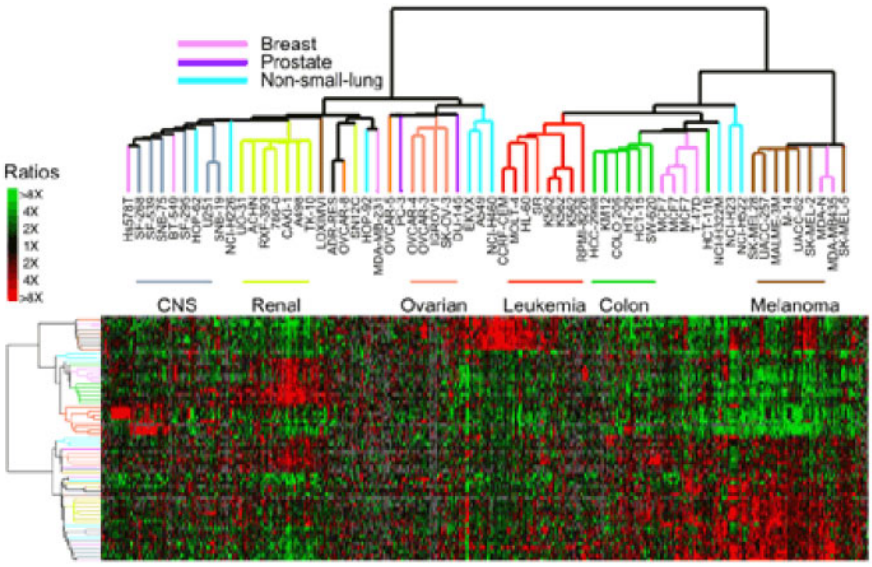
REGRESIÓN EN ESPACIOS DE ALTA DIMENSIÓN

P. Saavedra



Índice general

1. Modelos lineales **1**

1.1. Aproximación a los modelos de regresión lineal 1

1.2. Generalidades sobre los modelos lineales 3

1.2.1. Modelo de regresión lineal simple. 4

1.2.2. Regresión lineal múltiple 4

1.2.3. Análisis de la varianza con un factor de variación 5

1.2.4. Análisis de la varianza con dos factores de variación 5

1.2.5. Análisis de la covarianza 6

1.3. Estimación del modelo lineal 7

1.4. El teorema de Gauss-Markov 10

1.5. El problema de la alta dimensión 10

2. Regularización en regresión lineal **13**

2.1. Penalización con la norma L_2: regresión ridge 13

2.1.1. Existencia del estimador ridge 14

2.1.2. Grados efectivos de libertad 15

2.1.3. Perspectiva Bayesiana de la estimación ridge. 15

2.1.4. Selección del parámetro de contracción: método de validación cruzada 17

2.1.5. Intervalos de confianza para los parámetros: método bootstrap 19

2.2. Penalización L_1: el Lasso 24

2.3. Penalización mixta: elastic net 26

3. Regresión con variables latentes **31**

3.1. Introducción 31

3.2. Componentes principales 32

3.2.1. Aproximación al concepto de componente principal 32
3.2.2. Construcción de las componentes principales 33
3.2.3. Componentes principales como variables latentes 36
3.2.4. Regresión en componentes principales 38
3.3. Mínimos cuadrados parciales (PLS) 39

Capítulo 1

Modelos lineales

1.1. Aproximación a los modelos de regresión lineal

La capacidad de la función renal se mide normalmente mediante el nivel de aclaramiento de creatinina (CCr), la cual tiende a disminuir a partir de cierta edad, tal como se refleja en la siguiente figura.

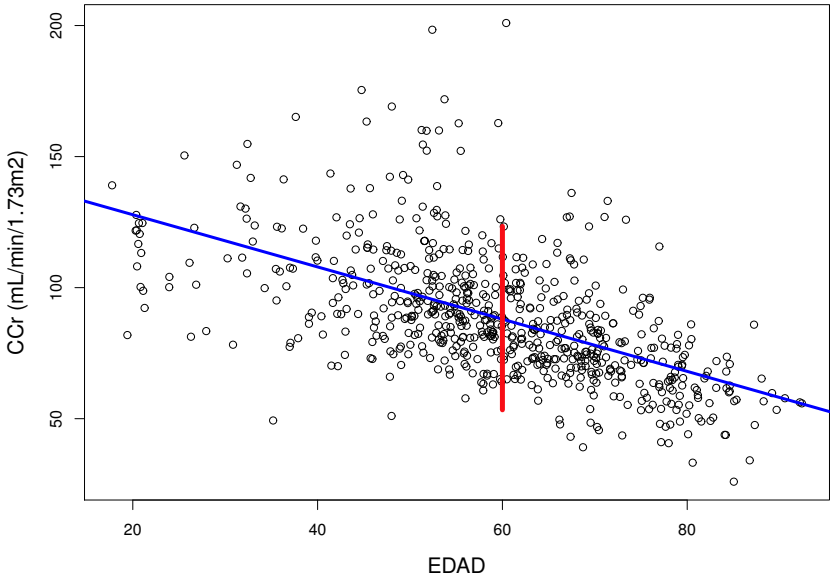


Figura 1.1: Evolución de los niveles de aclaramiento de creatinina según edad

El conjunto de datos que aparece en la figura puede expresarse en la forma:

$$\{(edad, CCr_i) \ : \ i = 1, \ldots, n\}$$

La figura sugiere modelizar estos datos en la forma:

$$CCr_i \sim N\left(\beta_0 + \beta_1 \cdot edad_i \, ; \, \sigma\right) \ : \ i = 1, \ldots, n$$

El diseño del estudio permite asimismo suponer que las variables aleatorias $CCr_1, \ldots, CCr_n$ son independientes. Nótese que $E\left[CCr_i\right] = \beta_0 + \beta_1 \cdot EDAD_i$. Una forma conveniente de expresar el modelo es la que sigue:

$$\begin{pmatrix} CCr_1 \\ \cdots \\ CCr_i \\ \cdots \\ CCr_n \end{pmatrix} \sim N_n \left(\begin{pmatrix} 1 & edad_1 \\ \cdots & \cdots \\ 1 & edad_i \\ \cdots & \cdots \\ 1 & edad_n \end{pmatrix} \cdot \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} \, ; \, \sigma^2 \begin{pmatrix} 1 & \cdots & 0 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & \cdots & 1 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & \cdots & 0 & \cdots & 1 \end{pmatrix} \right)$$

donde la matriz:

$$\mathbf{X} = \begin{pmatrix} 1 & edad_1 \\ \cdots & \cdots \\ 1 & edad_i \\ \cdots & \cdots \\ 1 & edad_n \end{pmatrix}$$

recibe el nombre de *matriz de diseño* y $(CCr_1, \ldots, CCr_n)'$ el de *vector de respuestas*. El problema de estimación consiste en seleccionar un vector $(\beta_0, \beta_1)'$ del espacio vectorial $\mathbb{R}^2$ que represente *adecuadamente* el conjunto de datos (esta es una expresión muy ambigua). En los problemas inferenciales, el llamado espacio paramétrico es el conjunto de *todos los posibles valores del parámetro desconocido*. Nótese que el vector $\boldsymbol{\mu} = \left(E\left[CCr_i\right] \ : \ i = 1, \ldots, n\right)$ puede expresarse en la forma:

$$\boldsymbol{\mu} = \beta_0 \begin{pmatrix} 1 \\ \cdots \\ 1 \\ \cdots \\ 1 \end{pmatrix} + \beta_1 \begin{pmatrix} edad_1 \\ \cdots \\ edad_i \\ \cdots \\ edad_n \end{pmatrix}$$

Por tanto, al recorrer $(\beta_0, \beta_1)'$ el espacio $\mathbb{R}^2$, el parámetro $\boldsymbol{\mu}$ recorre el espacio $\mathbb{R}^n$ pero dentro de la variedad lineal V_2 (dimensión 2) generada por la base:

$$V_2 = \ll (1, \ldots, 1)', (edad_1, \ldots, edad_n)' \gg$$

En este escenario decimos que la *dimensión del problema de regresión* es la dimensión del la variedad lineal en la que varía el parámetro $\boldsymbol{\mu}$. En la sección 1.3 veremos que el estimador de máxima verosimilitud (ML) del parámetro n dimensional $\boldsymbol{\mu}$ es la proyección ortogonal $\hat{\boldsymbol{\mu}}$ del vector de observaciones $\mathbf{CCr} = (CCr_1, \ldots, CCr_n)'$ sobre la variedad lineal V_2. Asimismo, el estimador ML de σ^2 es:

$$\hat{\sigma}^2 = \frac{1}{n} \|\mathbf{CCr} - \hat{\boldsymbol{\mu}}\|_2^2$$

donde $\| (v_1, \ldots, v_n) \|_2 = \sqrt{\sum_{j=1}^n v_j^2} \ : \ v_j \in \mathbb{R}$.

1.2. Generalidades sobre los modelos lineales

A lo largo de toda esta sección consideraremos un vector $\mathbf{Y} = (y_1, \ldots, y_n)'$ de variables aleatorias tales que:

$$\mathbf{Y} \sim \mathbf{N}_n \left(\boldsymbol{\mu}, \sigma^2 \mathbf{I}_n \right)$$

donde $\boldsymbol{\mu} = \boldsymbol{E}[\mathbf{Y}]$. En general, $\boldsymbol{\mu} \in \boldsymbol{V_p}$ (variedad lineal de dimensión p) donde $p < n$ (normalmente $p \ll n$). Si $\mathbf{X}$ es una matriz de dimensión $n \times p$, cuyas columnas corresponden a una base de la variedad V_p , puede entonces expresarse el vector de esperanzas en la forma $\boldsymbol{\mu} = \boldsymbol{X\beta}$ es un vector de parámetros p-dimensional. En el modelo considerado, $\mathbf{I}_n$ representa la matriz identidad de orden

n. Nótese que esta definición supone que las variables aleatorias $y_1, \ldots, y_n$ son incorreladas, y por estar normalmente distribuídas, independientes. Este modelo es el llamado *modelo lineal general* o simplemente, *modelo lineal*. En las siguientes subsecciones consideraremos diversos modelos del diseño experimental clásico que comparten la forma del modelo lineal.

1.2.1. Modelo de regresión lineal simple.

Considérese un conjunto de datos de la forma $\{(x_i, y_i) : i = 1, \ldots, n\}$ y sean las matrices:

$$\mathbf{Y} = \begin{pmatrix} y_1 \\ \cdots \\ y_i \\ \cdots \\ y_n \end{pmatrix} \qquad \mathbf{X} = \begin{pmatrix} 1 & x_1 \\ \cdots & \cdots \\ 1 & x_i \\ \cdots & \cdots \\ 1 & x_n \end{pmatrix} \qquad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}$$

Si estas matrices satisfacen las condiciones del modelo lineal general, se dice entonces que el conjunto de datos anterior obedece al modelo de *regresión lineal simple*. La forma habitual en la que se expresa el modelo es:

$$y_i = \beta_0 + \beta_1 \cdot x_i + e_i \ : \ i = 1, \ldots, n$$

donde $e_1, \ldots, e_n$ son variables aleatorias IID $N(0, \sigma)$.

1.2.2. Regresión lineal múltiple

Considérese ahora el conjunto de datos $\{(x_{i,1}, x_{i,2}, \ldots, x_{i,p}; y_i) : i = 1, \ldots, n\}$ donde $n \gg p$, y las matrices:

$$\mathbf{Y} = \begin{pmatrix} y_1 \\ \cdots \\ y_i \\ \cdots \\ y_n \end{pmatrix} \qquad \mathbf{X} = \begin{pmatrix} 1 & x_{1,1} & x_{1,2} & \ldots & x_{1,p} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 1 & x_{i,1} & x_{i,2} & \cdots & x_{i,p} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 1 & x_{n,1} & x_{n,2} & \cdots & x_{n,p} \end{pmatrix} \qquad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \cdots \\ \beta_p \end{pmatrix}$$

Los datos obedecen al modelo de *regresión lineal múltiple* si la matriz $\mathbf{X}$ es de rango completo ($\text{rango}\,(\mathbf{X}) = p+1$) y además, el vector aleatorio $\mathbf{Y}$ satisface las condiciones del modelo lineal general.

