squared
mod1     0.6105227
mod2     0.6154363
mod3     0.6083116
mod4     0.6085515
mod5     0.6131700
mod6     0.6141034
mod7     0.6062884
mod8     0.6117905
```

Por lo que el mejor modelo, según este criterio es el modelo 2, $price \sim kilometer + age + extras1$.

2.2. Coeficiente C_p de Mallows

Para calcular el estadístico de Mallows, partamos de un modelo $\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{u}$, ajustamos $\mathbf{Y}$ por otro modelo alternativo en el que se consideran p variables, el modelo ajustado en este caso se denota como $\boldsymbol{Y} = \widetilde{\boldsymbol{X}}\widetilde{\boldsymbol{\beta}} + \boldsymbol{u}$ y sean $\widehat{\boldsymbol{Y}}_{\boldsymbol{p}}$ las predicciones del modelo estimado. Veamos cual es su ECM:

$$ECM[\widehat{\boldsymbol{Y}}_{\boldsymbol{p}}] = E[(\widehat{\boldsymbol{Y}}_{\boldsymbol{p}} - \boldsymbol{Y})'(\widehat{\boldsymbol{Y}}_{\boldsymbol{p}} - \boldsymbol{Y})]$$

$$= E[(\widehat{\boldsymbol{Y}}_{\boldsymbol{p}} - \boldsymbol{X\beta})'(\widehat{\boldsymbol{Y}}_{\boldsymbol{p}} - \boldsymbol{X\beta})] \underbrace{=}_{\pm E[\widehat{\boldsymbol{Y}}_{\boldsymbol{p}}]} E[(\widehat{\boldsymbol{Y}}_{\boldsymbol{p}} - E[\widehat{\boldsymbol{Y}}_{\boldsymbol{p}}])'(\widehat{\boldsymbol{Y}}_{\boldsymbol{p}} - E[\widehat{\boldsymbol{Y}}_{\boldsymbol{p}}])]$$

$$+E[(E[\widehat{\boldsymbol{Y}}_{\boldsymbol{p}} - \boldsymbol{X\beta}]'(E[\widehat{\boldsymbol{Y}}_{\boldsymbol{p}}] - \boldsymbol{X\beta})] = Var(\widehat{\boldsymbol{Y}}_{\boldsymbol{p}}) + (sesgo[\widehat{\boldsymbol{Y}}_{\boldsymbol{p}}])^2.$$

Calculemos primero la varianza, tenemos:

$$\widehat{\boldsymbol{Y}}_{\boldsymbol{p}} = \widetilde{\boldsymbol{X}}(\widetilde{\boldsymbol{X}}'\widetilde{\boldsymbol{X}})^{-1}\widetilde{\boldsymbol{X}}'\boldsymbol{Y} = \widetilde{\boldsymbol{X}}(\widetilde{\boldsymbol{X}}'\widetilde{\boldsymbol{X}})^{-1}\widetilde{\boldsymbol{X}}'(\boldsymbol{X\beta} + \boldsymbol{u}) \Rightarrow E[\widehat{\boldsymbol{Y}}_{\boldsymbol{p}}] = \widetilde{\boldsymbol{X}}(\widetilde{\boldsymbol{X}}'\widetilde{\boldsymbol{X}})^{-1}\widetilde{\boldsymbol{X}}'\boldsymbol{X\beta}$$

luego,

$$(\widehat{\boldsymbol{Y}}_{\boldsymbol{p}} - E[\widehat{\boldsymbol{Y}}_{\boldsymbol{p}}])'(\widehat{\boldsymbol{Y}}_{\boldsymbol{p}} - E[\widehat{\boldsymbol{Y}}_{\boldsymbol{p}}]) = \boldsymbol{u}\widetilde{\boldsymbol{X}}(\widetilde{\boldsymbol{X}}'\widetilde{\boldsymbol{X}})^{-1}\widetilde{\boldsymbol{X}}'\boldsymbol{X}(\widetilde{\boldsymbol{X}}'\widetilde{\boldsymbol{X}})^{-1}\widetilde{\boldsymbol{X}}'\boldsymbol{u}$$

$$= \boldsymbol{u}\widetilde{\boldsymbol{X}}(\widetilde{\boldsymbol{X}}'\widetilde{\boldsymbol{X}})^{-1}\widetilde{\boldsymbol{X}}'\boldsymbol{u} \sim \sigma^2\chi^2_p. \Rightarrow Var(\widehat{\boldsymbol{Y}}_{\boldsymbol{p}}) = E[\sigma^2\chi^2_p].$$

Y ahora calculemos el sesgo, veamos que:

$$E[\underbrace{(\boldsymbol{Y} - \widehat{\boldsymbol{Y}}_{\boldsymbol{p}})'(\boldsymbol{Y} - \widehat{\boldsymbol{Y}}_{\boldsymbol{p}})}_{\sum_{i=1}^n (y_i - \hat{y}_{(p_i)})^2}] = \underbrace{E[(\boldsymbol{X\beta} - \widetilde{\boldsymbol{X}}(\widetilde{\boldsymbol{X}}'\widetilde{\boldsymbol{X}})^{-1}\widetilde{\boldsymbol{X}}'\boldsymbol{X\beta})'(\boldsymbol{X\beta} - \widetilde{\boldsymbol{X}}(\widetilde{\boldsymbol{X}}'\widetilde{\boldsymbol{X}})^{-1}\widetilde{\boldsymbol{X}}'\boldsymbol{X\beta})]}_{(sesgo[\widehat{\boldsymbol{Y}}_{\boldsymbol{p}}])^2}$$

$$+E[\boldsymbol{u}'(\boldsymbol{I} - \widetilde{\boldsymbol{X}}(\widetilde{\boldsymbol{X}}'\widetilde{\boldsymbol{X}})^{-1}\widetilde{\boldsymbol{X}}')\boldsymbol{u}].$$

Por tanto,

$$(sesgo[\widehat{\boldsymbol{Y}}_{\boldsymbol{p}}])^2 = E\left[\sum_{i=1}^n (y_i - \hat{y}_{p_i})^2\right] - E[\sigma^2\chi^2_{n-p}].$$

Teniendo la varianza y el sesgo calculados:

$$ECM[\widehat{\boldsymbol{Y}}_{\boldsymbol{p}}] = E[\sigma^2\chi^2_p] + E\left[\sum_{i=1}^n (y_i - \hat{y}_{p_i})^2\right] - E[\sigma^2\chi^2_{n-p}]$$

$$= \sigma^2 p + E\left[\sum_{i=1}^n (y_i - \hat{y}_{(p)i})^2\right] - \sigma^2(n-p)$$

dividiendo ambos lados por σ^2:

$$\frac{ECM[\widehat{\boldsymbol{Y}}_{\boldsymbol{p}}]}{\sigma^2} = E\left[\frac{\sum_{i=1}^{n}(y_i - \hat{y}_{p_i})^2}{\sigma^2}\right] - n + 2p.$$

Queremos minimizar esta expresión, o lo que es lo mismo:

$$\text{mín}\, E\left[\frac{\sum_{i=1}^{n}(y_i - \hat{y}_{p_i})^2}{\sigma^2}\right] + 2p.$$

Puesto que no conocemos el valor de σ, lo más que podemos hacer es utilizar su estimador, y lo que nos queda es lo que se denomina coeficiente C_p *de* $Mallows$:

$$C_p = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_{p_i})^2}{\hat{\sigma}^2} + 2p.$$

Esta estimación sólo es posible cuando $\hat{\sigma}^2 \simeq \sigma^2$, y esto lo conseguimos teniendo una muestra lo suficientemente grande y $\hat{\sigma}^2 = \frac{SCR_{n-p-k}}{n-p-k}$, estando los p regresores necesarios incluidos en los $(p+k)$, para que el estimador sea insesgado. Cuando el modelo es el adecuado, $C_p \approx p$, nos quedaremos con el modelo que tenga el más próximo a p y preferiblemente menor.

2.2.1. Aplicación

Usando la base de datos considerada en el apartado 2.1.1 y teniendo en cuenta el estudio previo realizado en el mismo, vamos a calcular los distintos modelos para ver cúal tiene menor C_p.
Para ello utilizaremos la librería *circular* y la librería *wle* de *R* con su función *mle.cp*, la cual calcula todos los posibles modelos y halla su coeficiente C_p para finalmente mostrar los 4 mejores modelos.
Lo primero es leer los datos y cargar las librerías:

```
> golf <- read.table("golffull.txt", header=TRUE)
> attach(golf)
> library(circular)
> library(wle)
```

Ahora indicamos el modelo con todas las variables para que la función *mle.cp* pueda trabajar, dicha función almacena en la variable *cp* los 4 modelos anteriormente mencionados:

```
> cp<-mle.cp(price~kilometer+age+extras1+extras2+TIA, data=golf)
> cp
```

```
Call:
mle.cp(formula = price ~ kilometer + age + extras1 + extras2 +
    TIA, data = golf)


Mallows Cp:
     (Intercept) kilometer age extras1 extras2 TIA    cp
[1,]           1         1   1       1       0   0 2.422
[2,]           1         1   1       1       0   1 4.005
[3,]           1         1   1       1       1   0 4.407
[4,]           1         1   1       1       1   1 6.000

Printed the first  4  best models
```

En nuestro caso, como tenemos que seleccionar los modelos cuyas variables *age* y *kilometer* estén presentes, hemos tenido que hacer alguna modificación

```
> misres<- subset(cp$cp, cp$cp[,2]==1 & cp$cp[,3]==1)
> head(misres)
     (Intercept) kilometer age extras1 extras2 TIA         cp
[1,]           0         1   1       0       0   0 622.527918
[2,]           1         1   1       0       0   0   3.551917
[3,]           0         1   1       1       0   0 590.044015
[4,]           1         1   1       1       0   0   2.422241
[5,]           0         1   1       0       1   0 608.241325
[6,]           1         1   1       0       1   0   5.505495
```

Vemos que los modelos que manejamos contienen a las variables deseadas. Ahora debemos escoger el que tenga menor *cp*, con lo que vamos a ordenarlos:

```
> ordenado<-misres[ order(misres[,7]), ]
> head(ordenado)
     (Intercept) kilometer age extras1 extras2 TIA       cp
[1,]           1         1   1       1       0   0 2.422241
[2,]           1         1   1       0       0   0 3.551917
[3,]           1         1   1       1       0   1 4.005040
[4,]           1         1   1       1       1   0 4.406582
[5,]           1         1   1       0       0   1 5.401711
[6,]           1         1   1       0       1   0 5.505495
```

En esta tabla se muestran los 6 mejores modelos, según este criterio, que tienen las variables explicativas impuestas por nosotros.
Luego el mejor modelo será: $price \sim intercept + kilometer + age + extras1$.

2.3. Validación cruzada

El primer paso y común a todos los tipos de validaciones cruzadas es dividir el conjunto de datos que tenemos en dos tipos de conjuntos:

- Un tipo de conjunto que nos sirve para estimar los parámetros del modelo, llamado *training set*.
- Un tipo de conjunto de validación que sirve para valorar la capacidad predictiva del modelo, llamado *testing set*.

Pasemos a ver las dos validaciones cruzadas más usadas.

2.3.1. Validación cruzada en r iteraciones

La validacion cruzada en r iteraciones (*r-fold cross validation*) comienza agrupando los datos en r subconjuntos de tamaño similar. Uno se utiliza para validar, y los restantes $(r-1)$ se consideran como conjuntos de estimación. Repetiremos este proceso r veces, una vez con cada uno de los subconjuntos de validación. Nos quedamos con el modelo en el que la suma de los errores de predicción al cuadrado sea más pequeño, es decir:

$$\text{mín}\{CV\} \quad \text{donde} CV = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_{iM})^2\}.$$

Normalmente se suele utilizar el 10-*fold cross validation* ó 5-*fold cross validation*, dependiendo del tamaño de datos que tengamos.

2.3.2. Validación crudada dejando uno fuera

Un caso importante del anterior, que cabe destacar, es la validación cruzada dejando una sola observación fuera (*leave-one-out cross validation*). En este caso, el error es muy pequeño, en cambio, el coste computacional es elevado, pues hay que calcular n iteraciones y analizar para cada iteración los datos de ambos conjuntos. El estadístico para decidir en este caso es:

$$CV = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_{iM}^{-i})^2$$

donde, $\hat{y}_{iM}^{-i} \equiv$ estimación cuando se ha eliminado la observación i-ésima.
Se tiene una expresión sencilla para este coeficiente, sin tener que rehacer todos los cálculos, basándonos en los $\hat{y}_{iM}$ originales:

$$CV = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - \hat{y}_{iM}}{1 - h_{iiM}}\right)^2$$

donde $1 - h_{ii,M}$ son los elementos diagonales de la matriz *hat*.
Es importante destacar que ambos tipos de validación tienen ciertas limitaciones:

- El *training set* y el *testing set* deben ser extraídos de la misma población, en caso de no serlo, la validación no produciría resultados significativos.

- Esta herramienta no es válida cuando tenemos un sistema que evoluciona con el tiempo, pues podría darse el caso de que ambos conjuntos mencionados anteriormente sufrieran cambios sistemáticos, por ejemplo: si tenemos un modelo que utilizamos para predecir el valor de las acciones, el cual ha sido calculado en un *training set* en un periodo de tiempo determinado, éste no será eficiente a la hora de predecir el valor de la misma población en el siguiente periodo de tiempo.

- Es importante notar que en los datos del *training set* debemos evitar que haya algún dato que esté también en el *testing set*.

2.3.3. Aplicación

Para ilustrar este método procederemos inicialmente de manera similar a como se hizo en 2.1.1 y teniendo en cuenta el estudio previo realizado en él, vamos a calcular los distintos modelos para ver cúal tiene menor CV.
Para calcular los distintos CV usaremos el paquete *lattice*, necesario para el paquete $DAAG$, el cual contiene a la función $CVlm$. Dicha función realiza las $m\text{-}fold\ cross\ validation$ según el valor que le asignemos a m.
Comencemos leyendo los datos, definiendo los modelos y cargando los paquetes necesarios:

```
> golf <- read.table("golffull.txt", header=TRUE)
> attach(golf)
> mod1 <- lm(price~kilometer+age, data=golf)
> mod2 <- lm(price~kilometer+age+extras1, data=golf)
> mod3 <- lm(price~kilometer+age+extras2, data=golf)
> mod4 <- lm(price~kilometer+age+TIA, data=golf)
> mod5 <- lm(price~kilometer+age+extras1+extras2, data=golf)
> mod6 <- lm(price~kilometer+age+extras1+TIA, data=golf)
> mod7 <- lm(price~kilometer+age+extras2+TIA, data=golf)
> mod8 <- lm(price~kilometer+age+extras1+extras2+TIA, data=golf)
> library(lattice)
> library(DAAG)
```

Una vez que hemos escrito los argumentos en la función, ésta nos muestra la tabla con el análisis de la varianza, los resultados de cada $fold$, y al final la suma de los errores al cuadrado (MS).