1.2.3. Análisis de la varianza con un factor de variación

Estos modelos surgen el el contexto de un factor F con p-niveles (valores de la variable F). Para el i-ésimo nivel del factor ($i = 1, \ldots, p$) se observan n_i variables aleatorias independientes $y_{i,1}, \ldots, y_{i,n_i}$. El modelo con un factor de variación tiene la forma $y_{i,j} \sim N\,(\theta + \alpha_j, \sigma) \;:\; i = 1, \ldots, p$ siendo $\alpha_1 = 0$ (el primer nivel del factor se elige como nivel de referencia). La forma matricial es ahora:

$$\mathbf{Y} = \begin{pmatrix} y_{1,1} \\ \vdots \\ y_{1,n_1} \\ y_{2,1} \\ \vdots \\ y_{2,n_2} \\ \vdots \\ y_{p,1} \\ \vdots \\ y_{p,n_p} \end{pmatrix} \qquad \mathbf{X} = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & \cdots & 0 \\ 1 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & \cdots & 1 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & \cdots & 1 \end{pmatrix} \qquad \boldsymbol{\beta} = \begin{pmatrix} \theta \\ \alpha_2 \\ \vdots \\ \alpha_p \end{pmatrix}$$

1.2.4. Análisis de la varianza con dos factores de variación

Supóngase ahora que se quiere evaluar el efecto simultáneo de dos factores F (con p niveles) y G (con q niveles) sobre una variable de respuesta Y. Ello daría lugar a $p \times q$ condiciones experimentales (celdas). Un diseño eficiente consistiría en realizar en todas las celdas el mismo número m de observaciones. Representaremos entonces por $y_{i,j,k}$ la k-ésima respuesta observada en la celda (i, j) (i-ésimo nivel para el factor F y j-ésimo nivel para el G. Para el correspondiente conjunto de datos, el modelo de análisis de la varianza con dos factores de variación e interacciones tiene la forma:

$$y_{i,j,k} \sim N\left(\theta + \alpha_i + \gamma_j + \delta_{i,j}, \sigma\right)$$

Aquí, $\alpha_1 = \gamma_1 = \delta_{1,j} = \delta_{i,1} = 0$. Los parámetros α_i y γ_j representan los llamados efectos principales mientras que los parámetros $\delta_{i,j}$ corresponden a las interacciones. Nótese que el vector de respuestas $\mathbf{Y}$ tiene dimensión $p \times q \times m$, mientras que el parámetro $\boldsymbol{\beta}$ es de dimensión $p \times q + 1$. Se deja al lector la formulación matricial del modelo.

1.2.5. Análisis de la covarianza

Estos modelos corresponden a estudios en los que se analiza simultáneamente la influencia de un factor y una covariable numérica. A menudo el propósito del investigador es analizar el efecto de un factor F sobre una respuesta Y, pero tales experimentos no se realizan sen escenarios que varían de acuerdo con una variable X. En general, el conjunto de datos obtenidos del estudio tiene la forma $\{(x_{i,j}, y_{i,j}) \ : \ i = 1, \ldots, p; \ j = 1, \ldots, n_i\}$, siendo $y_{i,j}$ variables aleatorias independientes tales que:

$$y_{i,j} \cong N\left(\theta + \alpha_i + \gamma x_{i,j}, \ \sigma\right) \quad i = 1, \ldots, p; \ j = 1, \ldots, n_i$$

La formulación matricial del modelo es entonces:

$$\mathbf{Y} = \begin{pmatrix} y_{1,1} \\ \vdots \\ y_{1,n_1} \\ y_{2,1} \\ \vdots \\ y_{2,n_2} \\ \vdots \\ y_{p,1} \\ \vdots \\ y_{p,n_p} \end{pmatrix} \quad \mathbf{X} = \begin{pmatrix} 1 & 0 & \cdots & 0 & x_{1,1} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & \cdots & 0 & x_{1,n_1} \\ 1 & 1 & \cdots & 0 & x_{2,1} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 1 & \cdots & 0 & x_{2,n_2} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & \cdots & 1 & x_{p,1} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & \cdots & 1 & x_{p,n_p} \end{pmatrix} \quad \boldsymbol{\beta} = \begin{pmatrix} \theta \\ \alpha_2 \\ \vdots \\ \alpha_p \\ \gamma \end{pmatrix}$$

1.3. Estimación del modelo lineal

Consideramos ahora el problema de la estimación del modelo lineal $\mathbf{Y} \sim \mathbf{N}_n(\boldsymbol{\mu}, \sigma^2\mathbf{I}_n)$, donde $\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}$, con $\boldsymbol{\beta} \in \mathbb{R}^p$. Supondremos que la matriz de diseño $\mathbf{X}$ es de rango completo p, lo que implica que $\mathbf{X}'\mathbf{X}$ es invertible. Puede deducirse que la función de log-verosimilitud tiene la forma:

$$\ell\left(\boldsymbol{\beta}, \sigma^2\right) = -\frac{n}{2}\log\left(2\pi\sigma^2\right) - \frac{1}{2\sigma^2} \parallel \mathbf{Y} - \mathbf{X}\boldsymbol{\beta} \parallel_2^2$$

El problema de maximizar la verosimilitud $\ell(\boldsymbol{\beta}, \sigma^2)$ supone resolver el problema de mínimos cuadrados:

$$\min_{\boldsymbol{\beta}} \parallel \mathbf{Y} - \mathbf{X}\boldsymbol{\beta} \parallel_2^2$$

Este problema es lineal y puede obtenerse fácilmente la siguiente estimación para el parámetro $\boldsymbol{\beta}$:

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

Se tiene entonces que:

$$\hat{\boldsymbol{\mu}} = \mathbf{X}\hat{\beta} = \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-\mathbf{1}}\mathbf{X}'\mathbf{Y} = \pi_{V_p}(\mathbf{Y})$$

donde V_p es la variedad lineal generada por las columnas de la matriz de diseño $\mathbf{X}$ y π_{V_p} representa la proyección ortogonal sobre V_p. Es interesante notar que:

$$\text{traza}\Big(\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\Big) = p$$

Esto es: *la traza de la matriz de proyección ortogonal sobre el subespacio V_p coincide con la dimensión de V_p.*

Es fácil comprobar que el estimador es centrado. En efecto:

$$E\left[\widehat{\boldsymbol{\beta}}\right] = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'E\left[\mathbf{Y}\right] = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{X}\boldsymbol{\beta} = \boldsymbol{\beta}$$

lo que supone obviamente que $E[\hat{\boldsymbol{\mu}}] = \boldsymbol{\mu}$.

La matriz de covarianzas del estimador se obtiene entonces como:

$$\text{var}\left(\widehat{\boldsymbol{\beta}}\right) = E\left[\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)'\right] =$$

$$(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{E}\left[(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})'\right]\mathbf{X}(\mathbf{X}'\mathbf{X})^{-\mathbf{1}} = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$$

Puede comprobarse que $\text{var}\left(\hat{\beta}_i\right) \to 0$ para $n \to \infty$, lo que significa que el estimador es consistente.

El estimador de máxima verosimilitud para la varianza del modelo σ^2 es:

$$\hat{\sigma}^2 = \frac{1}{n} \parallel \mathbf{Y} - \mathbf{X}\,\hat{\boldsymbol{\beta}} \parallel_{\mathbf{2}}^{\mathbf{2}}$$

Puede probarse que $\parallel \mathbf{Y} - \mathbf{X}\,\hat{\boldsymbol{\beta}} \parallel_{\mathbf{2}}^{\mathbf{2}} / \sigma^2 \cong \chi^2(n-p)$. Ello implica que el estimador $\hat{\sigma}^2$ no es centrado. En efecto:

$$E\left[\hat{\sigma}^2\right] = \frac{\sigma^2}{n} E\left[\parallel \mathbf{Y} - \mathbf{X}\,\hat{\boldsymbol{\beta}} \parallel_{\mathbf{2}}^{\mathbf{2}} / \sigma^2\right] = \frac{n-p}{n}\sigma^2$$

Ello significa que el sesgo del estimador es $-p\sigma^2/n$. Obviamente, un estimador alternativo centrado para σ^2 es:

$$S^2 = \frac{1}{n-p} \parallel \mathbf{Y} - \mathbf{X}\,\widehat{\boldsymbol{\beta}} \parallel_{\mathbf{2}}^{\mathbf{2}}$$

Ejemplo 1.1. Considérese un estudio experimental cuya finalidad es evaluar dos tratamientos T_1 y T_2. Para tal fin se dispone de tres unidades experimentales $\omega_1, \omega_2, \omega_3$. El ensayo se lleva a efecto de acuerdo con la siguiente diseño: ω_1 recibe T_1, ω_2 recibe T_2 y ω_3 no recibe ningún tratamiento. Supondremos entonces que el vector de respuestas $\mathbf{Y} = (y_1, y_2, y_3)'$ es tal que:

$$\begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix} \sim N_3\left(\begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{pmatrix}\begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix}; \sigma^2 \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}\right)$$

Nótese entonces que $E[\mathbf{Y}] = (\beta_1, \beta_2, 0)'$ y de esta forma, recorre el plano XY (espacio V_2).

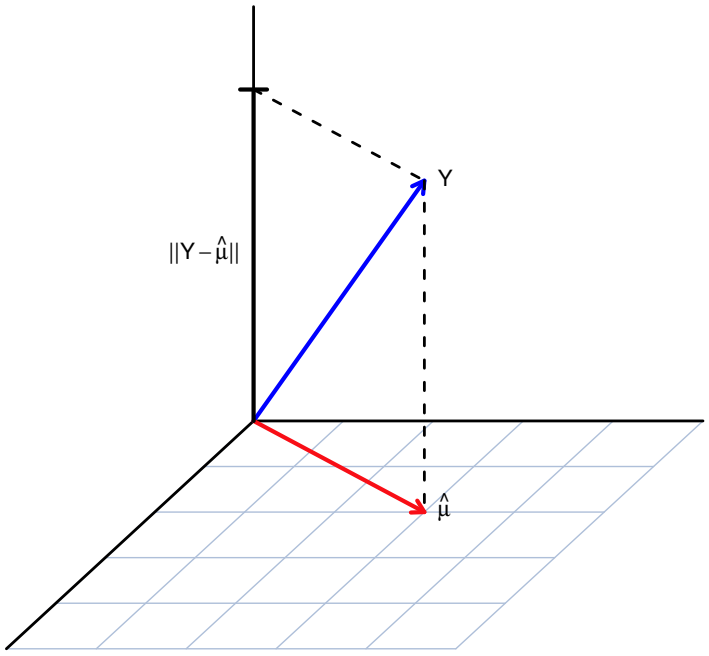


Figura 1.2:

```
library(MASS)
X = matrix(c(1, 0, 0, 0, 1, 0), ncol = 2)
b = c(2, 3)
dim(b) = c(2, 1)
I = diag(3)
sg = 0.5
Y = mvrnorm(1, X %*% b, sg * I)
```

1.4. El teorema de Gauss-Markov

Considérese la generalización del modelo lineal descrito en la sección 1.2 consistente en suprimir la hipótesis de normalidad. Más concretamente, supóngase que se observa un vector $\mathbf{Y} = (y_1, \ldots, y_n)'$ de variables aleatorias tal que:

1. $E[\mathbf{Y}] = \mathbf{X}\beta$, siendo la matriz de diseño $\mathbf{X}$ de dimensión $n \times p$ $(n > p)$ y rango p.

2. $\text{cov}(\mathbf{Y}) = \sigma^2 \mathbf{I}_n$.

En tal escenario, consideramos el problema de estimar el parámetro $\theta = a'\beta$, siendo $a \in \mathbb{R}^p$. El teorema de Gauss-Markov establece que el estimador $\hat{\theta} = a'\hat{\beta}$ es centrado para θ y *es el mejor entre todos los estimadores lineales centrados* de θ; esto es: $\forall \tilde{\theta} = b'\mathbf{Y}$ $(b \in \mathbb{R}^n)$ tal que $E\left[\tilde{\theta}\right] = \theta$ se satisface:

$$\text{var}\left(a'\hat{\beta}\right) \leq \text{var}\left(\tilde{\theta}\right)$$

1.5. El problema de la alta dimensión

En general, el problema de mínimos cuadrados para el modelo lineal descrito en la sección 1.2 conduce al sistema lineal de ecuaciones:

$$\mathbf{(X'X)}\ \boldsymbol{\beta} = \mathbf{X'Y}$$

Tal como se indicó en la sección 1.3, en aquellos casos en los que la matriz de diseño $\mathbf{X}$ es de rango máximo, el sistema tiene como solución única el estimador ordinario de mínimos cuadrados (OLS) $\hat{\boldsymbol{\beta}}$. Si ocurre que $n = p$, obviamente $\hat{\boldsymbol{\beta}} = \mathbf{Y}$, y por tanto, no es posible estimar σ^2. Si finalmente $n < p$, la matriz $\mathbf{X'X}$ no es invertible y por tanto, no existe el estimador OLS.

Consideramos ahora situaciones en las que, aún siendo $n > p$ ocurre que $n \sim p$. En tal caso, el estimador $\hat{\beta}$ se hace inestable en el sentido de que aumenta su varianza de forma inaceptable (aunque obviamente sigue siendo centrado). Para ilustrar esta situación consideramos B conjuntos de datos generados de acuerdo con la siguiente pauta:

$$y_i = \sum_{j=1}^{p} \beta_j x_{i,j} + e_i \quad : \quad i = 1, \ldots, n$$

Aquí, $e_1, \ldots, e_n$ son variables aleatorias IID . $N(0, \sigma)$. Se ha considerado $n = 100$ y $\beta_i = 1 \quad : \quad i = 1, \ldots, p$. Como valores de p se han tomado sucesivamente 5, 60 y 90.

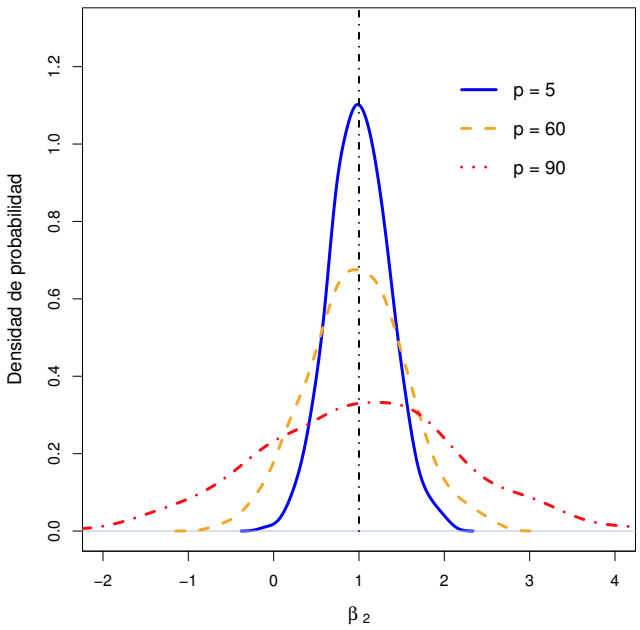


Figura 1.3: Distribuciones de probabilidad del estimador $\hat{\beta}_2$ para diferentes dimensiones del problema

Nótese que la dimensión $p = 90$ es excesivamente alta para un tamaño muestral $n = 100$ lo que llevará a estimaciones altamente inestables de los parámetros β_i; esto es, con una varianza muy elevada. En la figura 1.3 se muestra la densidad de probabilidad del estimador $\hat{\beta}_2$, para los tres valores de n considerados (las distribuciones de los estimadores se han obtenido tomando $B = 10,000$).

```
b2t <- function(p) {
    b = cbind(rep(1, p))
    X = matrix(runif(n * p), n, p)
    Y = X %*% b + rnorm(nrow(X))
```

```
    ml <- lm(Y ~ X)
    return(ml$coef[3])
}
B = 1000
n = 100
```

El hecho de que los estimadores $\hat{\beta}_j$ sean inestables no significa necesariamente que lo sean las predicciones $\hat{\mu}_i = \sum_{j=1}^{p} \hat{\beta}_j X_{i,j}$, pues normalmente se producen compensaciones entre las estimaciones $\hat{\beta}_j$. En cualquier caso, el investigador no podrá medir adecuadamente los efectos de las variables $X_{i,j}$ sobre las respuestas μ_i.

Capítulo 2

Regularización en regresión lineal

En el escenario de los problemas de alta dimensión ($p < n$, aunque $p \sim n$) la consecuencia más inmediata es que, aunque los estimadores de máxima verosimilitud de los parámetros del modelo son centrados, sus varianzas son muy elevadas y por tanto, tales estimaciones son muy inestables. La estimación *ridge* tiene como finalidad paliar este problema penalizando la verosimilitud mediante la métrica L_2. Tal procedimiento permitirá reducir notablemente la varianza de los estimadores aunque éstos sean sesgados, pero se reducirá el error cuadrático medio. La idea clave de la regresión ridge consiste en controlar la norma L_2 del estimador del vector de parámetros.

2.1. Penalización con la norma L_2: regresión ridge

El procedimiento de estimación ridge para el modelo lineal fue propuesto originalmente por Hoerl y Kennard (1962, 1970) para resolver los problemas de la multicolinealidad. Consideramos por tanto nuevamente el modelo $\mathbf{Y} \sim \mathbf{N}_n(\mu, \sigma^2\mathbf{I}_n)$, siendo $\mu = \mathbf{X}\boldsymbol{\beta}$. Supondremos sin pérdida de generalidad que $\sum_{i=1}^n y_i = 0$, $\sum_{i=1}^n x_{i,j} = 0$ y $\sum_{i=1}^n x_{i,j}^2/n = 1$ para $j = 1, \ldots, p$.

En este escenario, consideramos el problema de minimización de mínimos cuadrados penalizados de la forma:

$$\min_{\boldsymbol{\beta}} \parallel \mathbf{Y} - \mathbf{X}\boldsymbol{\beta} \parallel_2^2 + \lambda \parallel \boldsymbol{\beta} \parallel_2^2$$

siendo $\lambda \geq 0$. Este problema se puede formular alternativamente como:

$$\min_{\boldsymbol{\beta}} \parallel \mathbf{Y} - \mathbf{X}\boldsymbol{\beta} \parallel_2^2$$

sujeto a la restricción: $\parallel \boldsymbol{\beta} \parallel_2^2 \leq t$, para algún $t = t(\lambda)$.

Nótese que la solución del problema dependerá del valor seleccionado para el parámetro λ. Para $\lambda = 0$, el problema se reduce al estimador ordinario de mínimos cuadrados (máxima verosimilitud). El incremento del valor λ conducirá obviamente a la *contracción* del estimador del parámetro β. Por tanto, λ regula la cantidad de contracción que se pretende imponer al parámetro. Ello obviamente conducirá a la reducción de la varianza del estimador, pero éste será sesgado. El estimador ridge tiene la forma:

$$\hat{\boldsymbol{\beta}}^{\text{ridge}} = (\mathbf{X}'\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}'\mathbf{Y}$$

siendo $\mathbf{I}$ la matriz identidad.

Cuando la matriz se diseño $\mathbf{X}$ es ortogonal, es fácil comprobar que el estimador ridge es una contracción del estimador ordinario de mínimos cuadrados en la forma:

$$\hat{\boldsymbol{\beta}}^{\text{ridge}} = \frac{\hat{\boldsymbol{\beta}}}{1+\lambda}$$

donde $\hat{\boldsymbol{\beta}}$ es el estimador ordinario de mínimos cuadrados.

2.1.1. Existencia del estimador ridge

En orden a examinar las propiedades del estimador ridge reduciremos en primer lugar el modelo a la forma canónica en el modo que sigue:

1. Sean $d_1 \geq d_2 \geq \cdots \geq d_p \geq 0$ los autovalores de la matriz $\mathbf{X}'\mathbf{X}$.

2. Considérese la diagonalización $\mathbf{X}'\mathbf{X} = \mathbf{P}'\mathbf{D}\mathbf{P}$, donde $\mathbf{D} = \text{diag}(d_1, d_2, \ldots d_p)$ y $\mathbf{P}$ es una matriz ortonormal.

3. Sea $\mathbf{Z} = \mathbf{X}\mathbf{P}'$ y $\boldsymbol{\alpha} = \mathbf{P}\boldsymbol{\beta}$. Nótese que $\mathbf{Z}'\mathbf{Z} = \mathbf{P}\mathbf{X}'\mathbf{X}\mathbf{P}' = \mathbf{D}$.

   Se tiene entonces que $\mathbf{X}\boldsymbol{\beta} = \mathbf{Z}\boldsymbol{\alpha}$ y $\boldsymbol{\alpha}'\boldsymbol{\alpha} = \boldsymbol{\beta}'\boldsymbol{\beta}$. De esta forma se satisface la siguiente identidad:

$$\| \mathbf{Y} - \mathbf{X}\boldsymbol{\beta} \|_2^2 + \lambda \| \boldsymbol{\beta} \|_2^2 = \| \mathbf{Y} - \mathbf{Z}\boldsymbol{\alpha} \|_2^2 + \lambda \| \boldsymbol{\alpha} \|_2^2$$

Luego la existencia del estimador ridge para $\boldsymbol{\alpha}$ es equivalente a la existencia de $\hat{\boldsymbol{\beta}}^{\text{ridge}}$. El estimador ridge para $\boldsymbol{\alpha}$ es:

$$\hat{\boldsymbol{\alpha}}^{\text{ridge}} = (\mathbf{D} + \lambda \mathbf{I})^{-1} \mathbf{Z}'\mathbf{Y} = \text{diag}\left(\frac{1}{d_1 + \lambda}, \cdots, \frac{1}{d_p + \lambda} \right) \mathbf{Z}'\mathbf{Y}$$

Nótese que si $d_1 > 0$, entonces el estimador $\hat{\boldsymbol{\alpha}}^{\text{ridge}}$ está bien definido. En caso contrario, el estimador existe para cualquier $\lambda > 0$.