```
> RES1<-CVlm(data=golf, form.lm=mod1, m=10)
Analysis of Variance Table

Response: price
           Df Sum Sq Mean Sq F value Pr(>F)
kilometer   1   88.1    88.1     146 <2e-16 ***
age         1   75.2    75.2     124 <2e-16 ***
Residuals 169  102.2     0.6
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1



fold 1
Observations in test set: 17
               9     11    16    45   50    55   62    64     73
Predicted   4.93  4.502 4.557  4.17 2.96  4.12 5.12  4.88  3.756
cvpred      4.92  4.488 4.544  4.16 2.95  4.11 5.12  4.88  3.746
price       6.35  3.823 4.950  2.90 4.20  3.99 6.15  4.50  3.000
CV residual 1.43 -0.665 0.406 -1.26 1.25 -0.12 1.03 -0.38 -0.746
                85   101   118    125    148   150  155    164
Predicted   3.5638  4.31 2.827 3.8879  2.210 2.749  3.22  2.721
cvpred      3.5542  4.31 2.817 3.8861  2.202 2.744  3.22  2.721
price       3.6500  3.20 3.800 3.9500  1.900 3.100  4.20  2.400
CV residual 0.0958 -1.11 0.983 0.0639 -0.302 0.356  0.98 -0.321

Sum of squares = 11.1    Mean square = 0.65    n = 17

 .
 .
 .

fold 10
Observations in test set: 17
               5    10   13    20     31    32    34    40   43
Predicted   5.19 5.383 4.94  3.88  3.405  3.89 4.445 4.558 5.49
cvpred      5.11 5.274 4.87  3.90  3.444  3.88 4.391 4.489 5.34
```

```
price       6.20 5.900 5.55  2.50  3.250  2.40 4.600 5.450 7.00
CV residual 1.09 0.626 0.69 -1.40 -0.194 -1.48 0.209 0.961 1.66
               44     76     83   110    117    119     133    137
Predicted    3.33  3.690  3.185 3.766 2.7887  2.334  3.0203  2.594
cvpred       3.37  3.676  3.211 3.722 2.8252  2.411  3.0215  2.632
price        3.70  2.800  2.600 3.900 2.9000  1.450  2.9990  1.950
CV residual 0.33 -0.876 -0.611 0.178 0.0748 -0.961 -0.0225 -0.682

Sum of squares = 12.7    Mean square = 0.75    n = 17

Overall (Sum over all 17 folds)
   ms
0.617
```

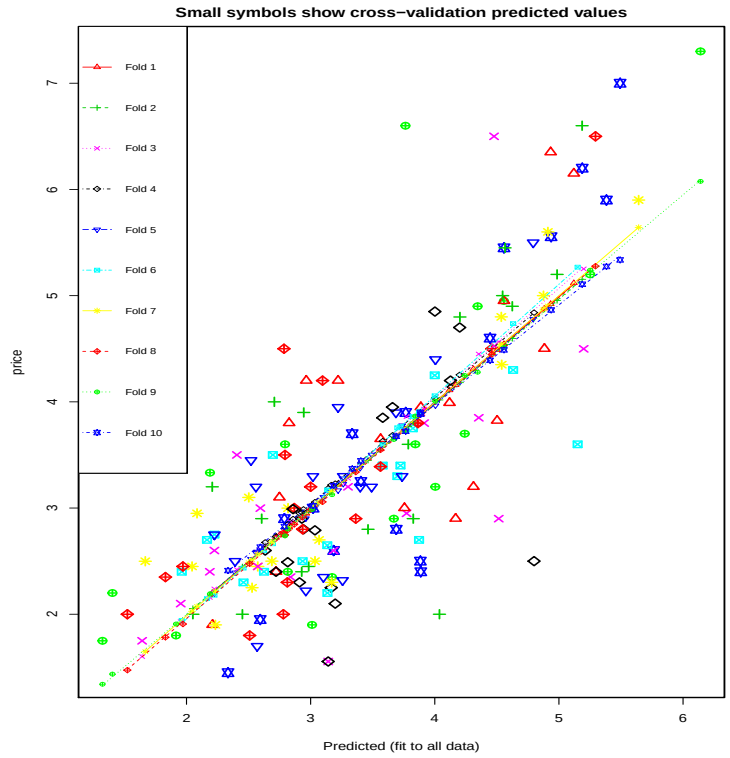


Figura 2.3: Cada recta de regresión está calculada a partir del subconjunto de entrenamiento de la base de datos que asignamos con la CV.

Realizamos la misma operación para los restantes 7 modelos, y colocamos en una matriz la suma de los errores de predicción al cuadrado, que es lo que nos interesa:

```
> nf <- 8
> nc <- 1
> resCV <- matrix(nrow=nf, ncol=nc, byrow=TRUE)
> rownames(resCV)<-c("mod1","mod2","mod3", "mod4",
+ "mod5", "mod6", "mod7", "mod8")
> colnames(resCV)<-c("ms")
> resCV[,1] <- c(attr(RES1, "ms"),
+ attr(RES2, "ms"),
+ attr(RES3, "ms"),
+ attr(RES4, "ms"),
+ attr(RES5, "ms"),
+ attr(RES6, "ms"),
+ attr(RES7, "ms"),
+ attr(RES8, "ms"))
> resCV
        ms
mod1 0.617
mod2 0.615
mod3 0.624
mod4 0.628
mod5 0.620
mod6 0.627
mod7 0.634
mod8 0.632
```

Por lo que el mejor modelo según la validación cruzada realizada es: modelo 2, $price \sim kilometer + age + extra1$.

2.4. Criterio de Información de Akaike

El criterio de información de Akaike, AIC, es un criterio relacionado con el criterio C_p *de* $Mallows$, aunque más general. La idea principal del AIC es maximizar la *log-verosimilitud* esperada de un modelo determinado, a través del EMV. El hecho de que se denomine criterio de información es porque está íntimamente relacionado con la llamada información de *Kullback - Leibler*. Este criterio no busca encontrar el mejor modelo, sino encontrar el modelo, de entre los que compiten, que mejor se ajuste a los datos con los que trabajamos.
Está definido por:

$$AIC(M) = -2ln\ \mathcal{L}(\hat{\boldsymbol{\beta}}_M, \hat{\sigma}^2) + 2(|M| + 1) \tag{2.1}$$

donde $ln\ \mathcal{L}(\hat{\boldsymbol{\beta}}_M, \hat{\sigma}^2)$ es el máximo valor del logaritmo de la función verosimilitud evaluado en $\hat{\boldsymbol{\beta}}_M$, donde $\hat{\boldsymbol{\beta}}_M$ es el EMV del modelo M, $\hat{\sigma}^2 = \frac{\sum(\boldsymbol{y}_i - \hat{\boldsymbol{y}}_i)^2}{n}$ es el EMV de σ^2, y $(|M| + 1)$ el número total de parámetros en el modelo incluyendo σ^2.
El primer término de la expresión 2.1 es una medida de bondad de ajuste, pues disminuye al aumentar $\hat{\boldsymbol{\beta}}_M$, y el segundo término es una penalización por el número de parámetros, exactamente igual que los términos del C_p *de Mallows*.
La generalidad que tiene el criterio AIC, viene del hecho de que podemos calcularlo siempre que tengamos una función de verosimilitud.
Entre los modelos que compiten, es mejor el que tenga el menor AIC.

Proposición 2.4.1. *Si se tiene normalidad y σ^2 es conocida entonces el criterio C_p es equivalente al criterio AIC.*

Demostración. Comencemos reescribiendo la expresión del AIC dada en 2.1:

$$ln\ \mathcal{L}(\hat{\boldsymbol{y}}_i, \hat{\boldsymbol{\beta}}, \hat{\sigma}^2) = -\frac{n}{2}ln(2\pi) - \frac{n}{2}ln(\hat{\sigma}^2) - \frac{n}{2\sum(\boldsymbol{y}_i - \hat{\boldsymbol{y}}_i)^2}(\boldsymbol{y}_i - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y}_i - \boldsymbol{X}\boldsymbol{\beta})$$

$$= -\frac{n}{2}ln(2\pi) - \frac{n}{2}ln(\hat{\sigma}^2) - \frac{n}{2}$$

y excluyendo términos que no dependen del número de parámetros p llegamos a:

$$-2ln\ \mathcal{L} + 2(|M| + 1) = n\ ln(\hat{\sigma}_p^2) + 2p$$

con $p = (|M| + 1)$.
Ahora, suponiendo conocida σ^2, observamos que minimizar el AIC es análogo a minimizar:

$$n\ ln\left(\frac{\hat{\sigma}_p^2}{\sigma^2}\right) + 2p$$

que puede escribirse, suponiendo normalidad ($\hat{\sigma}_p^2 \simeq \sigma^2$)

$$n\ ln\left(1 + \frac{\hat{\sigma}_p^2 - \sigma^2}{\sigma^2}\right) + 2p \underbrace{\simeq}_{ln(1+x)\simeq x \text{ para x pequeños}} \frac{n\hat{\sigma}_p^2}{\sigma^2} - n + 2p.$$

Luego, sustituyendo σ^2 por un estimador, quedándonos con los términos que dependen del número de parámetros p e ignorando las constantes aditivas en las que no aparece p, pues sólo dependen de n, llegamos a:

$$AIC \simeq \frac{\sum e_{(p)}^2}{\hat{\sigma}^2} + 2p = C_p$$

donde $\sum e^2_{(p)}$ es la suma de cuadrados de los residuos del modelo con p parámetros.

□

Como consecuencia de la Proposición 2.4.1, se obtienen expresiones más sencillas del AIC.

Corolario 2.4.1. *En el modelo lineal con errores gaussianos:*

$$-2ln\ \mathcal{L}(\hat{\boldsymbol{\beta}}_M, \hat{\sigma}^2) = nlog(\hat{\sigma}^2) + \frac{1}{\hat{\sigma}^2}(\boldsymbol{y}-\boldsymbol{X}_M\hat{\boldsymbol{\beta}}_M)'(\boldsymbol{y}-\boldsymbol{X}_M\hat{\boldsymbol{\beta}}_M)$$

$$= nlog(\hat{\sigma}^2) + \frac{n\hat{\sigma}^2}{\hat{\sigma}^2} = nlog(\hat{\sigma}^2) + n$$

donde $\hat{\sigma}^2$ es el EMV de σ^2.
Podemos ignorar la constante n, pues no depende de p, y escribir:

$$AIC = nlog(\hat{\sigma}^2) + 2(|M| + 1).$$

donde $\hat{\sigma}^2$ no es el estimador insesgado de σ^2.

Nota 2.4.1. *Las principales ventajas que tiene el criterio AIC, por lo cual es tan práctico son:*

- *No requiere de ninguna tabla para ver el correspondiente valor.*
- *Tiene una fácil implementación.*
- *No necesita un nivel de significación arbitrario para elegir entre dos modelos.*

Nota 2.4.2. *El término de penalización no depende del tamaño de la muestra, es decir, que el número de parámetros que seleccionamos con este criterio, es el mismo tanto para una muestra pequeña como para una grande. Lo que hace que AIC no sea consistente, es decir, que no se aproxima al modelo correcto conforme aumenta la muestra, como cabría esperar.*

2.4.1. Aplicación

Partimos de los mismos modelos expuestos en el Ejemplo 2.1.1. A diferencia de otros métodos, para usar el AIC no necesitamos ninguna librería extra de R, puesto que ya viene implementado en la función AIC.
Lo único que tendremos que hacer es cargar la base de datos y los modelos:

```
> golf <- read.table("golffull.txt", header=TRUE)
> attach(golf)
> mod1 <- lm(price~kilometer+age, data=golf)
> mod2 <- lm(price~kilometer+age+extras1, data=golf)
> mod3 <- lm(price~kilometer+age+extras2, data=golf)
> mod4 <- lm(price~kilometer+age+TIA, data=golf)
> mod5 <- lm(price~kilometer+age+extras1+extras2, data=golf)
> mod6 <- lm(price~kilometer+age+extras1+TIA, data=golf)
> mod7 <- lm(price~kilometer+age+extras2+TIA, data=golf)
> mod8 <- lm(price~kilometer+age+extras1+extras2+TIA, data=golf)
```

Y ahora, basta con calcular los AIC de los modelos y compararlos para ver cual es el mayor.
Colocaremos los resultados en una matriz como en ejemplos anteriores para verlo mejor:

```
> nf <- 8
> nc <- 1
> resAIC <- matrix(nrow=nf, ncol=nc, byrow=TRUE)
> rownames(resAIC)<-c("mod1","mod2","mod3", "mod4",
+ "mod5", "mod6", "mod7", "mod8")
> colnames(resAIC)<-c("AIC")
> resAIC[,1] <- c(AIC(mod1),
+ AIC(mod2),
+ AIC(mod3),
+ AIC(mod4),
+ AIC(mod5),
+ AIC(mod6),
+ AIC(mod7),
+ AIC(mod8))
> resAIC
          AIC
mod1 406.5904
mod2 405.3859
mod3 408.5433
mod4 408.4380
mod5 407.3697
mod6 406.9542
mod7 410.4026
mod8 408.9490
```

Por lo que el mejor modelo según este criterio sería el modelo 2, $price \sim kilometer + age + extra1$.

2.5. Criterio de Información Bayesiana

Schwarz (1978) ideó el criterio de información bayesiana, BIC, a raíz de la inconsistencia del estimador AIC, en éste se considera el tamaño de la muestra n en el término de penalización. Por lo tanto, fue diseñado con el objetivo de ser consistente, es decir, que a medida que el tamaño muestral aumenta, el criterio tiende a seleccionar el verdadero modelo que genera los datos.
BIC se basa en estudiar el comportamiento de la probabilidad *a posteriori* del modelo j-ésimo. Suponiendo ciertas hipótesis de las distribuciones *a priori* de los parámetros, tenemos:

$$ln\ f(\boldsymbol{X}|M_j) = \mathcal{L}_j(\hat{\boldsymbol{\beta}}_{\boldsymbol{j}}|\boldsymbol{X}) + ln\ P(\hat{\boldsymbol{\beta}}_{\boldsymbol{j}}|M_j) + \frac{p_j}{2}ln(2\pi) - \frac{p_j}{2}ln(n) + \frac{1}{2}ln|R_j|$$

donde $f(\boldsymbol{X}|M_j)$ es la verosimilitud marginal de los datos en el modelo M_j, $\mathcal{L}_j(\hat{\boldsymbol{\beta}}_{\boldsymbol{j}}|\boldsymbol{X})$ es la función de verosimilitud del modelo M_j, que tiene como EMV de $\boldsymbol{\beta}_j$ a $\hat{\boldsymbol{\beta}}_j$, $P(\hat{\boldsymbol{\beta}}_{\boldsymbol{j}}|M_j)$ es la probabilidad a *priori* de los parámetros, p_j es el número de parametros estimados y R_j es igual a nS_j, siendo S_j la matriz de covarianzas de $\hat{\boldsymbol{\beta}}_j$.
Haciendo tender n a infinito, se puede aproximar

$$ln\ f(\boldsymbol{X}|M_j) \simeq \mathcal{L}_j(\hat{\boldsymbol{\beta}}_{\boldsymbol{j}}|\boldsymbol{X}) - \frac{p_j}{2}ln(n)$$

que es equivalente a:

$$BIC(M_j) = -2ln\ \mathcal{L}_j(\hat{\boldsymbol{\beta}}_j|\boldsymbol{X}) + ln(n)(p_j).$$

Con lo que finalmente llegamos a la expresión siguiente:

$$BIC(M) = -2ln\ \mathcal{L}(\hat{\boldsymbol{\beta}}_{|M|+1}) + ln(n)(|M|+1). \tag{2.2}$$

Según este criterio, se obtiene el mejor modelo calculando los distintos BIC y quedándonos con el que tenga el menor.
Notemos que si añadimos más parámetros en el modelo, el ajuste se verá incrementado, pues el primer término mide la desviación del modelo estimado con el modelo saturado (con todas las variables), pero este efecto se compensa con el segundo término pera evitar el sobreajuste.

Corolario 2.5.1. *Bajo el supuesto de errores gaussianos, la expresión 2.2 se reduce a:*

$$BIC(M) = nlog(\hat{\sigma}^2) + log(n)(|M|+1).$$

2.5.1. Aplicación

Por último daremos el ejemplo de BIC. Como en los anteriores casos, consideramos los modelos introducidos en 2.1.1. En este caso tampoco hace falta ninguna librería adicional de R, pues también viene implementado el método en la función BIC.
Procedemos entonces de manera similar que en el Ejemplo 2.4.1

```
> golf <- read.table("golffull.txt", header=TRUE)
> attach(golf)
> mod1 <- lm(price~kilometer+age, data=golf)
> mod2 <- lm(price~kilometer+age+extras1, data=golf)
> mod3 <- lm(price~kilometer+age+extras2, data=golf)
> mod4 <- lm(price~kilometer+age+TIA, data=golf)
> mod5 <- lm(price~kilometer+age+extras1+extras2, data=golf)
> mod6 <- lm(price~kilometer+age+extras1+TIA, data=golf)
> mod7 <- lm(price~kilometer+age+extras2+TIA, data=golf)
> mod8 <- lm(price~kilometer+age+extras1+extras2+TIA, data=golf)
> nf <- 8
> nc <- 1
> resBIC <- matrix(nrow=nf, ncol=nc, byrow=TRUE)
> rownames(resBIC)<-c("mod1","mod2","mod3", "mod4", "mod5", "mod6",
>                     "mod7", "mod8")
> colnames(resBIC)<-c("BIC")
> resBIC[,1] <- c(BIC(mod1),
+ BIC(mod2),
+ BIC(mod3),
+ BIC(mod4),
+ BIC(mod5),
+ BIC(mod6),
+ BIC(mod7),
+ BIC(mod8))
> resBIC
          BIC
mod1 419.1804
mod2 421.1234
mod3 424.2808
mod4 424.1755
mod5 426.2547
mod6 425.8391
mod7 429.2876
mod8 430.9814
```

Por lo que el mejor modelo, con este criterio es el modelo 1, $price \sim kilometer + age$.

2.6. Comparación de criterios

En esta sección realizamos un estudio comparativo de los distintos criterios previamente introducidos. Comenzamos con R^2_{aj}, el cual penaliza muy poco la inclusión de nuevas variables con lo que tenderá al sobreajuste y la sobreparametrización, pues para maximizarlo basta con introducir variables para los que el estadístico Q_h sea mayor que 1, basándonos en el Teorema 2.1.1, lo que ocurre con probabilidad 0.5, ver [9]. Además, puede tomar valores negativos (cuando el estadístico F toma valores inferiores a 1) lo que puede causar problemas a la hora de su interpretación.
Por lo tanto, no es recomendable usar este criterio sin otro que lo complemente.
Para utilizar el *coeficiente C_p de Mallows* debemos suponer que entre los modelos que compiten está el correcto, lo que no tiene porqué ocurrir y es uno de sus inconvenientes, ver *Peña, D.* [2].
Podemos ver que C_p es más restrictivo que R^2_{aj}, pues para minimizar C_p sólo se incluirá un nuevo regresor si reduce en $2\hat{\sigma}^2$ veces la SCR. En cambio con el R^2_{aj}, incluiremos un nuevo regresor, con la intención de maximizarlo, si disminuye en $\hat{\sigma}^2$ veces la SCR, ver [9].
En cuanto a los dos últimos criterios, AIC y BIC, bajo ciertas condiciones, son similares.

Observación (comparación *AIC* y *BIC*):

Suponiendo normalidad, tenemos que el BIC está definido por:

$$BIC(p) = n\ ln(\hat{\sigma}^2) + p\ ln(n)$$

por tanto, podemos dividir ambos términos por n y considerar:

$$ln(\hat{\sigma}^2) + p\ \frac{ln(n)}{n}$$

supongamos ahora que tenemos dos modelos con p_1 y p_2 variables, con $k \leq p_1 < p_2$, entonces:

$$BIC(p_1) = ln(\hat{\sigma}^2_{p_1}) + p_1\ \frac{ln(n)}{n}$$

$$BIC(p_2) = ln(\hat{\sigma}^2_{p_2}) + p_1\ \frac{ln(n)}{n}$$

definamos ahora $\nabla BIC = BIC(p_2) - BIC(p_1)$, luego:

$$\nabla BIC = ln(\hat{\sigma}^2_{p_2}) - ln(\hat{\sigma}^2_{p_1}) + (p_2 - p_1)\ \frac{ln(n)}{n}$$

con lo que elegiremos el modelo con p_1 variables cuando $\nabla BIC \geq 0$.
De manera análoga se puede proceder con el AIC, obteniendo:

$$\nabla AIC = ln(\hat{\sigma}^2_{p_2}) - ln(\hat{\sigma}^2_{p_1}) + (p_2 - p_1)\ \frac{2}{n}$$

luego, cuando $ln(n) = 2$ ambos criterios serán similares, esto es cuando $n = 8$. Cuando $n > 8$ el BIC penaliza más la inclusión de variables irrelevantes que el AIC, luego es mejor.

2.7. Problemas en la regresión múltiple

2.7.1. Error de especificación

El error de especificación ocurre cuando omitimos variables explicativas importantes, introducimos variables explicativas innecesarias o suponemos que hay una relación lineal entre alguna variables explicativa y la variable respuesta cuando no la hay.
Si especificamos mal las variables, entonces los residuos tendrán esperanza no nula y afectará a todas las propiedades del modelo en cuestión. El hecho de omitir variables relevantes nos lleva a obtener un sesgo en los parámetros estimados. La varianza de la estimación de los parámetros se verá aumentada, lo que puede hacer que no detectemos alguna variable significativa.
En cambio, si lo que hacemos es incluir en el modelo variables irrelevantes. La varianza de los estimadores aumentará mucho si la variable que incluimos está muy correlada con las que ya estaban. Si por el contrario la variable incluida está incorrelada con las que estaban, este efecto no ocurrirá, sin embargo, los estimadores no serán eficientes pues hemos invertido un grado de libertad en estimar un parámetro totalmente innecesario.
Si suponemos que hay una relación lineal cuando no la hay, tendremos problemas al hacer predicciones fuera del marco de la muestra pues una relación no lineal se puede aproximar por una lineal en un entorno reducido, con lo que el modelo puede ser correcto en ese rango de valores de la muestra, pero no fuera.
Hay varias maneras de identificar estos errores, con gráficos de residuos respecto a distintas medidas, por ejemplo, respecto a: los valores estimados, las variables explicativas o variables candidatas a ser significativas.

Para resolver el tema de la no linealidad se puede recurrir a las transformaciones, por ejemplo, considerar:

$$y^{\lambda_0} = \beta_0 + \beta_1 x_1^{\lambda_1} + ... + \beta_k x_k^{\lambda_k} + u$$

siendo λ el parámetro de la transformación de *Box-Cox*. Hay varias maneras de hallar estos $k+1$ parámetros:

- Maximizando la función de verosimilitud respecto a ellos.
- Estimar únicamente λ_0 usando la máxima verosimilitud y después buscar las demás transformaciones mediante un algoritmo *(Box y Tidwell [1962]*.
- Mediante métodos gráficos e imponiendo que el modelo sea coherente y simple.

Otra manera de resolver la no linealidad es establecer una relación no lineal, $y = m(x_1, ..., x_k) + u$, y estimar los parámetros de $m(x_i)$ por la máxima verosimilitud o por los mínimos cuadrados.
También podemos recurrir a técnicas no paramétricas, aunque este método no se puede aplicar directamente cuando tenemos un número alto de variables explicativas, lo que se soluciona añadiendo hipótesis a $m(x_i)$, por ejemplo, aditividad, es decir:

$$m(x_1, ..., x_k) = g_1(x_1) + ... + g_k(x_k).$$

2.7.2. Hipótesis de normalidad

Cuando tenemos una muestra no normal, los estimadores obtenidos por el método de mínimos cuadrados no coinciden con los máximo verosímiles y dejan de ser eficientes, por lo que no podemos extraer toda la información que obtendríamos de la muestra si ésta fuera normal. Además cuando nuestra muestra no es normal los EMV suelen ser no lineales.
La falta de normalidad puede deberse a algunas observaciones extrañas, lo que nos llevaría a estudiar la influencia de éstas sobre el modelo, o a una fuerte asimetría.
Si tenemos normalidad, los residuos se comportarán como una combinación lineal de variables normales, ya que:

$$e_i = y_i - \hat{y}_i.$$

Se pueden expresar usando la *matriz de predicción o hat matrix* de forma:

$$\boldsymbol{e = y - \hat{y} = (I - H)y = (I - H)(X\beta + U)} \underbrace{=}_{\boldsymbol{HX=X}} \boldsymbol{(I - H)U}$$

$$= u_i - \sum_{j=1}^{n} h_{ij} u_j.$$

Usando el teorema central de límite, $e_i \underbrace{\sim}_{aprox.} N$, aunque las perturbaciones no sigan una distribución normal.

Cuando tenemos una muestra grande comparada con el número de parámetros, podemos suponer que los residuos son independientes para aplicarles contrastes de normalidad. Uno de los test más usados está basado en la asimetría y la curtosis de los residuos.

El coeficiente de asimetría de los residuos está definido por:

$$a = \frac{\sum \frac{e_i^3}{n}}{\hat{s}_r^3}$$

siendo $\hat{s}_r$ la varianza residual. Para muestras grandes, aproximadamente, $a \sim N(0, \frac{6}{n})$.

Y el coeficiente de curtosis es:

$$k = \frac{\sum \frac{e_i^4}{n}}{\hat{s}_r^4}$$

que para muestras grandes, $k \sim N(3, \frac{24}{n})$.

Bajo hipótesis de normalidad, ambos estadísticos son independientes, y podemos realizar el contraste:

$$J = n\left(\frac{a^2}{6} + \frac{(k-3)^2}{24}\right)$$

con $J \sim \chi_2^2$. *Bera y Jarque (1980)* propusieron este contraste.

La falta de normalidad en los residuos puede deberse a varias causas:

- Hay alguna observación atípica, con lo que deberemos estudiar si se debe a la omisión de variables explicativas en el modelo o sino, si éstas tienen una influencia grande en el modelo.

- Tienen una fuerte asimetría, en este caso deberíamos transformar la variable respuesta, pues es frecuente que la no linealidad y la falta de normalidad vayan unidas. O bien tenemos una muestra lo suficientemente grande como para aplicar el teorema central del límite y suponer normalidad. Si por el contrario tenemos una muestra pequeña, o bien renunciamos a realizar contrastes de significación, o bien podemos suponer un modelo de distribución para la perturbación que sea consistente con la muestra y basar los contrastes aproximados en los EMV del modelo.

Es importante insistir en la importancia de averiguar las causas de la no normalidad para detectar errores o deficiencias en nuestro modelo, puesto que si se han incluido todas las variables explicativas relevantes, por el teorema central del límite, es de suponer que la perturbación siga una distribución normal.

2.7.3. Robustez

Es bastante frecuente que en nuestra muestra haya observaciones atípicas. Estas observaciones tienen verdadera importancia puesto que pueden tener un gran peso en la estimación y también pueden indicarnos: errores de medición, omisión de variables explicativas relevantes, etc.
Es importante en un modelo estudiar hasta qué punto sus propiedades básicas son debidas a toda la muestra y no a un pequeño subconjunto de ella. Pues tendremos más confianza en un modelo cuyas propiedades se deduzcan de toda la muestra que no en uno que lo haga de unos pocos valores. Este análisis es el estudio de la robustez de un modelo y tiene dos partes: el estudio de la robustez del diseño de recogida de datos (robustez a priori); y el estudio de la robustez de los parámetros estimados (robustez a posteriori).
De la robustez a priori cabría destacar los efectos palanca de las observaciones

Definición 2.7.1. *Se denominan* ***puntos palanca*** *a aquellos puntos que tengan un alto valor en el peso de la pendiente de la recta de regresión definidos en los conceptos previos.*

El efecto palanca de cada observación es la capacidad de un punto de atraer la pendiente de la recta de regresión. Es importante notar que estos puntos pueden no detectarse en los gráficos de residuos anteriormente mencionados, pues puede ocurrir que las coordenadas del punto atípico, individualmente, estén cercanas a las medias y, sin embargo, sus coordenadas, conjuntamente, difieren mucho del resto de puntos.
Pasamos ahora a comentar la robustez a posteriori.
Un detalle a tener en cuenta es que el hecho de que una observación sea muy influyente a priori, no implica que lo sea realmente. Lo sería en el caso de que si la eliminásemos las propiedades del modelo variaran mucho. Se prueba que podemos considerar un punto influyente a nivel α si al eliminarlo obtenemos un valor $\hat{\beta}_{(i)}$ no admisible a nivel α, es decir, si $\hat{\beta}_{(i)}$ no está incluido en la región de confianza

$$\frac{(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})'\boldsymbol{X}'\boldsymbol{X}(\hat{\boldsymbol{\beta}}\boldsymbol{\beta})}{(k+1)\hat{s}_r^2} \leq F(k+1; n-k-1; 1-\alpha).$$

Consideraremos un dato como atípico cuando no se genere por el mismo procedimiento que el resto de la muestra. Por ejemplo, cuando haya un error de medida o si esa observación tiene un valor diferente del resto para una variable explicativa relevante omitida en el modelo. En ese caso, el modelo para esa observación sería: $y_i = \boldsymbol{x_i'}\boldsymbol{\beta} + \omega + u_i$, donde ω es el error de medida o el efecto de la variable explicativa omitida. Podemos modelizar el dato atípico como un desplazamiento en la media de la distribución. O alternativamente, como un desplazamiento en la varianza, de manera que la observación se genera con nuestro modelo, pero la varianza en ese punto será $c^2\sigma^2$ con $c >> 1$. Ambos modelos son equivalentes, pues con un sólo dato no es posible saber si la media o la varianza ha cambiado. Un dato atípico puede o no ser influyente, y viceversa.
Para buscar los valores atípicos se calculan los residuos estudentizados, $\hat{t}_j$, en todos los puntos. Pues el estadístico t asociado a $\hat{\omega}$ es precisamente $\hat{t}_j$.

Definición 2.7.2. *Se define el* ***residuo estudentizado*** *como:*

$$\hat{t}_i = \frac{e_i}{\hat{s}_{r_i}\sqrt{1 - h_{ii}}}$$

Para ver si existen valores atípicos se toma el máximo de los residuos estudentizados con H_0 : *todos los datos han sido generados por el mismo modelo*, éste seguirá la distribución del máximo de una variable t de Student, que depende de los grados de libertad de t y está tabulada.
Otro método para ver ésto es el *método de Bonferroni* y utilizar contrastes múltiples. Dicho procedimiento es simple y general, pero no es siempre óptimo. Se basa en la *desigualdad de Bonferroni*. Y se utiliza de la forma siguiente:
Sea c el número de comparaciones que construimos, sea $\overline{A}_i$ el suceso: *aceptar* $\mu_i \neq \mu_j$ *cuando realmente* $\mu_i = \mu_j$. Supongamos que hacemos las comparaciones de medias con un nivel de significación α:

$$P(\overline{A}_i) = \alpha.$$

Sea $B = \overline{A}_1 + \overline{A}_2 + ... + \overline{A}_c$. Los sucesos $\overline{A}_i$ no son mutuamente excluyentes, por tanto:

$$P(B) = P(\overline{A}_1 + \overline{A}_2 + ... + \overline{A}_c) \leq \sum P(\overline{A}_i) = c\alpha.$$

El método pretende garantizar un error de tipo I total para el conjunto de contrastes, α_T, por lo que $P(B) \leq \alpha_T$. Ésto se consigue calculando cada contraste individual a un nivel α de manera que:

$$\alpha = \frac{\alpha_T}{c}.$$

Lo que nos lleva a un procedimiento de aproximación bastante útil en la práctica.
Cuando c es grande, se necesitan niveles de significación muy pequeños, tanto que no están tabulados, por lo que se utiliza la aproximación:

$$t_\nu^\alpha \simeq \left(1 - \frac{z_\alpha + 1}{4\nu}\right)^{-1}$$

donde ν son los grados de libertad de t y z_α el valor de la distribución normal estándar (0,1) tal que $P(z \geq z_\alpha) = \alpha$.
Aunque si tenemos en nuestra muestra un grupo de datos atípicos pueden no detectarse con los procemientos vistos hasta ahora. A pesar de eliminar uno de los puntos, al haber otros parecidos hace que el punto eliminado no parezca influyente. Este fenómeno se llama *enmascaramiento* y se resuelve con la estimación robusta y técnicas más avanzadas.

2.7.4. Heterocedasticidad

Decimos que existe heterocedasticidad en las perturbaciones u_i cuando no se puede aplicar la hipótesis:

$$Var(u_i) = \sigma^2, \quad i = 1, .., n$$

con lo que inclumplimos una de las hipótesis básicas donde se asienta la regresión lineal.
En este caso las observaciones con la varianza baja son importantes, pues son más fiables a la hora de estimar la recta de regresión que las observaciones con varianza alta (en general, cuanto menor es su varianza, menos se desvían del valor medio que queremos estimar), y deberían tener más peso. Pero el método de mínimos cuadrados no tiene en cuenta esto, por lo que los estimadores calculados con este procedimiento dejan de ser eficientes y las fórmulas deducidas para calcular las varianzas de los estimadores ya no son correctas, por lo tanto, los contrastes basados en ellas dejan de ser válidos.
La pérdida de eficiencia de los estimadores depende de la magnitud de heterocedasticidad. Podemos medirla calculando el cociente entre la varianza máxima y la mínima de las observaciones, *Bloch y Moses, (1988)* recomiendan que cuando el cociente es menor que dos, podemos seguir utilizándolos puesto que la pérdida de eficiencia es pequeña. Cuando es mayor que dos, la pérdida de eficiencia es grande.
Si además de heterocedasticidad tenemos observaciones con alto efecto palanca, las consecuencias se agravan, pues también es más complicado estimar las perturbaciones del modelo, con lo que es más difícil estimar la varianza

de la muestra.
Para reconocer la heterocedasticidad basta con analizar los residuos. Mediante el gráfico de $e_i = f(\hat{y}_i)$ se puede detectar, y para identificar si la heterogeneidad en la variabilidad es debida a alguna variable explicativa podemos usar $e_i = f(x_i)$.
Uno de los contrastes para la heterocedasticidad es el de la razón de verosimilitudes. Para aplicar este contraste, dividimos los residuos, e_i en g grupos, cada uno de un tamaño n_i y estimamos la varianza en cada uno de ellos. Sea $\hat{\sigma}_i^2$ la estimación de la varianza del grupo i, y σ_i^2 el EMV de la varianza de los residuos. Entonces tenemos el contraste:
$H_0 : e_i \sim N(0, \sigma)$
$H_1 : e_i \sim N(0, \sigma_i)$, con g valores distintos de σ_i
luego el logaritmo de la razón de verosimilitudes de ambas hipótesis es:

$$log(\lambda) = -\sum_{i=1}^{g} \frac{n_i}{2} log(\hat{\sigma}_i^2) - \sum_{i=1}^{g} \frac{n_i}{2} - \left(-\frac{n}{2} log(\hat{\sigma}^2) - \frac{n}{2}\right)$$

por tanto,

$$2log(\lambda) = nlog(\hat{\sigma}^2) - \sum_{i=1}^{g} n_i log(\hat{\sigma}_i^2)$$

cuya distribución asintótica es χ^2_{g-1}.
El contraste anterior no tiene en cuenta la posibilidad de que los residuos sean sesgados por la heterocedasticidad. Para realizar un contraste más exacto en muestras pequeñas, tenemos el siguiente test:
$H_0 : y_i = \boldsymbol{x'}_i\boldsymbol{\beta} + u_i, \qquad u_i \sim N(0, \sigma)$
$H_1 : y_i = \boldsymbol{x'}_i\boldsymbol{\beta} + u_i, \qquad u_i \sim N(0, \sigma_i)$
que hace que las regresiones sean calculadas por separado en cada grupo al estimar las varianzas, $\hat{\sigma}_i^2$. Una vez definido este contraste, se procede de manera análoga al anterior.
El problema más básico que produce la heterocedasticidad es la formulación errónea del modelo. Por ejemplo, si nuestro modelo fuera:

$$y = kx_1^{\alpha_1} \cdot ... \cdot x_k^{\alpha_k} \cdot u$$

donde u sigue una distribución log-normal de media 1 y varianza desconocida. Estimamos por un modelo lineal $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + ... + \hat{\beta}_k x_k$, los residuos tendrán falta de normalidad, falta de linealidad y heterocedasticidad, aumentando la varianza de los errores conforme aumentan los valores de las variables explicativas. En este caso, deberíamos transformar la variable objetivo, y, con logaritmos. La heterocedasticidad más frecuente es que varianza aumente linealmente con el valor de $\overline{y}$. Aquí también se resuelve usando los logaritmos.

Si estamos en el caso donde la heterocedasticidad viene por una variable explicativa, x_k, y la desviación típica aumenta linealmente con dicha variable, el procedimiento a seguir es ajustar el siguiente modelo:

$$\frac{\hat{y}}{x_k} = \frac{\hat{\beta}_0}{x_k} + \hat{\beta}_1 \frac{x_1}{x_k} + \hat{\beta}_k + \frac{u}{x_k}$$

donde la perturbación ahora sí tiene varianza constante.
Otra herramienta útil para solucionar los problemas de heterocedasticidad es la de *mínimos cuadrados generalizados.*
Partamos de un modelo con heterocedasticidad en el que suponemos que:

$$E[\boldsymbol{U}\boldsymbol{U'}] = \sigma^2 \boldsymbol{G}$$

donde $\boldsymbol{G}$ es una matriz simétrica y definida positiva. En el caso que nos ocupa, para que las perturbaciones sean heterocedásticas, se supone que $\boldsymbol{G}$ es una matriz diagonal.
Ahora tenemos que diferenciar dos casos, cuando $\boldsymbol{G}$ sea conocida y cuando no.
1.- Si $\boldsymbol{G}$ es conocida entonces, $\boldsymbol{Y} \sim N_n(\boldsymbol{X\beta}, \sigma^2 \boldsymbol{G})$ y podremos estimar los parámetros por el método de máxima verosimilitud. Que es equivalente a hacer una transformación de las variables con el fin de que cumplan las hipótesis del modelo de regresión y luego aplicar los resultados ya dados.
Como $\boldsymbol{G}$ se supone conocida y definida positiva, podemos obtener una matriz simétrica, no singular, $\boldsymbol{A}$ tal que $\boldsymbol{G} = \boldsymbol{AA}$. Esta $\boldsymbol{A}$ se denomina matriz raíz cuadrada de $\boldsymbol{G}$ y en nuestro caso, su diagonal son los términos σ_i/σ. Multiplicando por la inversa de $\boldsymbol{A}$ nuestro sistema, tenemos que:

$$\boldsymbol{A}^{-1}\boldsymbol{Y} = \boldsymbol{A}^{-1}\boldsymbol{X\beta} + \boldsymbol{A}^{-1}\boldsymbol{U}.$$

Esta expresión puede reescribirse como:

$$\boldsymbol{Y}^* = \boldsymbol{X}^*\boldsymbol{\beta} + \boldsymbol{U}^*$$

con $\boldsymbol{Y}^* = \boldsymbol{A}^{-1}\boldsymbol{Y}$, $\boldsymbol{X}^* = \boldsymbol{A}^{-1}\boldsymbol{X}$ y $\boldsymbol{U}^* = \boldsymbol{A}^{-1}\boldsymbol{U}$. Observamos que esta nuevas variables están relacionadas entre ellas con el mismo $\boldsymbol{\beta}$.