2.1.2. Grados efectivos de libertad

A partir del estimador ridge para $\boldsymbol{\beta}$, el estimador ridge del parámetro μ es:

$$\hat{\boldsymbol{\mu}}^{\text{ridge}} = \mathbf{X} \left(\mathbf{X}'\mathbf{X} + \lambda \mathbf{I} \right)^{-1} \mathbf{X}'\mathbf{Y}$$

Obviamente $\hat{\boldsymbol{\mu}}^{\text{ridge}} \in V_p$, pero no es la proyección ortogonal sobre este subspacio. El número de los *grados efectivos de libertad* se define entonces por:

$$df\,(\lambda) = \text{traza}\left(\mathbf{X} \left(\mathbf{X}'\mathbf{X} + \lambda \mathbf{I} \right)^{-1} \mathbf{X}' \right)$$

Puede comprobarse que:

$$df\,(\lambda) = \sum_{j=1}^{p} \frac{d_j^2}{d_j^2 + \lambda}$$

De esta forma, entre mayor es el grado de penalización, *menor es el espacio* en el que se mueve la estimación ridge

2.1.3. Perspectiva Bayesiana de la estimación ridge.

Consideramos ahora el modelo de regresión lineal $\mathbf{Y} \sim N_n\left(\mathbf{X}\boldsymbol{\beta}, \sigma^2 \mathbf{I}_n \right)$ en un escenario bayesiano en el que $\beta_1, \ldots, \beta_p$ son variables aleatorias independientes e identicamente distribuidas, siendo $\beta_i \sim N\left(0, \tau\right)$. Entonces,

$$\hat{\beta}^{\text{bayes}} = E\left[\beta \mid \mathbf{X}, \mathbf{Y} \right] = \hat{\boldsymbol{\beta}}^{\text{ridge}}$$

con $\lambda = \sigma^2 / \tau^2$.

Ejemplo 2.1. Considérese ahora una simulación de $B = 2000$ réplicas del modelo de regresión descrito en el ejemplo 1.1. En la figura 2.1 se muestran las correspondientes estimaciones ridge para de $\lambda = 0$, 0.03, 0.07 y 0.2. Los correspondientes errores cuadráticos medios observados fueron proporcionales a 805.8, 758.9, 816.7 y 1284.1 (ver tabla 2.1).

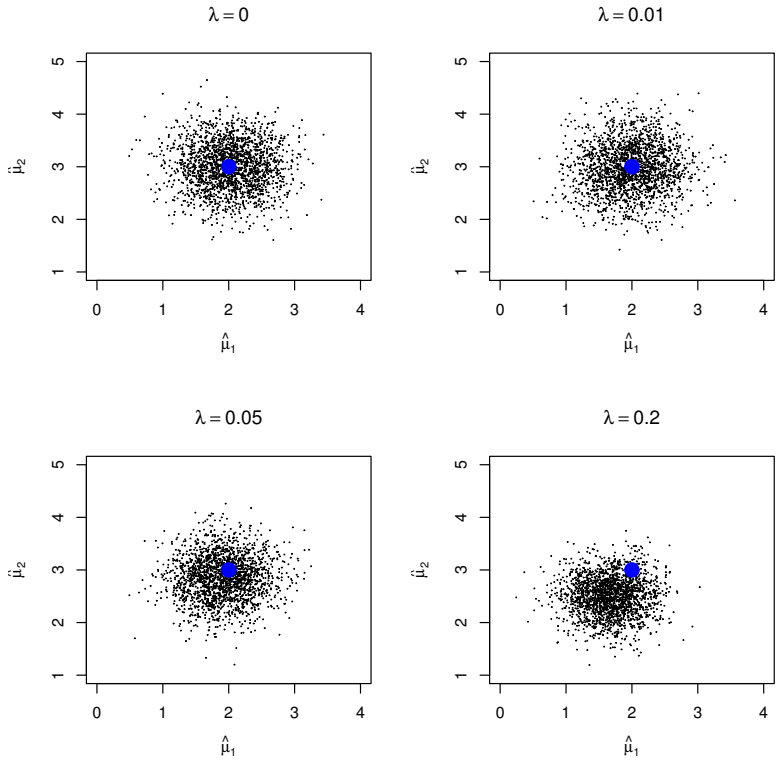


Figura 2.1: Estimación rige para diferentes valores del parámetro de *contracción*

λ	$df\,(\lambda)$	MSE
0	2	819.9
0.01	1.980	794.9
0.05	1.905	817.2
0.20	1.667	1292.0

Tabla 2.1: Error cuadrático medio según parámetro de *contracción*

Ejemplo 2.2. Consideramos ahora un modelo de regresión lineal con dos predictores entre los que existe una fuerte asociación lineal. El modelo se ha si-

mulado con la siguiente pauta: $x_{i,1} \sim N(0,1)$, $x_{i,2} \sim N(x_{i,1}, 0{,}2)$ y finalmente, $y_i \sim N(3 + x_{i,1} + x_{i,2}, 0{,}3)$. De esta forma, $\beta_0 = 3$, $\beta_1 = 1$ y $\beta_2 = 1$. Nótese la fuerte asociación entre los predictores $x_{i,1}$ y $x_{i,2}$. Ello da lugar a que los estimadores de los parámetros tengan una varianza muy elevada.

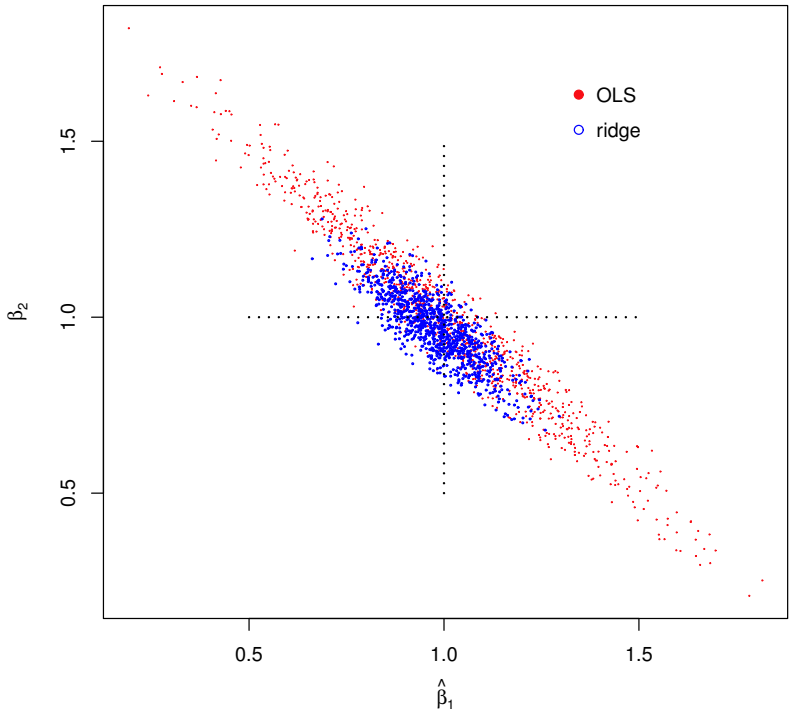


Figura 2.2:

2.1.4. Selección del parámetro de contracción: método de validación cruzada

La selección del parámetro de contracción λ es uno de los problemas fundamentales de la regresión ridge. Si un *oráculo* nos informase del verdadero valor de $\boldsymbol{\mu}$, una elección natural para λ sería la solución del problema:

$$\min_{\lambda} \parallel \hat{\boldsymbol{\mu}}^{\text{ridge}}(\lambda) - \boldsymbol{\mu} \parallel^2$$

donde $\hat{\boldsymbol{\mu}}^{\text{ridge}}(\lambda)$ expresa la dependencia del estimador ridge de λ. Dado que μ es desconocido, podría optarse por la solución *ingenua*:

$$\min_{\lambda} \parallel \mathbf{Y} - \hat{\boldsymbol{\mu}}^{\text{ridge}}(\lambda) \parallel^2$$

Esta solución sería la que mejor predeciría los datos de *entrenamiento* (los utilizados para la estimación del modelo), pero no la que predeciría un subconjunto independiente de datos. Lo ideal sería por tanto disponer de suficientes datos que permitieran reservar una parte (por ejemplo, el 30 %) para la validación del modelo.

Para fijar ideas, sea $(\mathbf{X}_l; \mathbf{Y}_l)$ y $(\mathbf{X}_v; \mathbf{Y}_v)$ los conjuntos de datos de entrenamiento y validación respectivamente. Para la selección de λ consideramos el siguiente procedimiento:

1. Para cada valor de λ, obtenemos con los datos de entrenamiento el estimador ridge: $\hat{\boldsymbol{\beta}}^{\text{ridge}}(\lambda)$.

2. Obtenemos entonces el error de predicción como:

$$\text{cv}(\lambda) = \parallel \mathbf{Y}_v - \mathbf{X}_v \hat{\boldsymbol{\beta}}^{\text{ridge}}(\lambda) \parallel^2$$

3. Seleccionamos entonces λ como solución del problema $\min_\lambda \text{cv}(\lambda)$.

La solución anterior no es viable cuando el número de datos es escaso. En tal caso, puede utilizarse alternativamente el método k-fold validación cruzada (*k-fold cross-validation*). Este método puede describirse brevemente de la siguiente forma:

1. Se fija un valor entero k (normalmente 5 ó 10). Se divide entonces aleatoriamente el conjunto de datos en r subconjuntos $\{(\mathbf{X}_1; \mathbf{Y}_1)\}, \ldots, \{(\mathbf{X}_j; \mathbf{Y}_j)\}, \ldots, \{(\mathbf{X}_r; \mathbf{Y}_r)\}$ de tal forma que $\{(\mathbf{X}_j; \mathbf{Y}_j)\}$ tiene tamaño k, $\forall j = 1, \ldots, r$.

2. Se fija un valor $\lambda \geq 0$.

3. Para $j = 1, \ldots, r$ consideramos como *datos de entrenamiento* $\cup_{i \neq j} \{(\mathbf{X}_i; \mathbf{Y}_i)\}$ y como *datos de validación* $\{(\mathbf{X}_j; \mathbf{Y}_j)\}$.

4. De acuerdo con el algoritmo descrito anteriormente, se obtiene entonces el estimador $\hat{\boldsymbol{\beta}}^{\text{ridge}}_{(-j)}(\lambda) : j = 1, \ldots, r$ y finalmente, el correspondiente error asociado a λ definido por:

$$\mathrm{cv}\,(\lambda) = \frac{1}{r}\sum_{j=1}^{r} \parallel \mathbf{Y}_j - \mathbf{X}_j\hat{\boldsymbol{\beta}}^{\mathrm{ridge}}_{(-j)}\,(\lambda) \parallel^2$$

5. El λ óptimo se define como $\lambda_{\mathrm{opt}} = \arg\min \mathrm{cv}\,(\lambda)$.

2.1.5. Intervalos de confianza para los parámetros: método bootstrap

Los paquetes estadísticos que incorporan procedimientos para la regresión ridge normalmente no proporcionan los errores estándar de los estimadores de los coeficientes. Ello se debe a que éstos no son muy útiles a la hora de obtener los intervalos de confianza dado los fuertes sesgos que en general producen las estimaciones ridge. En tal escenario, el *bootstrap* (Efron, 1979) parece un procedimiento prometedor para la obtención de los referidos interrvalos de confianza.

Para el modelo $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$, donde $\mathbf{ee}' = \sigma^{\mathbf{2}}\mathbf{I}_n$, proponemos ahora un algoritmo basado en el *remuestreo* de los residuales para obtener una aproximación a la ley de probabilidad del pivotal $\hat{\boldsymbol{\beta}}^{\mathrm{ridge}} - \boldsymbol{\beta}$.

1. Obtenemos el estimador ridge $\hat{\boldsymbol{\beta}}^{\mathrm{ridge}}$ y calculamos los residuales $\hat{\mathbf{e}} = \mathbf{Y} - \mathrm{X}\hat{\boldsymbol{\beta}}^{\mathrm{ridge}}$.

2. Del vector de residuales ê se selecciona una muestra aleatoria con reemplazamiento e* y obtenemos $\mathrm{Y}^* = \mathrm{X}\hat{\boldsymbol{\beta}}^{\mathrm{ridge}} + \mathrm{e}^*$.

3. Aplicando la regresión ridge al modelo $\mathrm{Y}^* = \mathrm{X}\hat{\boldsymbol{\beta}}^{\mathrm{ridge}} + \mathrm{e}^*$ se obtiene el estimador bootstrap $\boldsymbol{\beta}^*$.