Luego la nueva matriz de covarianzas es:

$$E[\boldsymbol{U}^*\boldsymbol{U}^{*\prime}] = \boldsymbol{A}^{-1}E[\boldsymbol{U}\boldsymbol{U'}]\boldsymbol{A}^{-1} = \sigma^2\boldsymbol{I}.$$

Con lo que queda arreglado el problema de la heterocedasticidad, pues el modelo $\boldsymbol{Y}^* = \boldsymbol{X}^*\boldsymbol{\beta} + \boldsymbol{U}^*$, es homocedástico, y ya podríamos aplicar el método

de *mínimos cuadrados* (que coincide con el de máxima verosimilitud) para calcular un estimador de $\boldsymbol{\beta}$, que será:

$$\hat{\boldsymbol{\beta}}_G = (\boldsymbol{X}^{*\prime}\boldsymbol{X}^*)^{-1}\boldsymbol{X}^{*\prime}\boldsymbol{Y}^* = \underbrace{(\boldsymbol{X}'\boldsymbol{G}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{G}^{-1}\boldsymbol{Y}}_{\boldsymbol{G}^{-1}=\boldsymbol{A}^{-1}\boldsymbol{A}^{-1}}$$

y se denomina *estimador de mínimos cuadrados generalizados* o MCG y tiene como matriz de covarianzas:

$$Var(\hat{\boldsymbol{\beta}}_G) = \sigma^2(\boldsymbol{X}*'\,\boldsymbol{X}*)^{-1} = \sigma^2(\boldsymbol{X}'\boldsymbol{G}^{-1}\boldsymbol{X})^{-1}.$$

2.- Veamos ahora el caso donde $\boldsymbol{G}$ es desconocida.
En general, no es posible resolver el caso en que todos los valores de la matriz $\boldsymbol{G}$ son desconocidos.
Lo habitual es suponer alguna estructura para $\boldsymbol{G}$, modelizar esta matriz introduciendo parámetros desconocidos adicionales de forma que el problema planteado sea tratable y utilizar métodos iterativos de estimación para el vector de parámetros $\boldsymbol{\beta}$ y los nuevos parámetros utilizados para modelizar la estructura de $\boldsymbol{G}$. Detalles adicionales pueden verse en *Peña, D.* [2] (*Cap. 9*).

2.7.5. Multicolinealidad

La multicolinealidad se da cuando las variables explicativas tienen una dependencia entre ellas fuerte, por tanto, es muy difícil ver el efecto que tiene cada una individualmente en la variable respuesta.
Este problema viene del hecho de intentar extraer más información de los datos que lo que contienen, por lo que dicho problema reside en la base de datos y no en el modelo.
En los modelos de regresión múltiple, para estimar el efecto de una variable explicativa debemos fijarnos en la parte de la variable que no está relacionada linealmente con las demás del modelo. En el caso de que sí lo estuviera no sería posible estimar su efecto, a esto se le llama el problema de la multicolinealidad.
Cuando nos disponemos a estimar los parámetros de los modelos de regresión, es necesario invertir la matriz $\boldsymbol{X}'\boldsymbol{X}$. Si tenemos una variable linealmente dependiente con el resto, la matriz $\boldsymbol{X}$ tendrá un rango menor que $k+1$, que es el número de parámetros, el determinante de $\boldsymbol{X}'\boldsymbol{X}$ será 0, por lo que la matriz no tendrá inversa y el sistema de ecuaciones determinado por los parámetros del modelo tendrá infinitas soluciones.
Puede darse también que las variables estén altamente correladas, sin ser exactamente combinación lineal de ninguna, en ese caso habrá una multicolinealidad alta, por ejemplo, en el caso de que tuvieramos dos variables

explicativas en nuestro modelo, x_1, x_2 con medias nulas, tal que:

$$\boldsymbol{X'X} = \begin{bmatrix} \sum x_1^2 & \sum x_1 x_2 \\ \sum x_1 x_2 & \sum x_2^2 \end{bmatrix} = n \begin{bmatrix} s_1^2 & s_{12} \\ s_{12} & s_2^2 \end{bmatrix}$$

invirtiendo la matriz y utilizando que $s_{12} = r s_1 s_2$ y $|\boldsymbol{X'X}| = s_1^2 s_2^2 (1 - r^2)$, tenemos:

$$(\boldsymbol{X'X})^{-1} = \frac{1}{n} \begin{bmatrix} \frac{1}{s_1^2(1-r^2)} & \frac{-r}{s_1 s_2(1-r^2)} \\ \frac{-r}{s_1 s_2(1-r^2)} & \frac{1}{s_2^2(1-r^2)} \end{bmatrix}.$$

Luego las varianzas de los estimadores serán:

$$Var(\hat{\beta}_i) = \frac{\sigma^2}{n s_i^2 (1 - r^2)}, \quad i = 1, 2$$

por tanto, cuando $r^2 \sim 1$ la varianza de los coeficientes estimados será muy alta. Además, las estimaciones tendrán una gran dependencia entre ellas, pues:

$$Cov(\hat{\beta}_1, \hat{\beta}_2) = \frac{-r\sigma^2}{n s_1 s_2 (1 - r^2)}.$$

El coeficiente de correlacion entre $\hat{\beta}_1$ y $\hat{\beta}_2$ será igual en valor absoluto, pero de signo contrario, a la correlación entre las variables explicativas, es decir

$$r(\hat{\beta}_1, \hat{\beta}_2) = \frac{Cov(\hat{\beta}_1, \hat{\beta}_2)}{\sqrt{Var(\hat{\beta}_1)}\sqrt{Var(\hat{\beta}_2)}} = -r.$$

Luego, las estimaciones serán tan dependientes entre sí, como lo sean las variables entre ellas.
En general, la varianza de un coeficiente de regresión es

$$Var(\hat{\beta}_i) = \sigma^2 / SCR(x_{i,R})$$

siendo $SCR(x_{i,R}) = \sum_{j=1}^{n} (x_{ij} - \hat{x}_{ij,R})^2$ la varianza residual de una regresión de x_i sobre el resto. Se tiene también que

$$SCR(x_{i,R}) = SCT(x_i) - SCE(x_{i,R}) = n s_i^2 (1 - R_{i,R}^2).$$

Llamando $R_{i,R}$ al coeficiente de correlación múltiple en la regresión de x_i en función del resto de variables, tenemos:

$$Var(\hat{\beta}_i) = \frac{\sigma^2}{n s_i^2 (1 - R_{i,R}^2)}$$

por lo que si el cuadrado del coeficiente de correlación es cercano a 1, la varianza será muy grande.
Para averiguar si tenemos o no multicolinealidad debemos examinar:

- La matriz de correlación entre las variables explicativas, $\boldsymbol{R}$, y $\boldsymbol{R}^{-1}$.
- Los factores de inflación de la varianza.
- Las raíces y vectores característicos de las matrices $\boldsymbol{X}'\boldsymbol{X}$, o $\boldsymbol{R}$.

Si tenemos una correlación alta entre variables explicativas es una clara señal de multicolinealidad. Puede ser que haya una relación perfecta entre una de las variables explicativas y el resto y, sin embargo, sus coeficientes de correlación sean bajos. Por ejemplo, supongamos las variables explicativas: $x_1, ..., x_k$ con media cero, varianza uno y ortogonales. Y definamos una nueva variable que sea la media de las anteriores, $x_{k+1} = (x_1, ..., x_k)/k$. Luego, $Var(x_{k+1}) = 1/k$ y $Cov(x_i, x_{k+1}) = 1/k$, con lo que su correlación es $1/\sqrt{k}$. Si k es grande, la correlación será pequeña, pero un modelo que incluya las $k+1$ variables $(x_1, ..., x_k, x_{k+1})$ tendrá una multicolinealidad exacta.
Sea $\boldsymbol{R}$ la matriz de correlación de las variables explicativas, la cual es cuadrada, simétrica de orden k y cuyo término (ij) es el coeficiente de correlación lineal simple entre x_i y x_j. Esta matriz, para dos variables sería:

$$\boldsymbol{R} = \begin{bmatrix} 1 & r \\ r & 1 \end{bmatrix}$$

luego $\boldsymbol{R}^{-1}$ es:

$$\boldsymbol{R}^{-1} = \begin{bmatrix} \frac{1}{1-r^2} & \frac{-r}{1-r^2} \\ \frac{-r}{1-r^2} & \frac{1}{1-r^2} \end{bmatrix}$$

podemos ver que los elementos de la diagonal, $1/(1-r^2)$, contienen al coeficiente de correlación. Para k variables, los elementos de la diagonal serían $1/(1-R^2_{i,R})$, siendo $R^2_{i,R}$ el coeficiente de correlación múltiple de la variable explicativa x_i con el resto de variables explicativas.
Por tanto, si tenemos elementos de la diagonal de $\boldsymbol{R}^{-1}$ grandes, nos indicará que hay alta multicolinealidad. En este caso no tenemos el problema que teníamos con los elementos de $\boldsymbol{R}$, donde podía no detectarse a simple vista la multicolinealidad, pues en los elementos de la diagonal de la matriz inversa se tienen en cuenta todas las variables explicativas, y entonces se detectará la multicolinealidad cuando una de las variables sea casi combinación lineal del resto. Aunque $\boldsymbol{R}^{-1}$ también tiene inconvenientes, cuando la matriz $\boldsymbol{R}$ sea casi singular, no podremos calcular su inversa con precisión.
Los términos de la diagonal de $\boldsymbol{R}^{-1}$ se interpretan como el aumento de la variabilidad en la estimación de los efectos de cada variable explicativa en la regresión múltiple, como consecuencia de la dependencia entre las variables, respecto a la regresión simple.
Veámoslo para dos variables explicativas de media cero:

La varianza de las estimaciones de los efectos de las variables mediante regresiones simples sería $\hat{s}_r^2(i)/s_i^2 n$, con $s_r^2(i)$ la varianza residual de la regresión simple que tiene por regresor la variable x_i. Si estimamos los efectos mediante regresión múltiple, la varianza sería $\hat{s}_r^2/s_i^2(1-r^2)n$. Por tanto:

$$\frac{Var(\text{efecto } x_i|\text{R. múltiple})}{Var(\text{efecto } x_i|\text{R. simple})} = \frac{\hat{s}_r^2(i)}{\hat{s}_r^2}\frac{1}{1-r^2}$$

dicha expresión nos indica que el cambio de la varianza de un coeficiente al pasar de la regresión simple a la regresión múltiple depende de dos factores. Uno, el cambio de la varianza residual de la regresión, que será mayor en la simple que en la múltiple, normalmente. Y dos, el $1/(1-r^2)$, denominado *factor de inflación de la varianza*, el cual mide el aumento de la varianza debido a la dependencia entre las variables.
Por ejemplo, al introducir una nueva variable explicativa en la regresión simple la cual esté muy correlada con la que ya había y no ayuda a explicar la variable objetivo, hace que el primer término, $\hat{s}_r^2(i)/\hat{s}_r^2$, esté cercano a 1 y la varianza del coeficiente de la primera variable estará multiplicada por el mencionado factor de inflación.
Se puede probar que el resultado anterior se puede generalizar como sigue:

$$\frac{Var(\text{efecto } x_i|\text{R. múltiple})}{Var(\text{efecto } x_i|\text{R. simple})} = \frac{\hat{s}_r^2(i)}{\hat{s}_r^2}FIV(i)$$

donde $FIV(i) = 1/(1-R_{i,R}^2)$ es el factor de inflación de la varianza.
Cuando $\boldsymbol{X'X}$ o $\boldsymbol{R}$ son singulares debemos recurrir a otras técnicas para tratar la multicolinealidad. Por ejemplo: el *índice de condicionamiento*, denotado por IC, el cual nos sirve para estos casos y está definido por:

$$IC = \sqrt{\frac{\text{máximo autovalor de la matriz}}{\text{mínimo autovalor de la matriz}}} \geq 1.$$

Normalmente se calcula este índice para $\boldsymbol{R}$ en vez de para $\boldsymbol{X'X}$, puesto que ésta no está afectada por las escalas de los regresores, pues en el caso de que un regresor tuviera una varianza grande y otro muy pequeña, la matriz $\boldsymbol{X'X}$ estaría mal condicionada, y por tanto, fuera de la diagonal tendría términos nulos.
Por convenio se admite que existe alta multicolinealidad cuando $IC > 30$. Cuando $10 < IC < 30$ tendremos una multicolinealidad moderada. Y en caso contrario tendremos bien definida la matriz y la multicolinealidad será lo suficientemente baja para no alterar la estimación por el método de mínimos cuadrados del modelo.

Antes de pasar a ver como solucionar la multicolinealidad, veamos el efecto de ésta en el error cuádratico medio, lo que nos será útil para ver una de sus soluciones.
Tenemos que el error cuádratico medio de $\hat{\boldsymbol{\beta}}$ está definido por:

$$ECM(\hat{\boldsymbol{\beta}}) = E[(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})'(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})]$$

$$= \sum_{i=0}^{k} (\hat{\beta}_i - \beta_i)^2 = \sigma^2 tr(\boldsymbol{X}'\boldsymbol{X})^{-1} = \sigma^2 \sum_{i=0}^{k} \frac{1}{\lambda_i}$$

donde los λ_i son los valores propios de la matriz $\boldsymbol{X}'\boldsymbol{X}$. Si esta matriz es casi singular, $\lambda_i \simeq 0$ para algún i, lo que lleva a tener un error cuadrático medio muy grande. Una vez visto esto, pasemos a mostrar como solucionar la multicolinealidad. Ésta no tiene solución sencilla pues, como mencionamos al principio de la sección, el problema reside en la muestra. Una de la alternativas es tomar las observaciones de manera que la matriz $\boldsymbol{X}'\boldsymbol{X}$ sea diagonal, lo que reduce la varianza de los estimadores. En caso de no poder diseñar la manera de recabar los datos, podemos eliminar regresores altamente correlados con otros, haciendo menor el número de parámetros a estimar, aunque dichos estimadores serán sesgados. Esta es una de las soluciones más simples. Veamos la manera de proceder.
Sea

$$y = \beta_1 x_1 + \beta_2 x_2 + u$$

nuestro modelo, donde vamos a suponer que las variables explicativas tienen media cero, por simplificar. Según lo visto anteriormente, $Var(\hat{\beta}_1) = \sigma^2/ns_1^2(1 - r_{12}^2)$.
Por lo que su error cuadrático medio es:

$$ECM(\hat{\beta}_1) = Var(\hat{\beta}_1).$$

Si eliminamos la variable explicativa x_2, nos queda el modelo:

$$y = b_1 x_1 + \varepsilon$$

por lo que la estimación de b_1 será:

$$\hat{b}_1 = \frac{\sum y x_1}{\sum x_1^2}$$

veamos que en efecto, es sesgada:

$$E[\hat{b}_1] = \frac{1}{\sum x_1^2} E\left[\beta_1 \sum x_1^2 + \beta_2 \sum x_2 x_1 + u \sum x_1\right]$$

$$= \beta_1 + \beta_2 \frac{\sum x_2 x_1}{\sum x_1^2} = \beta_1 + \beta_2 r_{12} \frac{s_2}{s_1}$$

por lo que, sí, es sesgado. Calculando su varianza, obtenemos:

$$Var(\hat{b}_1) = \frac{\sigma^2}{\sum x_1^2} = \frac{\sigma^2}{n s_1^2}.$$

Y por tanto, su error cuadrático medio será:

$$ECM(\hat{b}_1) = \left(\beta_2 r_{12} \frac{s_2}{s_1}\right)^2 + \frac{\sigma^2}{n s_1^2}$$

por lo que debería verificarse que $ECM(\hat{b}_1) < ECM(\hat{\beta}_1)$, es decir:

$$\frac{\beta_2^2 r_{12}^2 n s_2^2 + \sigma^2}{n s_1^2} < \frac{\sigma^2}{n s_1^2 (1 - r_{12}^2)}$$

de lo que se deduce que,

$$\frac{1}{1 - r_{12}^2} > \left(\frac{\beta_2}{\sigma}\right)^2 n s_2^2$$

luego cuando $r_{12} \simeq 1$, el $ECM(\hat{b}_1)$ será menor que $ECM(\hat{\beta}_1)$ y obtendremos una estimación mejor (aunque sesgada) del efecto de la variable explicativa x_1 eliminando de nuestro modelo la variable explicativa x_2.

Nota 2.7.1. *Reordenando la última expresión obtenemos un resultado bastante interesante:*

$$1 > \frac{\beta_2^2 n s_2^2 (1 - r_{12}^2)}{\sigma^2} = \frac{\beta_2^2}{Var(\hat{\beta}_2)} = \left(\frac{\beta_2}{DT(\hat{\beta}_2)}\right)^2$$

donde $DT(\hat{\beta}_2)$ es la desviación típica de $\hat{\beta}_2$. Sustituyendo los parámetros β_2 y σ^2 por sus estimaciones, obtenemos el estadístico t, al cuadrado, que se utiliza para contrastar si el parámetro es cero.

Teniendo en cuenta la Nota 2.7.1, eliminaremos de nuestro modelo las variables cuyo estadístico t sea menor que 1, para así tratar de mejorar el error cuadrático medio de estimación de los parámetros restantes y eliminar la multicolinealidad.
En vez de eliminar directamente las variables de nuestro modelo, podemos crear una nueva variable que agrupe las que están muy correladas entre sí.

Por ejemplo, supongamos que tenemos tres variables explicativas muy correladas. Podríamos crear una nueva variable que fuera la media aritmética de las tres, y sustituir en el modelo las tres variables por su media. Con esto ya no habría problema de dependencia, tendríamos un modelo con la misma capacidad predictiva y, además, más simple.
Si tenemos un número grande de variables correladas, podríamos intentar agruparlas en distintas variables. Las llamadas *componentes principales* se basan en esta idea. Supongamos ahora que tenemos k variables muy dependientes entre sí y de media cero. Creemos unas nuevas variables explicativas que sean combinaciones lineales de las que teníamos y que cumplan las siguientes propiedades:

- Son capaces de resumir la información contenida en las variables originales de manera óptima.
- Son ortogonales entre sí.

Sea $\boldsymbol{X}$ la matriz de los datos. Con lo que nos queda:

$$\boldsymbol{Z} = \boldsymbol{X}\boldsymbol{A}$$

siendo $\boldsymbol{A}$ una matriz de dimensiones $k \times k$ cuyas columnas son los vectores propios de la matriz $\boldsymbol{X'X}$, que supondremos normalizados a módulo uno. A estas nuevas variables $\boldsymbol{Z}$, de media cero, las llamaremos *componentes principales* de $\boldsymbol{X}$ y vendrán dadas por:

$$\boldsymbol{Z_i} = \boldsymbol{X}\boldsymbol{a_i}$$

donde $\boldsymbol{a_i}$ es el autovector asociado al autovalor λ_i. Las nuevas variables creadas están incorreladas, pues la matriz $\boldsymbol{X'X}$ es simétrica, sus vectores propios son ortogonales, por lo que si λ_i son los valores propios, $Cov(\boldsymbol{Z_i}\boldsymbol{Z_j}) = \boldsymbol{a_i'}\boldsymbol{X'}\boldsymbol{X}\boldsymbol{a_j} = \lambda_i\boldsymbol{a_i'}\boldsymbol{a_j} = 0$. Y la varianza de éstas será:

$$Var(\boldsymbol{Z_i}) = E[\boldsymbol{a_i'}\boldsymbol{X'}\boldsymbol{X}\boldsymbol{a_i}] = E[\lambda_i\boldsymbol{a_i'}\boldsymbol{a_i}] = \lambda_i.$$

Luego la variable asociada al mayor valor propio será la combinación lineal de mayor varianza, que es la primera componente principal, y así sucesivamente. Por tanto, en lugar de calcular nuestro modelo con las k variables explicativas, podemos calcularlo con las r componentes principales, pues $r < k$, asociadas con los r valores mayores de los λ_i. De esta manera obtendremos estimadores no centrados, pero posiblemente con menor varianza.

Capítulo 3

Métodos heurísticos para la selección de variables

En la práctica para seleccionar el modelo, podemos seguir las siguientes pautas:

- En base a la experiencia previa, fijar una serie de posibles modelos, pueden ser diferentes en cuanto al número de variables explicativas que consideran, y también en el tipo de las relaciones que se están proponiendo (por ejemplo, relaciones lineales frente a no lineales). Debe hacerse de modo que el número total de modelos bajo consideración sea tan pequeño como sea posible.
- Todos los modelos potenciales los evaluamos con alguno de los criterios anteriores. Como regla, nos puede ocurrir que haya un número de modelos que compiten y proporcionan aproximadamente la misma bondad de ajuste del modelo, y que difieren sólo en pequeños aspectos unos de otros.

Sin embargo, estas diferencias pueden causar ciertas incertidumbres en cuanto a las conclusiones que se obtengan de esos modelos. Aunque, estas recomendaciones prácticas no pueden seguirse siempre pues el número de variables regresoras puede ser muy grande y los modelos a considerar muy diferentes.
En este caso, podemos utilizar los métodos llamados parcialmente heurísticos. Aunque éstos tampoco están exentos de polémica, pues según el método utilizado, es posible que de lugar a modelos totalmente distintos. Por lo que la utilización de estos métodos nunca debe sustituir al estudio del estadístico, ni se puede considerar que las variables presentes en el modelo resultante son las principales variables explicativas de la variable respuesta.

3.1. Selección hacia delante

El método de la *selección hacia delante (o forward selection)* se basa en partir de un modelo inicial, con una sola variable, e ir incluyendo el resto, una a una, en cada iteración del algoritmo.
El algoritmo es el siguiente:

1. Escogemos como variable inicial la que esté más correlada con la variable respuesta, y, por ejemplo x_1.

2. Calculamos el modelo de regresión lineal simple entre x_1 e y, y las correlaciones parciales entre $(x_2, ..., x_k)$ e y habiendo eliminado de la variable respuesta el efecto de la variable x_1.

3. Introducimos la variable más correlada con y, sea esta variable x_2.

4. Calculamos de nuevo el modelo de regresión lineal con ambas variables y comprobamos si el estadístico t para el coeficiente de x_2, $\hat{\beta}_2$, es significativo. Si no lo es, excluímos del modelo la última variable y termina el algoritmo. Si en cambio, es significativo, introducimos en el modelo la variable que tenga el mayor coeficiente de correlación parcial con y habiendo eliminado el efecto de x_1 y x_2.

5. El algoritmo finaliza cuando obtenemos un t no significativo o cuando no quedan más variables que añadir.

Este método tiene la ventaja de requerir menos capacidad de cálculo que otros métodos, pues comienza trabajando con pocas variables. Por otro lado, el algoritmo no es capaz de eliminar variables cuando al incluir otras, éstas se vuelven innecesarias, por lo que puede causar error de especificación, tratado en la Sección 2.7.1. Además es posible que aparezca como no significativa alguna variable cuando realmente lo es, pero está relacionada con alguna variable no incluida. Por eso, este método tiene poco uso en la práctica.
También podemos aplicar el algoritmo con distintas técnicas de selección, por ejemplo, que entre en el modelo la variable que ofrece mayor ajuste según el criterio prefijado para seleccionar el modelo (C_p, AIC, CV, BIC).

3.1.1. Aplicación

Para mostrar un ejemplo de *selección hacia adelante* utilizaremos la librería $MASS$ con su función $stepAIC$, que además de utilizar el *criterio de*

información de Akaike, mencionado en la Sección 2.4 para contrastar los modelos, tiene como argumentos: el modelo con el que comienza el algoritmo, el modelo con el máximo número de variables, y la dirección en la que avanza, en este caso, hacia delante.
Comencemos leyendo los datos, cargando la librería y definiendo los dos modelos:

```
> golf <- read.table("golffull.txt", header=TRUE)
> attach(golf)
> library(MASS)
> mod0 <- lm(price~1, data=golf)
> mod8 <- lm(price~kilometer+age++extras1+extras2+TIA, data=golf)
```

Ya tenemos todos los argumentos para iniciar el algoritmo:

```
> mod.forward <- stepAIC(mod0, scope = list(upper = mod8),
>                        direction = "forward")
Start:  AIC=76.69
price ~ 1

            Df Sum of Sq    RSS     AIC
+ age        1   135.435 130.09 -44.030
+ kilometer  1    88.086 177.44   9.357
+ extras2    1     3.663 261.86  76.297
<none>                   265.53  76.686
+ TIA        1     0.286 265.24  78.501
+ extras1    1     0.257 265.27  78.520

Step:  AIC=-44.03
price ~ age

            Df Sum of Sq    RSS     AIC
+ kilometer  1   27.8855 102.21 -83.524
+ extras1    1    2.0038 128.09 -44.700
<none>                   130.09 -44.030
+ extras2    1    0.5429 129.55 -42.750
+ TIA        1    0.1996 129.89 -42.294

Step:  AIC=-83.52
price ~ age + kilometer

          Df Sum of Sq    RSS     AIC
```

```
+ extras1  1   1.88660 100.32 -84.729
<none>                 102.21 -83.524
+ TIA      1   0.09055 102.12 -81.677
+ extras2  1   0.02798 102.18 -81.572

Step:  AIC=-84.73
price ~ age + kilometer + extras1

          Df Sum of Sq    RSS     AIC
<none>                 100.32 -84.729
+ TIA      1  0.251492 100.07 -83.161
+ extras2  1  0.009439 100.31 -82.745
```

Por lo que el modelo resultante con este método es: $price \sim age + kilometer + extras1$.

3.2. Selección hacia atrás

Otro método es el de la *selección hacia atrás (o backward selection)*, el cual comienza con todas las variables explicativas disponibles en el modelo. El algoritmo es el siguiente:

1. Introducimos el modelo con todas las variables explicativas disponibles y fijamos un valor del estadístico t para cribar las variables.

2. Calculamos el modelo de regresión lineal y los estadísticos t para cada coeficiente.

3. Si alguno de ellos es menor que el valor prefijado, eliminamos dicha variable.

4. Volvemos a calcular el modelo de regresión lineal con las $k-1$ variables restantes y los estadísticos t para los coeficientes.

5. Repetimos el proceso hasta que no haya t significativos o no haya más variables.

Este método tiene varios inconvenientes inconvenientes, el primero es que es necesaria una gran capacidad de cálculo al trabajar con muchas variables. El segundo es que es bastante probable que tengamos el problema de la multicolinealidad, tratado en la Sección 2.7.5, pues si trabajamos con muchas variables es probable que algunas de ellas estén fuertemente relacionadas.

Por otro lado, es un método excelente para evitar que una variable significativa quede excluida de nuestro modelo, por lo que es bastante útil cuando el número de variables explicativas no es muy grande. Como en la sección anterior, podemos usar otro método para excluir variables, por ejemplo, eliminar la variable explicativa que proporcione un menor ajuste según el criterio elegido (C_p, AIC, CV, BIC).

3.2.1. Aplicación

Para el ejemplo de *selección hacia atrás* utilizamos también la librería $MASS$ con su función *stepAIC*, que además de utilizar el AIC dado en la Sección 2.4 para contrastar los modelos, tiene como argumentos: el modelo con el que comienza el algoritmo, el modelo con el mínimo número de variables, y la dirección en la que avanza, en este caso, hacia atrás.
Como ya teníamos los modelos definidos en el Ejemplo 3.1.1, basta con ejecutar la función:

```
> mod.backward <- stepAIC(mod8, scope = list(lower = mod0),
>                         direction = "backward")
Start:  AIC=-81.17
price ~ kilometer + age + +extras1 + extras2 + TIA

            Df Sum of Sq    RSS     AIC
- extras2    1     0.003 100.07 -83.161
- TIA        1     0.245 100.31 -82.745
<none>                   100.07 -81.166
- extras1    1     2.030 102.10 -79.712
- kilometer  1    27.222 127.29 -41.778
- age        1    76.403 176.47  14.412

Step:  AIC=-83.16
price ~ kilometer + age + extras1 + TIA

            Df Sum of Sq    RSS     AIC
- TIA        1     0.251 100.32 -84.729
<none>                   100.07 -83.161
- extras1    1     2.048 102.12 -81.677
- kilometer  1    27.592 127.66 -43.276
- age        1    76.711 176.78  12.715
```