4. Replicando B veces los pasos 2 y 3 se obtienen B valores de $\boldsymbol{\beta}^* - \hat{\boldsymbol{\beta}}^{\mathrm{ridge}}$, las cuales proporcionan una aproximación a la ley de probabilidad de $\hat{\boldsymbol{\beta}}^{\mathrm{ridge}} - \boldsymbol{\beta}$.

Freedman (1981), analiza la consistencia de este tipo de aproximaciones bootstrap utilizando las métricas de Mallows.

Consideramos ahora un conjunto de datos de la forma $\{(x_{i,1}, x_{i,2}; y_i) : i = 1, \ldots, 30\}$ generados por el modelo descrito en el ejemplo 2.2. Recuérdese que áquel presentaba una situación extrema de multicolinealidad. En la tabla 2.2 se

muestran las estimaciones de mínimos cuadrados de los coeficientes del modelo junto a sus intervalos de confianza al 95 %. Se muestran también sus estimaciones ridge y los intervalos de confianza al 95 % obtenidos por el método bootstrap descrito anteriormente. El parámetro de contracción λ se determinó utilizando el método k-fold de validación cruzada dado en 2.1.4. Nótese como las estimaciones de mínimos cuadrados de los parámetros β_1 y β_2 se desvían severamente de los verdaderos valores ($\beta_1 = \beta_2 = 1$). Las estimaciones ridge por el contrario están mucho más próximas a los verdaderos valores de los parámetros. Nótese además que el estimador $\hat{\beta}_2$ no tiene significación estadística, pero si la tiene $\hat{\beta}_2^{\text{ridge}}$. Obsérvese finalmente que los intervalos de confianza bootstrap obtenidos para la regresión ridge son notablemente más cortos que los obtenidos para la regresión ordinaria.

	Mínimos cuadrados		Ridge	
	Coeficiente	IC-95 %	Coeficiente	IC-95 %
X_1	1.694	[1.065 ; 2.323]	1.098	[1.121 ; 1.292]
X_2	0.415	[-0.166 ; 0.995]	0.871	[0.781 ; 0.932]

Tabla 2.2: Estimaciones de mínimos cuadrados y ridge para un conjunto de datos simulados por el modelo descrito en el ejemplo 2.2. Los intervalos de confianza para la rebresión ridge se obtuvieron utilizando el algoritmo bootstrap dado en 2.1.5.

Ejemplo 2.3. La hormona paratiroidea o parathormona (PTH) favorece la activación de los osteoclastos y secundariamente la de los osteoblastos, aumentando el recambio óseo. No obstante, el equilibrio final es ligeramente negativo, por lo que el resultado global de la PTH sobre la masa ósea es perjudicial. Sin embargo, cuando la acción de la PTH es intermitente, el efecto es positivo. Se cree que los osteoclastos maduros no responden a la PTH directamente, sino a través de las señales que envían las células de estirpe osteoblástica, que poseen receptores para esta hormona. Hoy se sabe que los osteoclastos y preosteoclastos expresan receptores para la PTH, por lo que no puede excluirse la posibilidad de un efecto directo de esta hormona.

Para analizar este problema hemos utilizado un conjunto de datos obtenidos de un estudio transversal en el que se incluyeron 1097 mujeres postmenopáusicas con edades comprendidas entre los 29 y 90 años (media de 57.9 años). Consideramos

el subconjunto de la forma:

$$\{(edad_i, pth_i\,;\, l2l4_i) \;:\; i = 1, \ldots, 1097\}$$

Aquí, PTH representa el valor de la parathormona (pg/ml) y L2L4 es la determinación media de la densidad mineral ósea (DMO) en las lumbares 2, 3 y 4 obtenida por el método de absorciometría dual de rayos X (DXA). Tal como cabía esperar, los niveles de PTH se asocian con la disminución de la DXA (tabla 2.3, modelo 1). En cualquier caso, tal asociación podría explicarse por un efecto de confusión de la edad, pues con ésta, la DMO tiende a disminuir y la PTH a aumentar. Por tal motivo, realizamos el ajuste por edad utilizando el método de los mínimos cuadrados ordinarios (modelo 2). Puede observarse que al realizar el referido ajuste, la asociación PTH-DMO pierde la significación estadística (el intervalo de confianza al 90 % contiene al cero). Tal pérdida de significación estadística podría ser consecuencia de la correlación existente entre la edad y la PTH (el coeficiente de correlacion de Pearson es de 0.25). Al realizar la regresión ridge (tabla 2.3, modelo 3), el intervalo de confianza para el coeficiente de la PTH se reduce, pero sigue conteniendo al cero. Este resultado es coherente con el método lasso que se estudiará en la siguiente sección.

	Mínimos cuadrados (1)		Mínimos cuadrados (2)		Ridge (3)	
	Coef.	IC-90 %	Coef.	IC-90 %	Coef.	IC-90 %
PTH	-0.992	-1.488 ; -0.496	-0.438	-0.939 ; 0.062	-0.442	-0.920 ; 0.049
Edad	-	-	-3.527	-4.326 ; -2.728	-3.411	-4.415 ; -2.760

Tabla 2.3: Modelos para DMO en L2-L4 según valores de la PTH y Edad

Ejemplo 2.4. Stamey *et al* (1989) analizaron la correlación entre el nivel de antígeno prostático específico (PSA) y una serie de marcadores clínicos en 97 varones a los que se les iba a practicar una prostatectomía radical. Analizaremos los efectos simultáneos de los marcadores sobre los niveles del PSA (lpsa) a partir de una serie de medidas como el volumen del tumor (lcavol), el log-peso de la próstata (lweight), la edad (age), la cantidad de hiperplasia prostática benigna (lbph), el grado de infiltración del tumor en la vesícula seminal (svi), el grado de penetración capsular (lcp) y el score de Gleason (gleason). El conjunto de los 97

registros fue aleatoriamente dividido en un subconjunto de datos de entrenamiento ($n = 67$) y otro de test ($n = 30$).

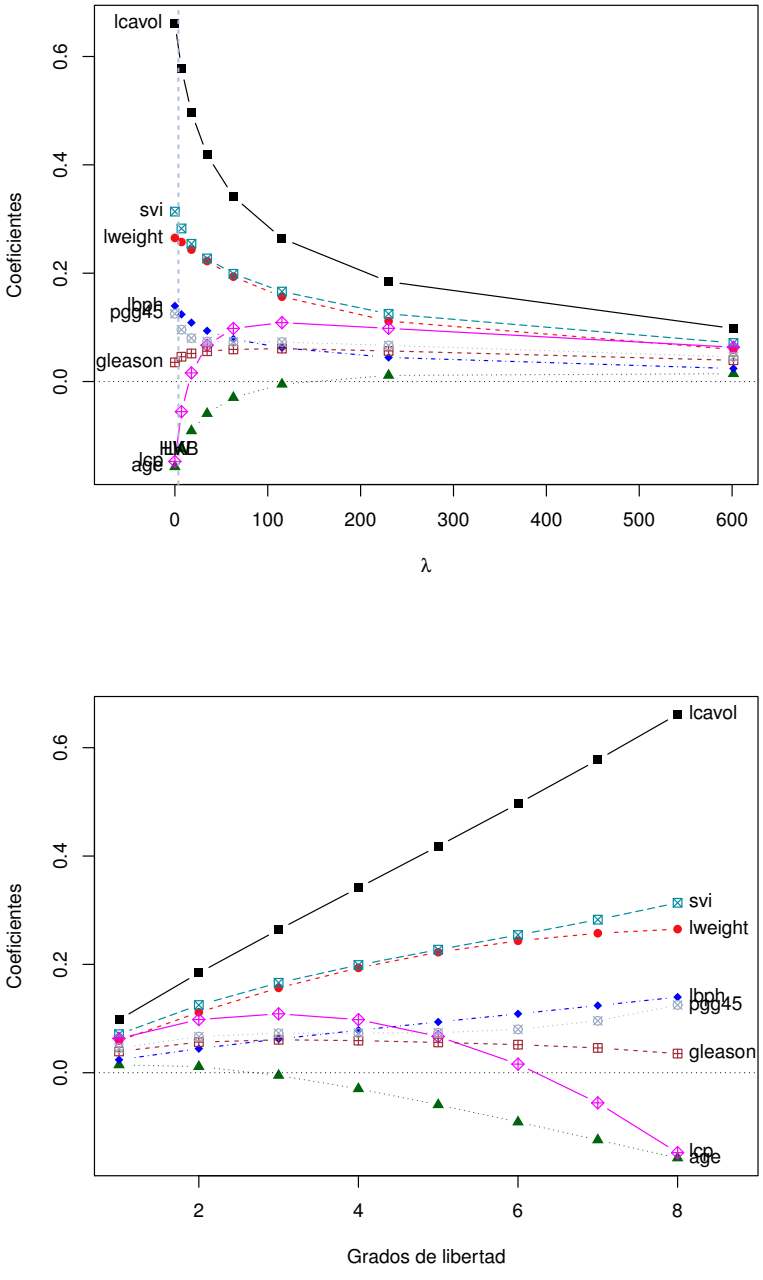


Figura 2.3:

```
library(ElemStatLearn)
data(prostate)
attach(prostate)
learn = subset(prostate, train == TRUE)
test = subset(prostate, train == FALSE)
p = 8  #N<U+00BA> de variables
X <- data.matrix(prostate[, 1:p])
lpsa_l = learn[, p + 1]
X_l = data.matrix(learn[, 1:p])
lpsa_t = test[, p + 1]
X_t = data.matrix(test[, 1:p])
correl = cor(learn[, 1:8])
```

	lcavol	lweight	age	lbph	svi	lcp	gleason	pgg45
lcavol	1.0000	0.3002	0.2863	0.0632	0.5929	0.6920	0.4264	0.4832
lweight	0.3002	1.0000	0.3167	0.4370	0.1811	0.1568	0.0236	0.0742
age	0.2863	0.3167	1.0000	0.2873	0.1289	0.1730	0.3659	0.2758
lbph	0.0632	0.4370	0.2873	1.0000	-0.1391	-0.0885	0.0330	-0.0304
svi	0.5929	0.1811	0.1289	-0.1391	1.0000	0.6712	0.3069	0.4814
lcp	0.6920	0.1568	0.1730	-0.0885	0.6712	1.0000	0.4764	0.6625
gleason	0.4264	0.0236	0.3659	0.0330	0.3069	0.4764	1.0000	0.7571
pgg45	0.4832	0.0742	0.2758	-0.0304	0.4814	0.6625	0.7571	1.0000

```
library(MASS)
lambda = seq(0, 7, 0.1)
pr <- lm.ridge(lpsa_l ~ X_l, lambda = lambda)
select(pr)

## modified HKB estimator is 3.356
## modified L-W estimator is 3.051
## smallest value of GCV  at 4.9

nGCV <- which.min(pr$GCV)
(lGCV <- pr$lambda[nGCV])

## [1] 4.9

pro <- lm.ridge(lpsa_l ~ X_l, lambda = lGCV)
```