```
Step:  AIC=-84.73
price ~ kilometer + age + extras1

            Df Sum of Sq    RSS     AIC
<none>                   100.32 -84.729
- extras1    1     1.887 102.21 -83.524
- kilometer  1    27.768 128.09 -44.700
- age        1    76.600 176.92  10.852
```

Por tanto, según este método obtenemos el siguiente modelo: $price \sim kilometer + age + extras1$.

3.3. Selección paso a paso

El método de *selección paso a paso (o stepwise selection)*, es una combinación de los dos anteriores, así, evita los inconvenientes de la selección hacia adelante y no requiere de una capacidad de cálculo tan grande como la de la selección hacia atrás. En cada paso se contrasta si entra una nueva variable explicativa o sale una que ya esté en el modelo. El algoritmo requiere fijar dos reglas, una para las variables de entrada y otra para las variables de salida. El proceso termina cuando no haya mejoras significativas a la hora de añadir o eliminar alguna variable.
Este método es el más utilizado de los 3.

3.3.1. Aplicación

Por último veremos un ejemplo de la *selección paso a paso*, en el cual utilizaremos la misma librería $MASS$ con su función $stepAIC$, que recibirá de argumentos: el modelo con el que comienza el algoritmo, el modelo con el máximo número de variables, y la dirección en la que avanza, en este caso, hacia ambos lados.
Como en el ejemplo anterior, ejecutamos solamente la función:

```
> mod.step <- stepAIC(mod0, scope = list(upper = mod8),
>                     direction = "both")
Start:  AIC=76.69
price ~ 1

            Df Sum of Sq    RSS     AIC
+ age        1   135.435 130.09 -44.030
+ kilometer  1    88.086 177.44   9.357
```

```
+ extras2    1     3.663 261.86  76.297
<none>                  265.53  76.686
+ TIA        1     0.286 265.24  78.501
+ extras1    1     0.257 265.27  78.520

Step:  AIC=-44.03
price ~ age

            Df Sum of Sq    RSS     AIC
+ kilometer  1    27.886 102.21 -83.524
+ extras1    1     2.004 128.09 -44.700
<none>                  130.09 -44.030
+ extras2    1     0.543 129.55 -42.750
+ TIA        1     0.200 129.89 -42.294
- age        1   135.435 265.53  76.686

Step:  AIC=-83.52
price ~ age + kilometer

            Df Sum of Sq    RSS     AIC
+ extras1    1     1.887 100.32 -84.729
<none>                  102.21 -83.524
+ TIA        1     0.091 102.12 -81.677
+ extras2    1     0.028 102.18 -81.572
- kilometer  1    27.886 130.09 -44.030
- age        1    75.234 177.44   9.357

Step:  AIC=-84.73
price ~ age + kilometer + extras1

            Df Sum of Sq    RSS     AIC
<none>                  100.32 -84.729
- extras1    1     1.887 102.21 -83.524
+ TIA        1     0.251 100.07 -83.161
+ extras2    1     0.009 100.31 -82.745
- kilometer  1    27.768 128.09 -44.700
- age        1    76.600 176.92  10.852
```

Con este método obtenemos el siguiente modelo: $price \sim age + kilometer + extras1$.

Capítulo 4

Técnicas de regularización

Para hallar los estimadores por el método de *mínimos cuadrados ordinarios* de los parámetros en el modelo lineal clásico, hay que resolver el sistema de ecuaciones:

$$\boldsymbol{X'X\beta = X'y}.$$

Para que este sistema tenga solución única, la matriz $\boldsymbol{X}$ debe tener rango máximo, $rango(\boldsymbol{X}) = p$.
Sin embargo, puede haber problemas:

- Cuando haya columnas en la matriz $\boldsymbol{X}$ que sean casi combinación lineal de otras, es decir, cuando se presenta el problema de la colinealidad.
- Cuando el número de regresores, p, es grande. En este caso, la solución es numéricamente inestable, aunque los coeficientes sigan siendo identificables en teoría.

Además, en muchas de las aplicaciones que están surgiendo en nuestros días, por ejemplo en genética, ocurre que el número de covariables es mucho mayor que el número de observaciones de las que disponemos. Esto se conoce como problemas con "n pequeño, y p grande". En todas estas situaciones son útiles las técnicas de regularización. Puede decirse que las técnicas de regularización se aplican para obtener estimaciones de los coeficientes de regresión cuando la matriz $\boldsymbol{X'X}$ es singular o está muy próxima a serlo. En este contexto, regularizar significa, hacer el problema tratable, imponiendo una serie de restricciones al conjunto de soluciones admisibles.
En las técnicas de regularización se plantea un problema de optimización que considera como función objetivo una obtenida por mínimos cuadrados penalizados (Penalized Least Squares, PLS)

$$\boldsymbol{PLS(\beta) = (y - X\beta)'(y - X\beta)}+\lambda pen(\boldsymbol{\beta})$$

donde $\lambda \geq 0$ es un parámetro de penalización que controla el efecto de la penalización y pen($\boldsymbol{\beta}$) es el término de penalización.
Si $\lambda \simeq 0$ entonces $\hat{\boldsymbol{\beta}}_{PLS}$ está próximo al MCO($\hat{\boldsymbol{\beta}}_{LS}$). En cambio, si λ es grande, se le da mucha importancia a la penalización.
Luego, se plantea ahora el problema de hallar $\hat{\boldsymbol{\beta}}_{PLS}$:

$$\hat{\boldsymbol{\beta}}_{PLS} = \arg\min_{\beta}[(\boldsymbol{y} - X\boldsymbol{\beta})'(\boldsymbol{y} - X\boldsymbol{\beta}) + \lambda pen(\boldsymbol{\beta})].$$

Que es equivalente a resolver el problema de optimización:

$$\hat{\boldsymbol{\beta}}_{PLS} = \arg\min_{\boldsymbol{\beta}}[(\boldsymbol{y} - X\boldsymbol{\beta})'(\boldsymbol{y} - X\boldsymbol{\beta})]$$

$$s.a.\ pen(\boldsymbol{\beta}) \leq t$$

donde t es una constante relacionada con el parámetro de penalización (smoothing) λ en una relación uno a uno.

4.1. Regresión contraída

La regresión contraída (o *ridge regression)* fue introducida en 1970 por Hoerl y Kennard.
Recordemos que

$$\boldsymbol{\beta} = (\beta_1, ..., \beta_p)'.$$

En el caso de la regresión ridge se considera la siguiente penalización:

$$pen(\boldsymbol{\beta}) = \|\boldsymbol{\beta}\|^2 = \sum_{j=0}^{k} \beta_j^2 = \boldsymbol{\beta}'\boldsymbol{\beta}$$

por tanto nos queda:

$$PLS(\boldsymbol{\beta}) = (\boldsymbol{y} - \boldsymbol{X\beta})'(\boldsymbol{y} - \boldsymbol{X\beta}) + \lambda\boldsymbol{\beta}'\boldsymbol{\beta}.$$

De donde puede comprobarse que:

$$\hat{\boldsymbol{\beta}}_{PLS} = (\boldsymbol{X}'\boldsymbol{X} + \lambda\boldsymbol{I}_p)^{-1}\boldsymbol{X}'\boldsymbol{y}.$$

Observación:

Recordemos que

$$\hat{\boldsymbol{\beta}}_{LS} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}.$$

Para valores de λ cercanos a cero, el impacto de pen($\boldsymbol{\beta}$) es prácticamente nulo, y $\hat{\boldsymbol{\beta}}_{PLS} \simeq \hat{\boldsymbol{\beta}}_{LS}$. Sin embargo, si λ es grande, esta técnica permite resolver el problema de la multicolinealidad, porque hace que la matriz $(\boldsymbol{X'X} + \lambda \boldsymbol{I}_p)^{-1}$ sea invertible en el caso de que $\boldsymbol{X'X}$ no lo fuera.
Además, $\hat{\boldsymbol{\beta}}_{PLS}$ es una contracción de $\hat{\boldsymbol{\beta}}_{LS}$ hacia cero. Esto puede verse observando la función objetivo a minimizar, y el papel que en ella desempeña $\lambda \boldsymbol{\beta'\beta}$. Si λ es grande, el $\min_{\boldsymbol{\beta}}\{PLS(\boldsymbol{\beta})\}$, estará determinado por el término que minimice $\lambda pen(\boldsymbol{\beta}) = \lambda \boldsymbol{\beta'\beta}$, que claramente se minimiza cuando $\boldsymbol{\beta} = 0$.
En la práctica no interesa penalizar la ordenada en el origen (*intercept*) del modelo de regresión, β_0. Para ello existen dos alternativas:

- Centrar todas las covariables y la variable respuesta, Y, para que los nuevos valores de éstas tengan media cero, $\overline{y} = 0$, $\overline{x} = 0$, lo que automáticamente produce que $\hat{\beta}_0 = 0$. Esto implica que el *intercept* (o término constante) se elimina del modelo, y por lo tanto no se penaliza.

- Modificar la penalización a:

$$pen(\boldsymbol{\beta}) = \sum_{j=1}^{k} \beta_j^2 = \boldsymbol{\beta' K \beta}$$

  donde $\boldsymbol{K} = diag(0, 1, ..., 1)$, es decir, se introduce una matriz de penalización que excluye al coeficiente β_0, y sigue siendo la identidad para el resto de los coeficientes. Esta segunda opción es la que adoptaremos, y nos conduce al estimador de regresión contraída (*ridge estimate*) dado por:

$$\hat{\boldsymbol{\beta}}_{PLS} = (\boldsymbol{X'X} + \lambda \boldsymbol{K})^{-1} \boldsymbol{X'y}.$$

Proposición 4.1.1. *Propiedades de* $\hat{\boldsymbol{\beta}}_{PLS}$*:*

- $\boldsymbol{E}(\hat{\boldsymbol{\beta}}_{PLS}) = (\boldsymbol{X'X} + \lambda \boldsymbol{K})^{-1} \boldsymbol{X'X\beta}$. Por tanto, no es insesgado salvo que $\lambda = 0$. Normalmente ocurrirá $|\hat{\boldsymbol{\beta}}_{j,PLS}| \leq |\hat{\boldsymbol{\beta}}_{j,LS}|$, $j = 1, ..., k$. Aunque no siempre se tiene.

- $Cov(\hat{\boldsymbol{\beta}}_{PLS}) = \sigma^2 (\boldsymbol{X'X} + \lambda \boldsymbol{K})^{-1} \boldsymbol{X'X} (\boldsymbol{X'X} + \lambda \boldsymbol{K})^{-1}$.

Proposición 4.1.2. *Comparación* $\hat{\boldsymbol{\beta}}_{PLS}, \hat{\boldsymbol{\beta}}_{LS}$*:*

- $\boldsymbol{E}(\hat{\boldsymbol{\beta}}_{LS}) = \boldsymbol{E}((\boldsymbol{X'X})^{-1} \boldsymbol{X'y}) = (\boldsymbol{X'X})^{-1} \boldsymbol{X'X\beta} = \boldsymbol{\beta}$, luego es insesgado.

- $Cov(\hat{\boldsymbol{\beta}}_{LS}) = \sigma^2 (\boldsymbol{X'X})^{-1}$.

Luego:

$$\hat{\boldsymbol{\beta}}_{PLS} = (\boldsymbol{X'X} + \lambda \boldsymbol{K})^{-1}\boldsymbol{X'y} = (\boldsymbol{X'X} + \lambda \boldsymbol{K})^{-1}(\boldsymbol{X'X})\underbrace{(\boldsymbol{X'X})^{-1}\boldsymbol{X'y}}_{\hat{\boldsymbol{\beta}}_{LS}}$$

$$= (\boldsymbol{X'X} + \lambda \boldsymbol{K})^{-1}(\boldsymbol{X'X})\hat{\boldsymbol{\beta}}_{LS}.$$

En el caso de matrices ortogonales se puede obtener una relación entre los coeficientes de ambos estimadores que ilustra porqué se llama regresión contraída:

$$Cov(\hat{\boldsymbol{\beta}}_{PLS}) = \sigma^2(\boldsymbol{X'X} + \lambda \boldsymbol{K})^{-1}(\boldsymbol{X'X})(\boldsymbol{X'X} + \lambda \boldsymbol{K})^{-1}.$$

Puede probrarse que la matriz $Cov(\hat{\boldsymbol{\beta}}_{LS}) - Cov(\hat{\boldsymbol{\beta}}_{PLS})$ es definida positiva para $\lambda > 0$, lo que implica que: $Var(\hat{\boldsymbol{\beta}}_{j,PLS}) < Var(\hat{\boldsymbol{\beta}}_{j,LS}), \quad j = 1, ..., k.$
En resumen, con la regresión ridge el estimador que se obtiene es sesgado, pero tiene menor ECM.
Lo único que quedaría es calcular el parámetro λ adecuado, lo que se hace normalmente por el método de la *validación cruzada* comentado en la Sección 2.3.
Nótese por último que la escala de las covariables es importante cuando estamos regularizando. La penalización formada por el cuadrado de los coeficientes de regresión asume que todos los coeficientes pueden ser comparados en valor absoluto. Sin embargo, la escala tiene un impacto directo en la interpretacion de esos valores absolutos. Por ejemplo, el coeficiente asociado a una covariable que mide la distancia estará escalada por un factor de 1.000 cuando la variable se mida en metros en vez de kilómetros. Por lo tanto, es importante hacer que todas las variables sean comparables en su escala antes de aplicar la aproximación por mínimos cuadrados penalizados. La solución más común es la de normalizar todas las variables.

4.1.1. Aplicación

Para ilustrar la regresión *ridge* usaremos los paquetes de *R*: *car* y $MASS$. Y un fichero de datos utilizado en *Tibshirani, R. et al.* [12]. El objetivo de este estudio es determinar qué variables influyen en la presencia de un antígeno prostático específico, el cual se utiliza para detectar el cáncer de próstata.

```
> url <- "http://www-stat.stanford.edu/~tibs/ElemStatLearn
+ /datasets/prostate.data"
> cancer <- read.table(url, header=TRUE)
> library(car)
> library(MASS)
```

Nuestro fichero de datos cuenta con 97 observaciones y 10 variables, las cuales son:

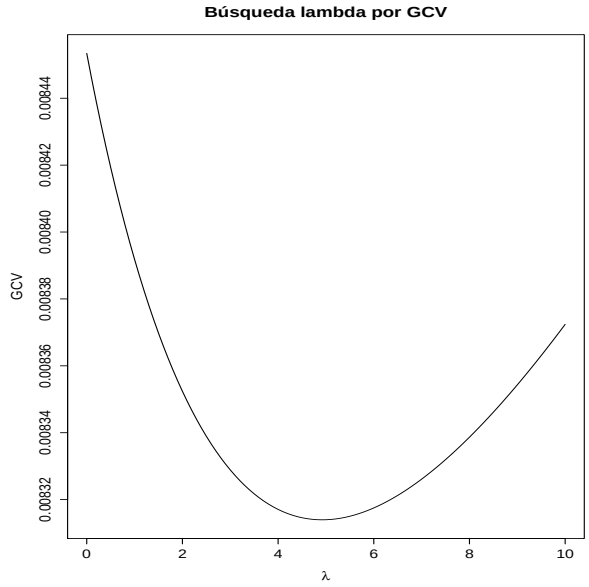
- lcavol: log-volúmen del cáncer.
- lweight: log-tamaño de la próstata.
- age: edad del paciente.
- lbhp: log-cantidad de hiperplasia benigna.
- svi: toma el valor 1 si está invadida la vesícula seminal y 0 si no.
- lcp: log-penetración capsular.
- gleason: puntuación Gleason.
- pgg45: porcentaje de la puntuación Gleason 4 ó 5.
- lpsa: log-análisis del antígeno prostático específico.
- train: variable para distinguir el conjunto de entrenamiento y el de test.

Seleccionemos ahora el conjunto *test* y el conjunto *train* valiéndonos de la variable *"train"* antes mencionada. Tal como está conformado el fichero de datos, el 70 % del conjunto está destinado al entrenamiento del modelo.

```
> train = subset(cancer,train=="TRUE")
> test = subset(cancer,train=="FALSE")
```

Calculemos ahora el modelo de regresión *ridge* con la función *lm.ridge* la cual tiene implementada una búsqueda del λ óptimo a través de la *validación cruzada generalizada*, es importante remarcar que este término puede inducir a error, pues no es una generalización de la *validación cruzada* mencionada en la sección 2.3, aunque se utiliza por convenio dicho nombre, se podría hablar de "aproximación".