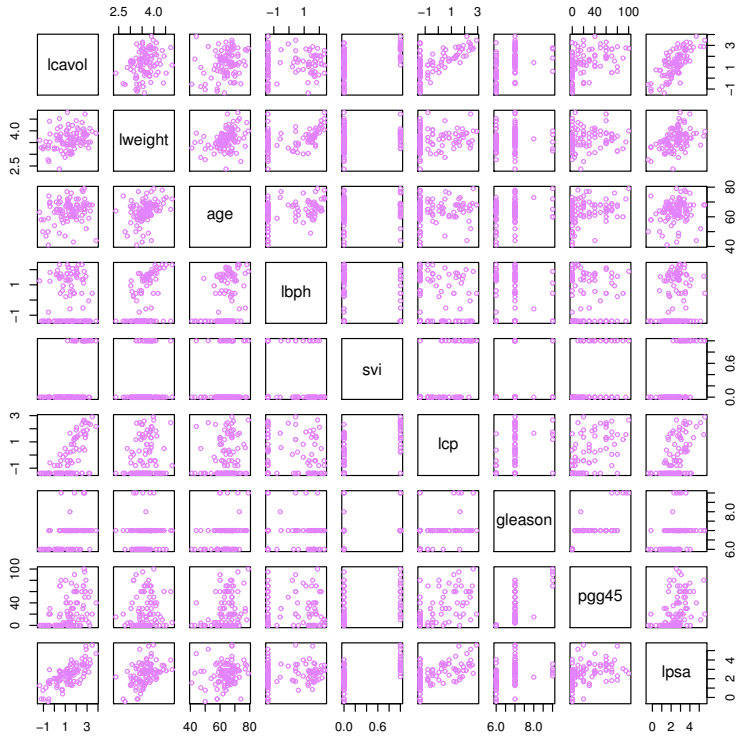


Figura 2.4:

2.2. Penalización L_1: el Lasso

Consideraremos el modelo lineal $\mathbf{Y} \cong \mathbf{N}_n(\mu, \sigma^2\mathbf{I}_n)$, donde $\mu = \mathbf{X}\boldsymbol{\beta}$. Supondremos adicionalmente que las variables están estandarizadas en la forma $\sum_{i=1}^n y_i = 0$, $\sum_{i=1}^n x_{i,j} = 0$ y $\sum_{i=1}^n x_{i,j}^2/n = 1$ para $j = 1, \ldots, p$. El estimador *lasso* $\hat{\boldsymbol{\beta}}^{\text{lasso}}$ se define como la solución del siguiente problema de minimización de mínimos cuadrados penalizados, ahora en la forma (Tibshirani, 1996):

$$\min_{\boldsymbol{\beta}} \| \mathbf{Y} - \mathbf{X}\boldsymbol{\beta} \|_2^2 + \lambda \| \boldsymbol{\beta} \|_1$$

donde ahora $\| (\beta_1, \ldots, \beta_p) \|_1 = \sum_{j=1}^p | \beta_j |$.

Este problema se puede formular alternativamente como:

$$\min_{\boldsymbol{\beta}} \| \mathbf{Y} - \mathbf{X}\boldsymbol{\beta} \|_2^2$$

sujeto a la restricción: $\| \boldsymbol{\beta} \|_1 \leq t$, para algún $t = t(\lambda)$.

Para el análisis de existencia del estimador lasso remitimos al lector al Obsborne *et al*, 2000.

El estimador lasso $\hat{\boldsymbol{\beta}}^{\text{lasso}}$, a diferencia del ridge, no es lineal en $\mathbf{Y}$; esto es, no existe una matriz $\mathbf{H}$,tal que $\hat{\beta}^{\text{lasso}} = \mathbf{HY}$.

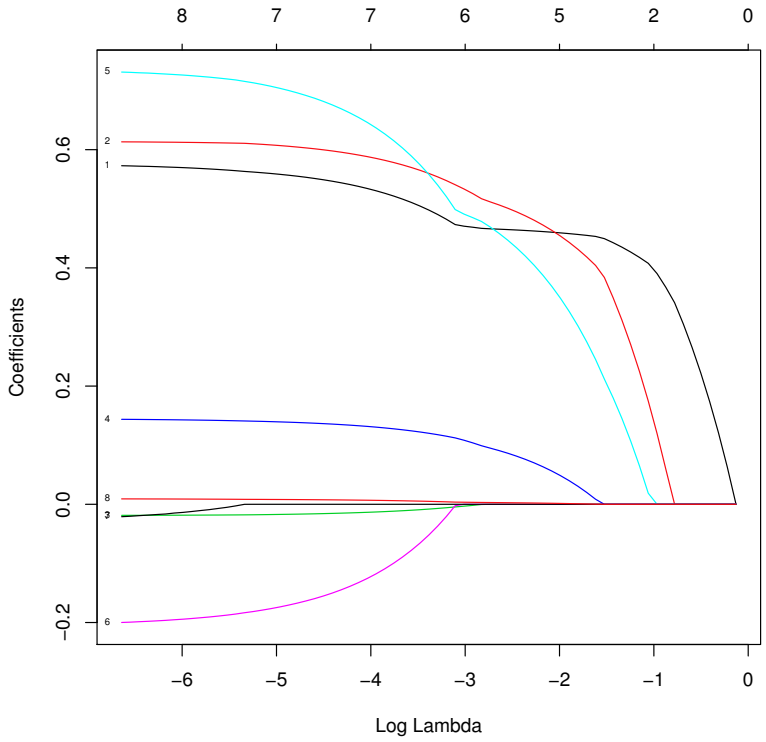


Figura 2.5: Coeficientes obtenidos por el Lasso en función del parámetro de *contracción* λ

De la misma forma que en la regresión ridge, la penalización de los mínimos cuadrados por la norma L_1 produce una *contracción* de los valores absolutos de los coeficientes hacia cero. Sin embargo, mientras que en la regresión ridge los coeficientes se encogen progresivamente hacia cero, el lasso hace que algunos coeficientes se hagan nulos. De esta forma, el lasso puede considerarse una selección continua de variables.

La figura 2.6 tiene como finalidad la interpretación geométrica de los métodos ridge y Lasso

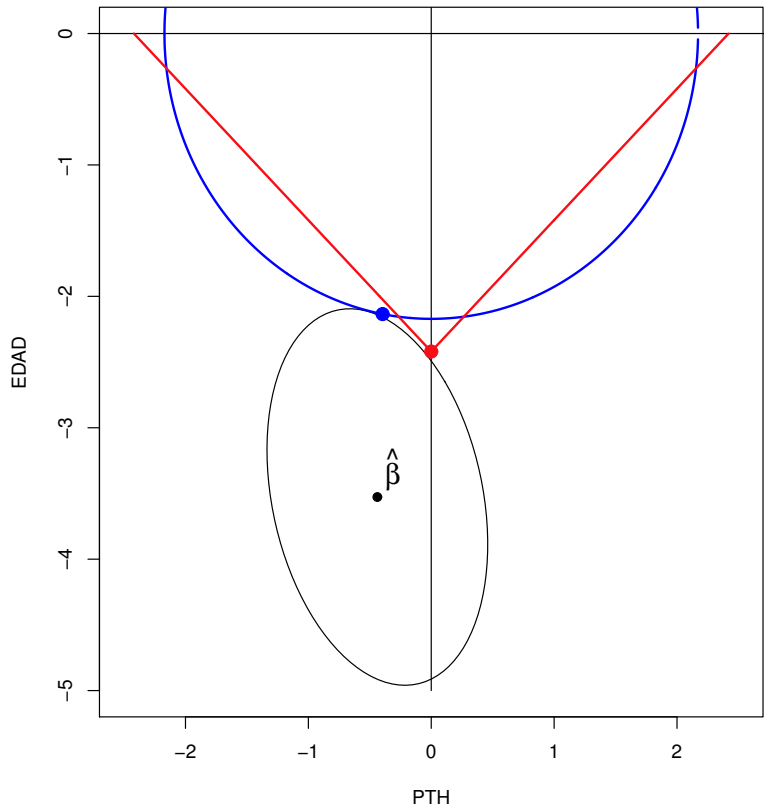


Figura 2.6: Geometrías ridge y Lasso

2.3. Penalización mixta: elastic net

Consideraremos nuevamente el modelo lineal $\mathbf{Y} \cong \mathbf{N}_n(\mu, \sigma^2\mathbf{I}_n)$, siendo $\mu = \mathbf{X}\boldsymbol{\beta}$. Supondremos adicionalmente que las variables están estandarizadas en la forma $\sum_{i=1}^n y_i = 0$, $\sum_{i=1}^n x_{i,j} = 0$ y $\sum_{i=1}^n x_{i,j}^2/n = 1$ para $j = 1, \ldots, p$. En tal escenario, consideramos el siguiente problema de minimización de mínimos cuadrados penalizados de la forma:

$$\min_{\boldsymbol{\beta}} \| \mathbf{Y} - \mathbf{X}\boldsymbol{\beta} \|_2^2 + \lambda_2 \| \boldsymbol{\beta}\|_2^2 + \lambda_1 \| \boldsymbol{\beta}\|_1$$

donde λ_1 y λ_2 son valores fijos no negativos.

Definiendo $\alpha = \lambda_1 / (\lambda_1 + \lambda_2)$, el problema puede expresarse alternativamente en la forma:

$$\min_{\boldsymbol{\beta}} \| \mathbf{Y} - \mathbf{X}\boldsymbol{\beta} \|_1^2$$

sujeto a la restricción: $\alpha \| \beta\|_1 + (1-\alpha) \| \beta\|_2^2 \leq t$, para algún t. El estimador $\hat{\beta^{en}}$ recibe el nombre de elastic-net (Zou y Hastie, 2005) y supone un compromiso entre la estimación ridge ($\alpha = 0$) y la estimación Lasso ($\alpha = 1$).

En la tabla 2.4 se muestran los coeficientes estandarizados de los predictores para el logaritmo del PSA correspondientes a los diversos métodos de regresión empleados.

Variable	OLS	Mejor sub-conjunto	ridge	LASSO	Elastic-net
lcavol	0.716	0.780	0.517	0.573	0.573
lweight	0.293	0.352	0.273	0.225	0.225
age	-0.143		-0.077	0	0
lbph	0.212		0.187	0.093	0.093
svi	0.310		0.260	0.163	0.163
lcp	-0.289		-0.064	0	0
gleason	-0.021		0.031	0	0
pgg45	0.277		0.162	0.058	0.058
Test error	0.5213	0.492	0.4873	0.4523	0.4523

Tabla 2.4: Comparación de los estimadores OLS, mejor subconjunto, ridge, Lasso y elastic net para los datos del estudio del cáncer de próstata. Los coeficientes son los estandarizados

Apéndice. El procedimiento "glmnet" para la regresión regularizada

```
library(glmnet)
data(prostate)
data.train <- prostate[prostate$train, ]
y <- data.train$lpsa
data.train <- as.matrix(data.train[, -c(9, 10)])
data.test <- prostate[!prostate$train, ]
yt <- data.test$lpsa
data.test <- as.matrix(data.test[, -c(9, 10)])
Xs = apply(data.train, 2, scale)
f_ols <- glmnet(data.train, y, lambda = 0)
y_ols = predict(f_ols, newx = data.test)
sum((y_ols - yt)^2)/length(yt)
fols <- glmnet(Xs, y, lambda = 0)
(bols <- as.vector(t(coef(fols))))
library(bestglm)
Xy = prostate[prostate$train, -10]
bm <- bestglm(Xy, IC = "BIC")
xl = data.train[, 1:2]
f_bm <- glmnet(xl, y, lambda = 0)
y_bm = predict(f_bm, newx = data.test[, 1:2])
sum((y_bm - yt)^2)/length(yt)
xls = apply(xl, 2, scale)
f_bm <- glmnet(xls, y, lambda = 0)
(b_bm <- as.vector(t(coef(f_bm))))
cvr <- cv.glmnet(data.train, y, alpha = 0)
mse = NULL
for (i in 1:length(cvr$lambda)) {
    fr <- glmnet(data.train, y, alpha = 0, lambda = cvr$lambda[i])
    y_r <- predict(fr, newx = data.test)
    mse = c(mse, sum((y_r - yt)^2)/length(yt))
}
```

```
plot(cvr$lambda, mse, type = "l")
j = which.min(mse)
fr <- glmnet(Xs, y, alpha = 0, lambda = cvr$lambda[j])
(bridge <- as.vector(t(coef(fr))))
cvl <- cv.glmnet(data.train, y, alpha = 1)
plot(cvl)
mse = NULL
for (i in 1:length(cvl$lambda)) {
    fl <- glmnet(data.train, y, alpha = 1, lambda = cvl$lambda[i])
    y_l <- predict(fl, newx = data.test)
    mse = c(mse, sum((y_l - yt)^2)/length(yt))
}
plot(cvl$lambda, mse, type = "l")
j = which.min(mse)
fl <- glmnet(Xs, y, alpha = 1, lambda = cvl$lambda[j])
(blasso <- as.vector(t(coef(fl))))
cve <- cv.glmnet(data.train, y)
plot(cve)
mse = NULL
for (i in 1:length(cve$lambda)) {
    fe <- glmnet(data.train, y, lambda = cve$lambda[i])
    y_en <- predict(fe, newx = data.test)
    mse = c(mse, sum((y_en - yt)^2)/length(yt))
}
plot(cve$lambda, mse, type = "l")
j = which.min(mse)
fen <- glmnet(Xs, y, lambda = cve$lambda[j])
(belastic <- as.vector(t(coef(fen))))
```