```
> modelo_ridge <- lm.ridge(lpsa ~ ., data=train[,-10],
>                          lambda = seq(0,10,0.1))
> plot(seq(0,10,0.1), modelo_contraida$GCV,
>      main="Búsqueda lambda por GCV",
+ type="l", xlab=expression(lambda), ylab="GCV")
```

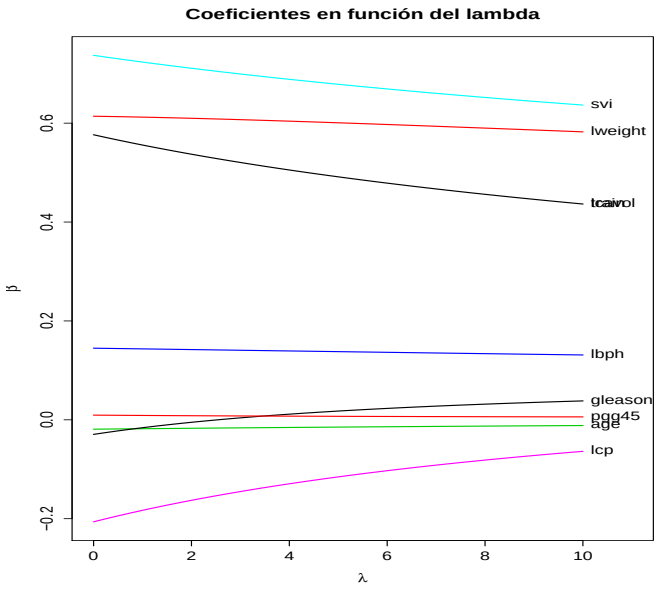


Vemos que el λ debe estar próximo a 5, para averiguar el valor óptimo podemos emplear la función *select*:

```
> select(lm.ridge(lpsa ~ ., data=train[,-10], lambda = seq(0,10,0.1)))
modified HKB estimator is 3.355691
modified L-W estimator is 3.050708
smallest value of GCV  at 4.9
```

luego el valor óptimo es $\lambda = 4{,}9$. Podemos ver también cómo varían los coeficientes al modificar el λ

```
> matplot(seq(0,10,0.1), coef(modelo_ridge)[,-1], xlim=c(0,11), type="l",
+ xlab=expression(lambda), ylab=expression(hat(beta)), lty=1, lwd=2,
+ main="Coeficientes en función del lambda")
> text(rep(10, 9), coef(modelo_ridge)[length(seq(0,10,0.1)),-1],
+ colnames(train)[-9], pos=4)
```



Se aprecia que al aumentar el λ los coeficientes tienden a 0, pero debemos tener en cuenta que a mayor λ, mayor es el sesgo de nuestro modelo.
Tenemos ya definido nuestro modelo:

```
> modelo_ridge <- lm.ridge(lpsa ~ ., data=train[,-10], lambda = 4.9)
> coefficients(modelo_ridge)
                  lcavol      lweight          age         lbph
 0.096814771  0.492787412  0.601103227 -0.014821787  0.138019854
         svi           lcp      gleason        pgg45
0.679632580  -0.116790333  0.017113954  0.007081258
```

Para finalizar calculemos el error cuadrático medio del modelo obtenido por el método de *mínimos cuadrados ordinarios* y el error del modelo penalizado.

```
> modelo_mco <- lm(lpsa~ . , data=train[,-10])
> ajuste_mco <- predict(modelo_mco,test)
> sum((test$lpsa-ajuste_mco)^2)
[1] 15.63822
```

Este modelo tiene una suma de errores cuadráticos medios de 15.63822. Veamos ahora el modelo obtenido por la regresión *ridge*:

```
> coeficientes <- as.vector(coef(modelo_ridge))
> matriz <- as.matrix(test[,-9:-10])
> matriz <- cbind(rep(1,length=nrow(test)),matriz)
```

```
> ajuste_ridge <- matriz%*%coeficientes
> sum((test$lpsa- ajuste_ridge)^2)
[1] 14.8323
```

tiene un error de 14.8323, por tanto, mejora al modelo obtenido por los mínimos cuadrádos ordinarios según el criterio de la GCV.

4.2. Regresión LASSO

En esta sección estudiamos la regresión *LASSO* (*Least Absolute Shrinkage and Selection Operator*). Este método realiza tanto la selección de variables como la regularización de la solución con el fin de mejorar la precisión de la predicción y la interpretación del modelo estadístico que produce. Fue introducido por *Robert Tibshirani* en 1996.
Hemos visto que la *regresión ridge* permite la estimación de los coeficientes de regresión cuando la matriz $\boldsymbol{X'X}$ es casi singular, pero no es una solución completamente satisfactoria, pues todos los coeficientes de regresión estimados pueden ser distintos de cero. Para la interpretación sería conveniente, no sólo reducir los coeficientes y hacerlos cercanos a cero, sino tener la posibilidad de que alguno fuera exactamente cero. Esto nos permitiría hacer la estimación del modelo con la selección de variables de una sola vez, lo cual se consigue sustituyendo la penalización de los coeficientes de regresión al cuadrado, β_j^2, por una penalización basada en el valor absoluto:

$$pen(\boldsymbol{\beta}) = \sum_{j=1}^{k} |\beta_j|.$$

Por lo que

$$\hat{\boldsymbol{\beta}}_{LASSO} = \arg\min_{\boldsymbol{\beta}} \left[(\boldsymbol{y} - \boldsymbol{X\beta})'(\boldsymbol{y} - \boldsymbol{X\beta}) + \lambda \sum_{j=1}^{k} |\beta_j| \right]$$

donde de nuevo no penalizaremos al intercept.
Como en la *regresión ridge*, el criterio de mínimos cuadrados penalizados realiza un balance entre ajustar los datos por el criterio de mínimos cuadrados y regularizar la solución según determina la penalización. La compensación entre estos dos objetivos está controlada por el parámetro de penalización. El nombre que le dio *Tibshirani* es debido a que $\hat{\boldsymbol{\beta}}_{LASSO}$ está definido en función de una penalización con valor absoluto y permite seleccionar las covariables de una manera determinada (lo veremos más adelante).

La diferencia entre ambas regresiones puede verse en su penalización. La *regresión ridge* impone una penalización cuadrática que tiene un fuerte impacto en los coeficientes grandes pero una penalización pequeña en los coeficientes cercanos a cero. Por contra, la penalización en valor absoluto de la regresión *LASSO* aumenta de manera más lenta para los coeficientes grandes, pero se aleja más rápido del cero para coeficientes que están cercanos a cero. Consecuentemente, esperamos que se comporte de manera que los coeficientes pequeños tiendan más fuerte a cero, mientras que los coeficientes grandes no se vean afectados apenas por la penalización. Lo vemos mejor en la Figura 4.1:

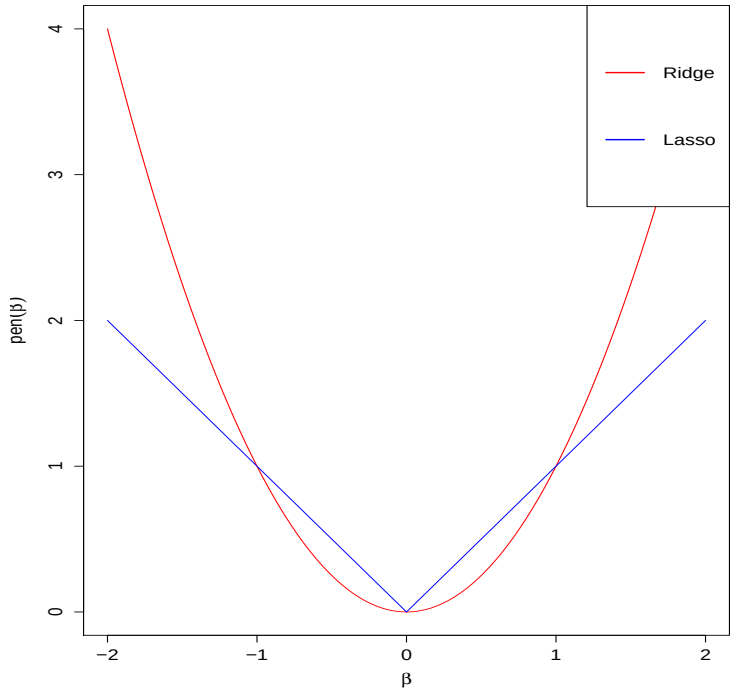


Figura 4.1: Penalizaciones de la *regresión ridge* (en rojo) y de la *regresión LASSO* (en azul).

A diferencia de la *regresión ridge*, no hay una solución analítica disponible para la *regresión LASSO* (salvo en el caso de un diseño ortogonal $\boldsymbol{X'X} = \boldsymbol{I}_p$). Aunque el criterio de mínimos cuadrados penalizados no es diferenciable, por la inclusión de la penalización con el valor absoluto, podemos seguir obteniendo ecuaciones para estimar similares a las ecuaciones normales en el

modelo lineal clásico. Éstas vienen dadas por:

$$\boldsymbol{2X'X\beta + 2X'y} + \lambda \sum_{j=1}^{k} sign(\beta_j) = 0.$$

Sin embargo, la solución no explícita no se puede calcular porque la función *sign* de la expresión anterior, a diferencia de lo que ocurre en la *ridge regresión*, no es lineal en los datos. A pesar de esto, no puede expresarse como un estimador lineal $\hat{\boldsymbol{\beta}} = \boldsymbol{Ay}$ con $\mathbf{A}$ $(p \times n)$. Esto también implica que las propiedades estadísticas de $\hat{\boldsymbol{\beta}}_{LASSO}$ son mucho más complicadas de deducir que en el caso de la *regresión ridge*. Aunque, estas son, en principio, análogas a la *regresión ridge*, es decir, $\hat{\boldsymbol{\beta}}_{LASSO}$ es sesgado, y tiene menor ECM que el estimador por *mínimos cuadrados ordinarios*. Una comparación entre *ridge* y *LASSO* es más complicada y no existe una respuesta general en la cual se establezca cuál es preferible en términos del error cuadrático medio.
El criterio *LASSO* puede escribirse de manera equivalente como:

$$\hat{\boldsymbol{\beta}}_{LASSO} = \arg\min_{\boldsymbol{\beta}} [\boldsymbol{(y - X\beta)'(y - X\beta)}]$$

sujeto a la restricción:

$$\sum_{j=1}^{k} |\beta_j| \leq t$$

donde t y el parámetro de penalización λ están relacionados uno a uno [*Sydsaeter et al. (2005)*].
La estimación *LASSO* se basa en aproximaciones de optimización numérica. En el enfoque original que le dió *Tibshirani*, fue aplicada una aproximación de programación cuadrática para optimizar el criterio de mínimos cuadrados penalizados sujeto a la restricción de desigualdad de la suma de valores absolutos, lo que no era eficiente para un número grande de variables por su alto coste computacional. Por lo que surgieron algoritmos alternativos como el *LARS (least angle regression)* (Efron y otros, 2004) y el *cyclical coordinate descent* (Friedman y otros, 2010) que permitieron reducir enormemente este coste.

4.2.1. Aplicación

En el ejemplo de la regresión *LASSO* usaremos el paquete de *R penalized* y el mismo fichero de datos de la Sección 4.1.1.
El paquete tiene la función *penalized* que puede usarse para ajustar un modelo penalizado para predecir la variable respuesta. Por ejemplo, vamos a predecir la variable *lpsa* de nuestro fichero de datos en función de las demás variables.

```
> url <- "http://www-stat.stanford.edu/~tibs/ElemStatLearn
+ /datasets/prostate.data"
> cancer <- read.table(url, header=TRUE)
> data=cancer[,-10]
> attach(data)
> library(penalized)
> ajustado <- penalized(lpsa, data[,1:8], lambda1=1, steps=50,
>                       trace = FALSE)
```

En este primer paso hemos modificado la base de datos porque no hacemos uso de la columna que nos indicaba las observaciones destinadas al conjunto *train* y al conjunto *test*. Cuando usamos el argumento *"steps"*, la función comienza a ajustar el modelo en el valor máximo de λ, que es el valor que hace que todos los coeficientes sean 0. Desde dicho valor continúa ajustando el modelo haciendo decrecer el λ hasta el valor especificado. El argumento $lambda1 = 1$ se utiliza para indicar que usaremos la penalización de la regresión $LASSO$ y no la de *ridge* que sería con la misma función y $lambda2 = 1$. Veamos en un gráfico el efecto de elegir un λ u otro:

```
> plotpath(ajustado, log="x")
```

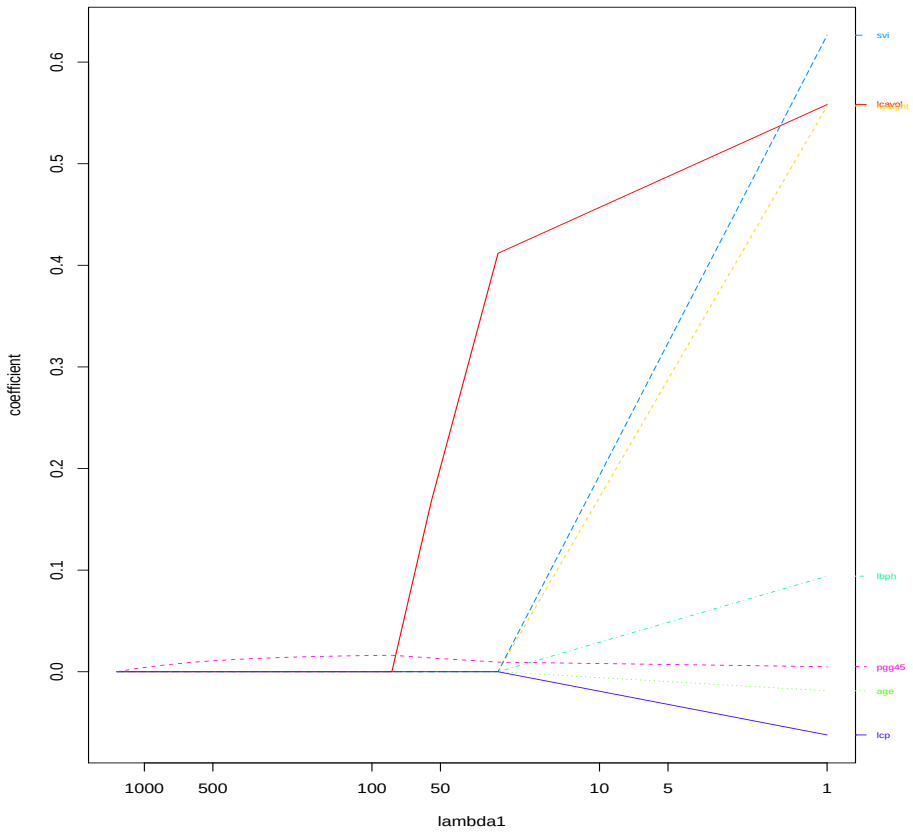


Figura 4.2: Evolución de los coeficientes según el λ escogido.

Dicho paquete tiene implementado un método basado en la CV para seleccionar el λ adecuado. Se consigue con la función *optL1* la cual calcula el óptimo y lo guarda en el elemento *lambda*:

```
> opt1 <- optL1(lpsa, data[,1:8], fold=10)
lambda= 516.6546  cvl= -147.4131
lambda= 835.9647  cvl= -150.0562
lambda= 319.3101  cvl= -146.4278
lambda= 197.3445  cvl= -146.0811
lambda= 47.2665  cvl= -135.4336
lambda= 104.5912  cvl= -145.9538
lambda= 69.16259  cvl= -145.6992
lambda= 29.2123  cvl= -125.7249
lambda= 18.0542  cvl= -121.2387
```

```
lambda= 11.15811  cvl= -118.9514
lambda= 6.89609  cvl= -116.5056
lambda= 4.262018  cvl= -112.5124
lambda= 2.634072  cvl= -111.4239
lambda= 1.766723  cvl= -111.8711
lambda= 2.743628  cvl= -111.389
lambda= 3.323602  cvl= -111.3628
lambda= 3.090396  cvl= -111.3362
lambda= 3.082834  cvl= -111.3357
lambda= 2.953269  cvl= -111.3415
lambda= 3.045993  cvl= -111.3337
lambda= 3.010576  cvl= -111.3323
lambda= 2.988686  cvl= -111.3356
lambda= 3.022384  cvl= -111.3326
lambda= 3.002215  cvl= -111.3335
lambda= 3.014719  cvl= -111.3323
lambda= 3.007382  cvl= -111.3327
lambda= 3.012158  cvl= -111.3323
lambda= 3.011457  cvl= -111.3322
lambda= 3.011386  cvl= -111.3322
lambda= 3.011077  cvl= -111.3322
> opt1$lambda
[1] 3.011386
```

una vez tenemos el λ ya podemos calcular nuestro modelo con sus coeficientes:

```
> ajustado <- penalized(lpsa, data[,1:8], lambda1=opt1$lambda)
# nonzero coefficients: 7
> coef(ajustado)
(Intercept)      lcavol     lweight         age        lbph
 0.88505882  0.54580818  0.44087198 -0.01492084  0.08791441
        svi       pgg45
 0.39953795  0.00468374
```

4.3. Propiedades geométricas de los estimadores regularizados

Como veremos en el ejemplo, la regularización *LASSO* da vectores con los coeficientes estimados de manera que algunos de los parámetros se estiman para que sean exactamente cero. Para entender este comportamiento, investigaremos las propiedades geométricas de la estimación por mínimos cuadrados

penalizados. Para ilustrar este hecho, vamos a considerar un vector de coeficientes $\boldsymbol{\beta} = (\beta_1, \beta_2)'$; pero todos los resultados se generalizan fácilmente. Notemos que no hemos incluido al *intercept* lo que supone que consideremos las covariables estandarizadas y una variable respuesta centrada.

Proposición 4.3.1. *El criterio de mínimos cuadrados, $LS(\boldsymbol{\beta})$, puede reescribirse como:*

$$\begin{aligned} LS(\boldsymbol{\beta}) &= (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})'\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) + \boldsymbol{y}'(\boldsymbol{I_n} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}')\boldsymbol{y} \\ &= (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})'\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) + \hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}. \end{aligned} \tag{4.1}$$

Demostración. Partamos del producto del criterio de mínimos cuadrados

$$(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) = \boldsymbol{y}'\boldsymbol{y} - 2\boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{y} + \boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta}.$$

Veamos ahora la expansión de la forma cuadrática en $\boldsymbol{\beta}$:

$$(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})'\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta} - 2\boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{X}\hat{\boldsymbol{\beta}} + \hat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{X}\hat{\boldsymbol{\beta}}.$$

Teniendo en cuenta que $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$, el segundo sumando es

$$2\boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{X}\hat{\boldsymbol{\beta}} = 2\boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} = 2\boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{y}$$

y el tercer sumando es

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

Con lo que llegamos a

$$(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})'\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta} - 2\boldsymbol{\beta}\boldsymbol{X}'\boldsymbol{y} + \boldsymbol{y}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$$

con esto tendríamos el primer sumando, el segundo la obtenemos viendo en la Expresión 4.1 que

$$\begin{aligned} \hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}} &= (\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})'(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) \\ &= (\boldsymbol{y} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y})'(\boldsymbol{y} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}) \\ &= \boldsymbol{y}'(\boldsymbol{I_n} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}')\boldsymbol{y}. \end{aligned}$$

□

Como hemos visto que $LS(\boldsymbol{\beta})$ es equivalente a una forma cuadrática, los valores de $\boldsymbol{\beta}$ que resuelven $LS(\boldsymbol{\beta}) = c$, para una constante c, es decir, sus curvas de nivel, son elipses con una forma determinada por la matriz $\boldsymbol{X}'\boldsymbol{X}$. Por otro lado, en dos dimensiones, la restricción:

$$|\beta_1| + |\beta_2| = t$$

define curvas de nivel con forma de diamente de lado $\sqrt{2}t$. Por lo tanto, el estimador *LASSO* regularizado, dada una t, es el punto de corte de las dos regiones geométricas definidas por la restricción y por el criterio de los mínimos cuadrados. Si el punto de corte está en uno de los vértices del diamante, algunos coeficientes se estimarán como cero.
Las curvas de nivel que se definen en la regresión *ridge* son círculos de la forma:

$$\beta_1^2 + \beta_2^2 = t$$

con lo que no se puede dar el caso que el estimador corte con la región en un vértice, pues es un círculo, con lo que no conseguiremos que ningún coeficiente se estime como cero.
Lo vemos mejor en los gráficos recogidos en la Figura 4.3:

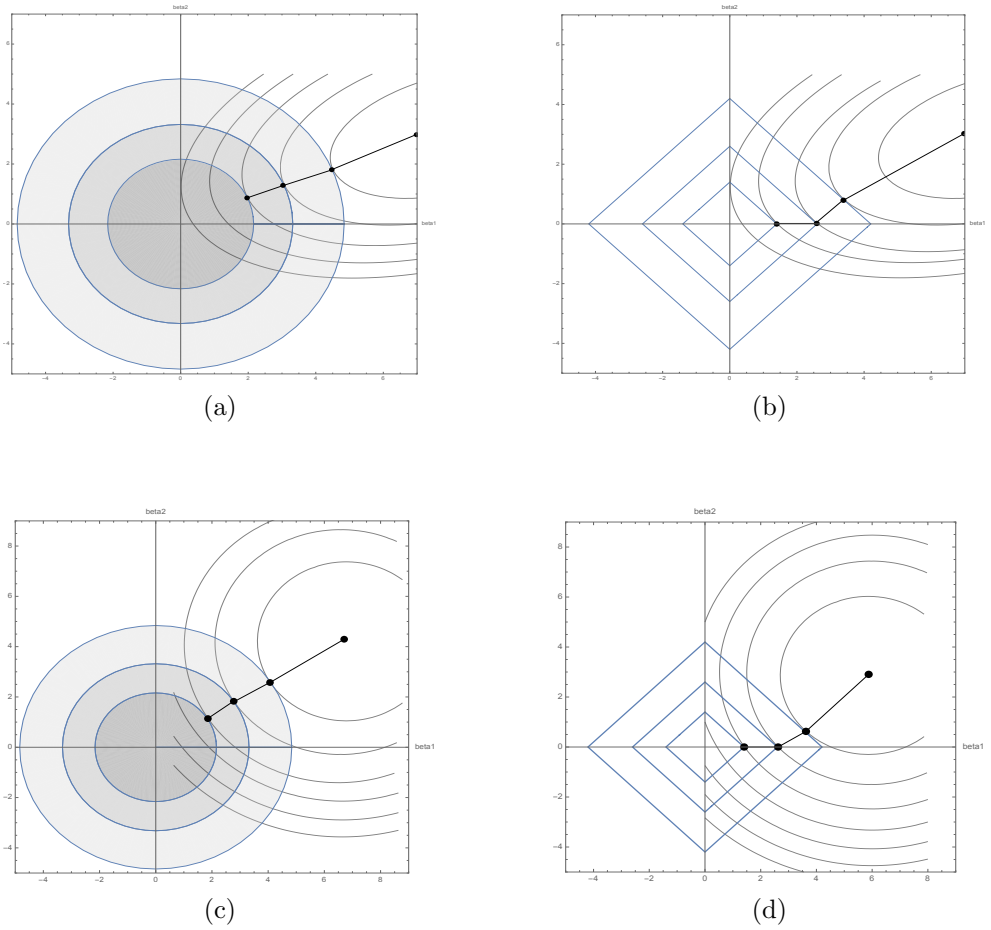


Figura 4.3: Interpretación geométrica del criterio de mínimos cuadrados penalizados para la regresión *ridge* (figuras (a) y (c)) y para la regresión *LASSO* (figuras (b) y (d)). En las figuras superiores, se considera una matriz $\boldsymbol{X'X}$ no diagonal, mientras que las figuras inferiores corresponden a una matriz de diseño $\boldsymbol{X'X} = \boldsymbol{I}_2$.

Resumen

Mínimos cuadrados penalizados

La estimación regularizada en el modelo lineal permite penalizar el criterio de los mínimos cuadrados:

$$PLS(\boldsymbol{\beta}) = (\boldsymbol{y} - \boldsymbol{X\beta})'(\boldsymbol{y} - \boldsymbol{X\beta}) + \lambda pen(\boldsymbol{\beta})$$

con el parámetro de penalización, $\lambda \geq 0$.

Regresión *ridge*

Para la regresion *ridge*, la penalización viene dada por la suma de los coeficientes al cuadrado:

$$pen(\boldsymbol{\beta}) = \sum_{j=1}^{k} \beta_j^2 = \boldsymbol{\beta' K \beta}$$

con la matriz de penalización $\boldsymbol{K} = diag(0, 1, ..., 1)$. El resultado de la estimación de los mínimos cuadrados penalizados es:

$$\hat{\boldsymbol{\beta}}_{PLS} = (\boldsymbol{X'X} + \lambda \boldsymbol{K})^{-1} \boldsymbol{X'y}.$$

Regresión *LASSO*

Para la regresión *LASSO*, la penalización viene dada por la suma de los valores absolutos de los coeficientes:

$$pen(\boldsymbol{\beta}) = \sum_{j=1}^{k} |\beta_j|.$$

El resultado de la estimación no tiene una fórmula analítica y debe determinarse numéricamente, por ejemplo utilizando técnicas de programación cuadrática.

Elección del parámetro de penalización

El parámetro de penalización, λ, puede determinarse por los métodos *r-fold cross validation* o con la *generalized cross validation.*

Apéndice A

Anexo

A.1. Comandos en R de las gráficas

A.1.1. Figura 2.1

```
> par(mfrow=c(2,2), las=1)
> plot(age,price,ylab="sales price in 1000$",xlab="age in months",
+ main="Sales price vs age")
> plot(kilometer,price,ylab="sales price in 1000$",
+ xlab="kilometer reading in 1000 km", main="Sales price vs kilometer")
> plot(TIA,price,ylab="sales price in 1000$",
+ xlab="months until next TIA appointment", main="Sales price vs TIA")
```

A.1.2. Figura 2.2

```
> par(mfrow=c(1,2), las=2)
> boxplot(price ~ extras1, main="Sales price vs no ABS/ABS",
+ ylab="sales in 1000$", col="gold", names=c("no ABS", "ABS"))
> boxplot(price ~ extras2, main="Sales price vs no sunroof/sunroof",
+ ylab="sales in 1000$", col="gold", names=c("no sunroof", "sunroof"))
```

A.1.3. Figura 4.1

```
> library(gplots)
> curve(x^2, from=-2, to=2, xlab=expression(beta),
+ ylab=expression(pen(beta)), col="red", ylim=c(0, 4))
> curve(abs(x), from=-2, to=2, col="blue", add=T)
> legend("topright", c("Ridge", "Lasso"),
+ lwd=2, col=c("red", "blue"))
```

A.2. Comandos en Mathematica de las gráficas

A.2.1. Figura 4.3

```
Penalización Ridge y matriz no -diagonal

circ1 := ParametricPlot[{r*Cos[t], r*Sin[t]}, {t, 0, 2*Pi}, {r, 0,
   2.16}, PlotStyle -> {GrayLevel[0.25]},
  PlotRange -> {{-5, 7}, {-5, 7}}, AxesLabel -> {"beta1", "beta2"}]
circ2 := ParametricPlot[{r*Cos[t], r*Sin[t]}, {t, 0, 2*Pi}, {r, 2.16,
   3.32}, PlotStyle -> {GrayLevel[0.55]},
  PlotRange -> {{-5, 7}, {-5, 7}}, AxesLabel -> {"beta1", "beta2"}]
circ3 := ParametricPlot[{r*Cos[t], r*Sin[t]}, {t, 0, 2*Pi}, {r, 3.32,
   4.84}, PlotStyle -> {GrayLevel[0.8]},
  PlotRange -> {{-5, 7}, {-5, 7}}, AxesLabel -> {"beta1", "beta2"}]
X = {{0.5, -1}, {0.45, 0.5}};
beta = {{beta1}, {beta2}};
LS = (beta - {{8}, {3}})\[Transpose].X\[Transpose].X.(beta - {{8}, \
{3}});
elipse :=
 ContourPlot[(beta - {{8}, {3}})\[Transpose].X\[Transpose].X.(beta - \
{{8}, {3}}), {beta1, 0, 7}, {beta2, -2, 5}, ContourShading -> False,
  PlotRange -> {{-5, 7}, {-5, 7}}, AxesLabel -> {"beta1", "beta2"}]
cont1 := Show[elipse, circ1, circ2, circ3, Axes -> True]
cont1


Penalización LASSO y matriz no - diagonal

rombo1 :=
 ContourPlot[{Abs[beta1] + Abs[beta2] == 1.4}, {beta1, -4,
   4}, {beta2, -4, 4}, PlotRange -> {{-5, 7}, {-5, 7}},
  AxesLabel -> {"beta1", "beta2"}]
rombo2 :=
 ContourPlot[{Abs[beta1] + Abs[beta2] == 2.6}, {beta1, -4,
   4}, {beta2, -4, 4}, PlotRange -> {{-5, 7}, {-5, 7}},
  AxesLabel -> {"beta1", "beta2"}]
rombo3 :=
 ContourPlot[{Abs[beta1] + Abs[beta2] == 4.2}, {beta1, -4.5,
```

```
   4.5}, {beta2, -5, 5}, PlotRange -> {{-5, 7}, {-5, 7}},
  AxesLabel -> {"beta1", "beta2"}]
cont2 := Show[elipse, rombo1, rombo2, rombo3, Axes -> True]
cont2


Penalización Ridge y matriz ortonormal

X = {{1, 0}, {0, 1}};
circulos :=
 ContourPlot[(beta - {{6}, {3}})\[Transpose].X\[Transpose].X.(beta - \
{{6}, {3}}), {beta1, 0, 8}, {beta2, -5, 10}, ContourShading -> False,
  PlotRange -> {{-5, 9}, {-5, 9}}, AxesLabel -> {"beta1", "beta2"}]
cont3 := Show[circulos, circ1, circ2, circ3, Axes -> True]
cont3


Penalización LASSO y matriz ortonormal

cont4 := Show[circulos, rombo1, rombo2, rombo3, Axes -> True]
cont4
```

A.3. Paquetes de R

- C. Agostinelli and U. Lund (2013). R package 'circular': Circular Statistics (version 0.4-7). URL `https://r-forge.r-project.org/projects/circular/`.
- Claudio Agostinelli and SLATEC Common Mathematical Library (2015). wle: Weighted Likelihood Estimation. R package version 0.9-91. `https://CRAN.R-project.org/package=wle`.
- John H. Maindonald and W. John Braun (2015). DAAG: Data Analysis and Graphics Data and Functions. R package version 1.22. `https://CRAN.R-project.org/package=DAAG`.
- Sarkar, Deepayan (2008) Lattice: Multivariate Data Visualization with R. Springer, New York. ISBN 978-0-387-75968-5.
- Venables, W. N. & Ripley, B. D. (2002) Modern Applied Statistics with S. Fourth Edition. Springer, New York. ISBN 0-387-95457-0.

- John Fox and Sanford Weisberg (2011). An R Companion to Applied Regression, Second Edition. Thousand Oaks CA: Sage. URL: `http://socserv.socsci.mcmaster.ca/jfox/Books/Companion`.
- Jelle Goeman, Rosa Meijer and Nimisha Chaturvedi (2014). penalized: L1 (lasso and fused lasso) and L2 (ridge) penalized estimation in GLMs and in the Cox model. R package version 0.9-45. `https://CRAN.R-project.org/package=penalized`.

Bibliografía

[1] Fahrmeir, L.; Kneib, Th.; Lang, S.; Marx, B. *Regression: Models, Methods and Applications.* New York: Springer. 2013.

[2] Peña, D. *Regresión y diseño de experimentos.* Madrid: Alianza Editorial, S.A. 2002.

[3] Harrell, F. E. Jr. *Regression Modeling Strategies.* New York: Springer. 2001.

[4] Apuntes de Modelos lineales y diseño de experimentos. Tercer curso de Grado en Matemáticas, 2013-14. Universidad de Sevilla. (Profesores D. Juan M. Muñoz Pichardo y D. Joaquín Antonio García de las Heras).

[5] Apuntes de Inferencia estadística. Tercer curso de Grado en Matemáticas, 2014-15. Universidad de Sevilla. (Profesores D. Emilio Carrizosa Priego y D. Joaquín Antonio García de las Heras).

[6] Carmen García Olaverri. (1996). Estabilidad de algunos criterios de selección de modelos. *Qüestiió, Vol 20, 2 pp. 147-166.*

[7] Andrew W. Moore. *Cross-validation for detecting and preventing overfitting.* Apuntes. Carnegie Mellon University.

[8] Pilar Cacheiro Martínez. (2011). *Métodos de selección de variables en estudios de asociación genética. Aplicación a un estudio de genes candidatos en Enfermedad de Parkinson.* Fin de máster. A Coruña: Universidad de Santiago de Compostela.

[9] Dña. María Jesús Bárcena Ruíz. Universidad del País Vasco. Economía Aplicada III (Estadística y Econometría). `http://campusvirtual.ehu.es/open_course_ware/castellano/experimentales/estadistica/materiales-de-estudio/index.html`.

[10] D. Juan M. Vilar Fernández. Universidad de Santiago de Compostela. `http://dm.udc.es/asignaturas/estadistica2/indice_res.html`.

[11] Francisco Félix Caballero Díaz. (2011). *Selección de modelos mediante criterios de información en análisis factorial. Aspectos teóricos y computacionales*. Tesis Doctoral. Granada: Universidad de Granada.

[12] Hastie, T.; Tibshirani, R.; Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference and Prediction* New York: Springer. 2001.

[13] Raúl Vaquerizo. `http://analisisydecision.es/regresion-ridge-o-contraida-con-r/`.

[14] R Core Team (2014). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL `http://www.R-project.org/`.

[15] R Studio `http://www.rstudio.org/`.