Capítulo 3

Regresión con variables latentes

3.1. Introducción

A lo largo de este capítulo consideraremos nuevamente el modelo de regresión lineal general, el cual expresamos en la forma:

$$E\left[\mathbf{Y} \mid X_{1,}, \ldots, X_p\right] = \beta_0 + \sum_{j=1}^{p} \beta_j X_j$$

donde $X_j' = (x_{1,j}, \ldots, x_{n,j})$, para $j = 1, \ldots, p$. El teorema de Gauss-Markov establece que para el parámetro $a'\beta$ ($a \in \mathbb{R}^{p+1}$), *el estimador* $a'\hat{\beta}$ *es el de varianza mínima entre todos los estimadores lineales centrados* (ver sección 1.4). Sin embargo, en aquellos casos en los que la matriz de diseño no es de rango máximo o las variables $X_1, \ldots, X_p$ están fuertemente correladas (problema mal condicionado), el estimador de mínimos cuadrados $\hat{\beta}$ (si existe), aunque es centrado, puede tener una varianza inaceptable. Por tal motivo, habrá que admitir un cierto sesgo en cualquier estimador en orden a que su varianza sea moderada. En el capítulo anterior se utilizaron métodos de estimación basados en *contracciones* del estimador de mínimos cuadrados. En el presente, consideramos dos métodos basados en la construcción de p variables incorreladas $Z_1, \ldots, Z_p$ a partir de las variables originales $X_1, \ldots, X_p$. Las variables Z_i reciben el nombre de *variables latentes*. En las siguientes secciones describimos el método de regresión basado en las *componentes principales* (PC) y el método de los *mínimos cuadrados parciales* (PLS).

Sin pérdida de generalidad asumimos que las variables X_j están estandarizadas

en la forma $\sum_{i=1}^{n} x_{i,j} = 0$ y $\sum_{i=1}^{n} x_{i,j}^2/n = 1$ para $j = 1, \ldots, p$. De esta forma, la matriz de covarianzas de $\mathbf{X} = (X_1, \ldots, X_p)$ es:

$$\mathbf{C} = \frac{1}{n}\mathbf{X}'\mathbf{X}$$

Nótese que la diagonal principal de $\mathbf{C}$ son las varianzas de las variables $X_1, \ldots, X_n$, y de esta forma, su traza de la es p. Recordamos también que $\mathbf{C}$ es definida no negativa, lo que supone que sus autovalores son no negativos.

3.2. Componentes principales

3.2.1. Aproximación al concepto de componente principal

Considérense las variables X_1 y X_2 de la figura 3.1 entre las cuales existe claramente una fuerte correlación lineal ($\rho_{1,2} \approx 0{,}9$). Esto supone que ambas variables, en algún sentido, pueden resumirse a una única variable esencial (componente principal).

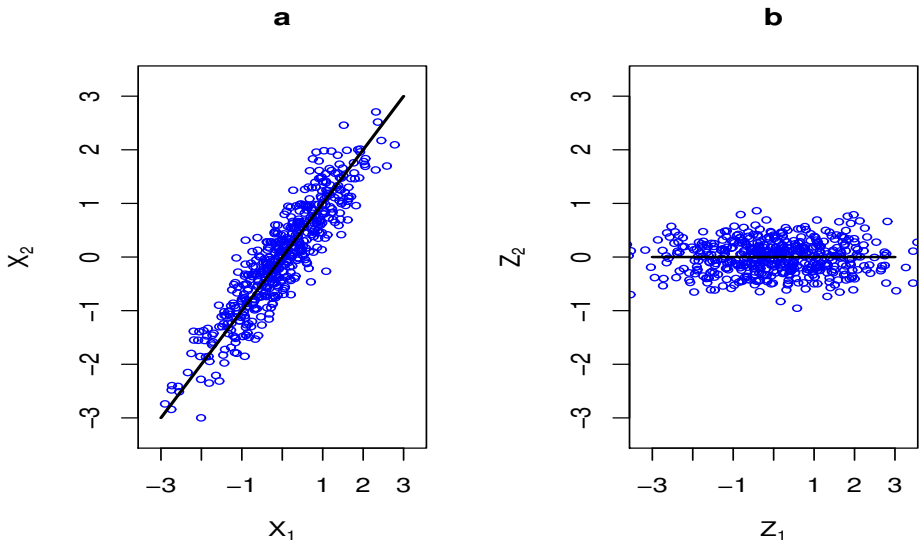


Figura 3.1:

Para obtenerla, imaginemos que la nube de puntos (a) gira en el sentido del eje que da la *máxima variabilidad* de los datos hasta alcanzar la posición (b). Definimos entonces como variable Z_1 la que corresponde al nuevo eje de abscisas

y Z_2, al nuevo eje de ordenadas. Obsérvese que la variabilidad de Z_2 es tan escasa que podría dejar de considerarse propiamente como variable. De esta forma puede admiterse que la variabilidad del vector original de variables X_1 y X_2 estaría determinada por la única variable esencial Z_1.

3.2.2. Construcción de las componentes principales

En general, para un vector de variables $\mathbf{X} = (X_1, \ldots, X_p)$, el vector de ***componentes principales*** $\mathbf{Z} = (Z_1, \ldots, Z_p)$ es el resultado de una transformación ortogonal $\mathbf{Z} = \mathbf{XW}$, definida de acuerdo con el siguiente algoritmo:

1. La primera *componente principal* se define por:

$$Z_1 = w_{1,1}X_1 + \cdots + w_{1,p}X_p = \mathbf{X}\mathbf{w}_1$$

donde $\mathbf{w}_1' = (w_{1,1}, \ldots, w_{1,p})$ es tal que $\mathbf{w}_1'\mathbf{w}_1 = 1$ y además:

$$\text{var}\,(Z_1) = \frac{1}{n}\mathbf{w}_1'\mathbf{X}'\mathbf{X}\mathbf{w_1} = \mathbf{w_1'}\mathbf{C}\mathbf{w_1}$$

es ***máxima***. El objetivo de varianza máxima para Z_1 se corresponde con la idea de girar los ejes coordenados hasta que el primero coincida con la *máxima variabilidad* de la nube de puntos.

Para resolver este problema, consideramos la función lagrangiana:

$$\mathcal{L}_1\,(\mathbf{w}) = \mathbf{w}'\mathbf{C}\mathbf{w} - \lambda\,(\mathbf{w}'\mathbf{w} - 1)$$

Derivando respecto de $\mathbf{w}$ e igualando a cero se obtiene:

$$\frac{\partial \mathcal{L}_1}{\partial \mathbf{w}} = 2\mathbf{C}\mathbf{w} - 2\lambda\mathbf{w} = 0$$

De lo anterior se deduce que:

$$\mathbf{C}\mathbf{w} = \lambda\mathbf{w}$$

De esta forma, la solución $\mathbf{w}_1$ es un autovector de la matriz de covarianzas

$\mathbf{C}$, mientras que λ_1 es el autovalor correspondiente. Nótese que

$$\text{var}\,(Z_1) = \mathbf{w}_1'\mathbf{C}\mathbf{w_1} = \lambda_1$$

2. La segunda componente principal tiene también la forma: $Z_2 = \mathbf{X}\mathbf{w}_2$ donde $\mathbf{w}_2$ es tal que, $\mathbf{w}_2'\mathbf{w}_2 = 1$, $\text{var}\,(Z_2)$ es máxima y Z_2 es incorrelada con Z_1. Nótese que $\text{cov}\,(Z_1, Z_2) = \mathbf{w}_2'\mathbf{C}\mathbf{w}_1 = \lambda_1\mathbf{w}_2'\mathbf{w_1}$. Por tanto, la incorrelación de Z_1 y Z_2 es equivalente a $\mathbf{w}_2'\mathbf{w_1} = 0$. La función langrangiana tiene entonces la forma:

$$\mathcal{L}_2\,(\mathbf{w}) = \mathbf{w}'\mathbf{C}\mathbf{w} - \lambda\,(\mathbf{w}'\mathbf{w} - 1) - \mu\mathbf{w}'\mathbf{w}_1$$

Derivando nuevamente respecto de $\mathbf{w}$ e igualando a cero se obtiene:

$$\frac{\partial\mathcal{L}_2}{\partial\mathbf{w}} = 2\mathbf{C}\mathbf{w} - 2\lambda\mathbf{w} - \mu\mathbf{w_1} = 0$$

Veamos ahora que $\mu = 0$. Efectivamente, multiplicando la ecuación anterior por la derecha por $\mathbf{w_1'}$ queda:

$$\mathbf{2}\mathbf{w_1'}\mathbf{C}\mathbf{w} - 2\lambda\mathbf{w_1'}\mathbf{w} - \mu\mathbf{w_1'}\mathbf{w_1} = 0$$

El resultado se sigue del hecho de ser $\mathbf{w_1'}\mathbf{C}\mathbf{w} = \mathbf{0}$, $\mathbf{w_1'}\mathbf{w} = 0$ y $\mathbf{w_1'}\mathbf{w_1} = 1$. Ello supone que:

$$\mathbf{C}\mathbf{w_2} = \lambda_2\mathbf{w_2}$$

Esto es, $\mathbf{w_2}$ es otro autovector de $\mathbf{C}$, orotgonal a $\mathbf{w_1}$. Nótese que:

$$\text{var}\,(Z_2) = \mathbf{w_2'}\mathbf{C}\mathbf{w_2} = \lambda_2$$

Dado que la varianza de Z_1 era máxima, $\lambda_1 > \lambda_2$.

3. El vector de coeficientes $\mathbf{w}_j$ correspondiente a la j-ésima componente principal $Z_j = \mathbf{X}\mathbf{w}_j$ $(j \leq p)$ se obtiene de tal forma que $\mathbf{w}_j'\mathbf{w}_j = 1$, $\text{var}\,(Z_j)$ sea máxima y $\text{cov}\,(Z_i, Z_j) = 0 \;:\; i < j$. De esta forma es posible construir hasta p componentes principales.

El procedimiento anterior puede resumirse finalmente en la transformación ortogonal:

$$\mathbf{Z} = \mathbf{XW}$$

donde las columnas de $\mathbf{W}$ son los autovectores de la matriz de covarianzas $\mathbf{C}$, siendo los correspondientes autovalores las varianzas de las componentes principales; esto es: $\lambda_j = \text{var}\,(Z_j)$. Nótese que por construcción, $\lambda_1 > \lambda_2 > \cdots > \lambda_p$.

Nótese que por ser nulas las medias de las variables $X_1, \ldots, X_p$, también lo son las de las componentes principales $Z_1, \ldots, Z_p$. De esta forma, la matriz de covarianzas de $\mathbf{Z}$ es:

$$\frac{1}{n}\mathbf{Z}'\mathbf{Z} = \frac{1}{n}\mathbf{W}'\mathbf{X}'\mathbf{X}\mathbf{W} = \mathbf{W}'\mathbf{C}\mathbf{W}$$

Ello supone que las matrices de covarianzas de $\mathbf{X}$ y $\mathbf{Z}$ son semejantes, lo que implica que tienen la misma traza. De esta forma, la *varianza total* de ambos vectores es la misma. Más concretamente se tiene:

$$\sum_{j=1}^{p} \text{var}\,(Z_j) = \sum_{j=1}^{p} \lambda_j = \sum_{j=1}^{p} \text{var}\,(X_j) = p$$

por ser $\text{var}\,(X_j) = 1 \;\; : \;\; j = 1, \ldots, p$.

En la siguiente salida de R se resume el análisis de componentes principales correspondiente al vector de variables predictoras del antígeno prostático específico (ejemplo 2.4). Las variables fueron estandarizadas a media cero y varianza uno. Puede comprobarse que la suma de las varianzas de las ocho componentes principales es ocho.

```
mp <- prcomp(X_1, scale = TRUE)
summary(mp)

## Importance of components:
##                           PC1   PC2   PC3    PC4    PC5    PC6   PC7    PC8
## Standard deviation     1.851 1.278 1.018 0.7863 0.6747 0.6138 0.529 0.4173
## Proportion of Variance 0.428 0.204 0.130 0.0773 0.0569 0.0471 0.035 0.0218
## Cumulative Proportion  0.428 0.632 0.762 0.8392 0.8962 0.9432 0.978 1.0000
```

3.2.3. Componentes principales como variables latentes

Sea nuevamente el vector $\mathbf{X} = (X_1, \ldots, X_p)$ y su vector de componentes principales $\mathbf{Z} = \mathbf{XW}$. Dado que $\mathbf{W}$ es una matriz ortogonal, puede escribirse $\mathbf{X} = \mathbf{ZW}'$, o lo que es igual:

$$X_j \approx w_{1,j}Z_1 + \cdots + w_{p,j}Z_p \ : \ j = 1, \ldots, p$$

Esta expresión sugiere que las variables *observables* $\mathbf{X}$ son dependientes de las variables *latentes* (no observables) $\mathbf{Z}$. Nótese que si el vector $\mathbf{X}$ está mal condicionado (fuerte correlación entre sus variables), un número menor de componentes k ($k \ll p$) podría explicar razonablemente bien la variabilidad del vector $\mathbf{X}$. Ello conduciría por tanto a la siguiente representación:

$$X_j = w_{1,j}Z_1 + \cdots + w_{k,j}Z_k + U_j \ : \ j = 1, \ldots, p$$

donde las varibles U_j reciben el nombre de *especificidades*. Esta representación es equivalente a:

$$X_j = w_{1,j} \cdot \sqrt{\lambda_1}\frac{Z_1}{\sqrt{\lambda_1}} + \cdots + w_{k,j}\sqrt{\lambda_k}\frac{Z_k}{+\sqrt{\lambda_k}} + U_j \ : \ j = 1, \ldots, p$$

Nótese que la variable latente $F_m = Z_m/\sqrt{\lambda_m}$ está estandarizada ($\text{var}\,(F_m) = 1$), y de ahí, $\alpha_{m,j} = w_{m,j}\sqrt{\lambda_m}$ mide su efecto sobre la variable X_j. Esta representación puede expresarse en la forma del modelo factorial clásico:

$$X_j = \alpha_{1,j}F_1 + \cdots + \alpha_{k,j}F_k + U_j$$

Dado que $\text{var}\,(X_j) = 1$ y los factores F_m son incorrelados, se tiene finalmente para $j = 1, \ldots, p$:

$$1 = \alpha_{1,j}^2 + \cdots + \alpha_{k,j}^2 + \text{var}\,(U_j)$$

La forma matricial del modelo factorial es:

$$\mathbf{X} = \mathbf{FA} + \mathbf{U}$$

donde las columnas de la matriz $\mathbf{F} = (F_1, \ldots, F_k)$, $\mathbf{A}$ la matriz de coeficientes, y finalmente, $\mathbf{U} = (U_1, \ldots, U_p)$. Obviamente, cuando $\text{var}\,(U_j) \approx 0$, la variable X_j

se explica bien por los k factores comunes.

Consideramos ahora nuevamente los datos correspondientes al estudio sobre el cáncer de próstata. En el siguiente cuadro se muestra la representación de los ocho predictores del log (psa) en función de dos factores deducidos de las primeras componentes principales.

```
##          Dim.1    Dim.2
## lcavol  0.8071  0.04124
## lweight 0.3102  0.72149
## age     0.4460  0.53644
## lbph    0.0562  0.83081
## svi     0.7325 -0.22421
## lcp     0.8515 -0.21723
## gleason 0.7303 -0.07043
## pgg45   0.8258 -0.17241
```

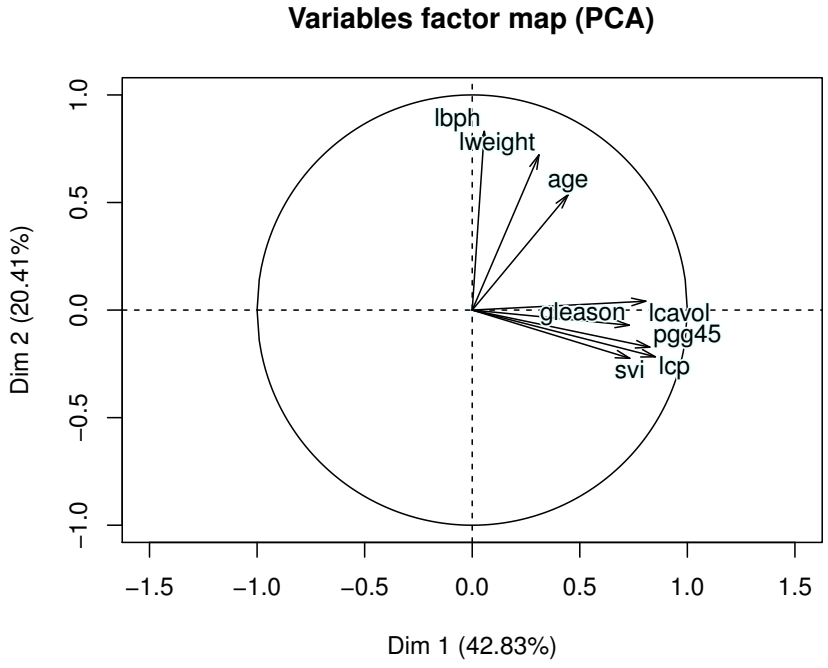


Figura 3.2: Representación de los predictores del PSA según las dos primeras componentes principales

Nótese que las posiciones de las variables están cerca de la circunferencia uniad, lo que significa que los dos factores explican razonablemente bien la variabilidad de los datos

La proporción de *variabilidad explicada* por las k primeras componentes principales se define por:

$$\frac{1}{p}\sum_{j=1}^{k}\lambda_j$$

3.2.4. Regresión en componentes principales

Considérese ahora que se desea estimar el modelo lineal descrito en la sección 3.1. . El método de regresión en componentes principales puede resumirse en el siguiente algoritmo:

1. A partir del vector $(X_1, \ldots, X_p)$, obtener el vector de componentes principales $(Z_1, \ldots, Z_p)$.

2. Centralizar el vector $\mathbf{Y}' = (y_1, \ldots, y_n)$ en la forma $\tilde{\mathbf{Y}} = (y_1 - \bar{y}, \ldots, y_n - \bar{y})$, con $\bar{y} = (1/n)\sum_{i=1}^{n} y_i$.

3. Para un valor $K \leq p$, obtener la regresión de $\widetilde{\mathbf{Y}}$ sobre $Z_1, \ldots, Z_K$. Dado que las variables Z_j son ortogonales, esta regresión es de la forma:

$$\hat{\mathbf{Y}}_{(K)}^{\text{pcr}} = \bar{y}\mathbf{1} + \sum_{m=1}^{K}\theta_m Z_m$$

   donde $\theta_j =< Z_j, \mathbf{Y} > / < Z_j, Z_j >$.

4. Finalmente, expresando las componentes principales en función de las variables originales X_j ($Z_m = \sum_{j=1}^{p} w_{m,j}X_j$), se obtiene el estimador basado en las componentes principales en la forma:

$$\hat{\mathbf{Y}}_{(K)}^{\text{pcr}} = \bar{y}\mathbf{1} + \sum_{j=1}^{p}\hat{\beta}_j{}^{\text{pcr}}X_j$$

El estimador $\hat{\boldsymbol{\beta}}^{\text{pcr}} = \left(\hat{\beta}_1{}^{\text{pcr}}, \ldots, \hat{\beta}_p{}^{\text{pcr}}, \right)$ es el de la regresión en componentes principales.

```
library(pls)
K = 8
mpc <- pcr(lpsa_l ~ X_l, scale = TRUE, K, validation = "CV")
summary(mpc)

## Data:  X dimension: 67 8
##  Y dimension: 67 1
## Fit method: svdpc
## Number of components considered: 8
##
## VALIDATION: RMSEP
## Cross-validated using 10 random segments.
##        (Intercept)  1 comps  2 comps  3 comps  4 comps  5 comps  6 comps
## CV           1.217   0.9232   0.8811   0.8063   0.7961   0.8055   0.8255
## adjCV        1.217   0.9205   0.8796   0.8033   0.7931   0.8026   0.8221
##        7 comps  8 comps
## CV      0.7981   0.7493
## adjCV   0.7936   0.7446
##
## TRAINING: % variance explained
##         1 comps  2 comps  3 comps  4 comps  5 comps  6 comps  7 comps
## X         42.83    63.24    76.20    83.92    89.61    94.32    97.82
## lpsa_l    45.18    50.84    59.58    61.00    61.17    62.08    66.36
##         8 comps
## X        100.00
## lpsa_l    69.44
```

3.3. Mínimos cuadrados parciales (PLS)

Este método se basa también en obtener una representación factorial para la matriz de diseño $\mathbf{X}$, de la forma:

$$\mathbf{X} = \mathbf{FA} + \mathbf{U}$$

donde $\mathbf{F} = (Z_1, \ldots, Z_k)$ (normalmente $k \ll p$), siendo $Z_1, \ldots, Z_k$ variables incorreladas y $\mathbf{U} = (U_1, \ldots, U_p)$. El objetivo ahora es construir los factores de tal forma que *sean relevantes para la predicción de* $\mathbf{Y}$. Ello supone que en su determinación se tiene en cuenta la respuesta $\mathbf{Y}$. De la misma forma que con las componentes principales, obtendremos inicialmente p variables ortogonales $\mathbf{Z} = (Z_1, \ldots, Z_p)$, para proceder posteriormente a la selección de un subconjunto reducido de los mismos. Nótese que para $k = p$, $\mathbf{F} = \mathbf{Z}$.

Describimos ahora el método de los mínimos cuadrados parciales

1. La primera componente (dirección) *pls,* $Z_1 = \mathbf{X}\mathbf{w_1}$, $(\mathbf{w}_1' = (w_{1,1}, \ldots, w_{1,p})$) se obtiene como solución del problema:

$$\max_{\mathbf{w}} \operatorname{cov}(Z_1, \mathbf{Y})^2$$

sujeto a $\mathbf{w}'\mathbf{w} = \mathbf{1}$. Para ello consideramos la función lagrangiana:

$$\mathcal{L}_1(\mathbf{w}) = \mathbf{w}'\mathbf{X}'\mathbf{Y}\mathbf{Y}'\mathbf{X}\mathbf{w} - \lambda(\mathbf{w}'\mathbf{w} - 1)$$

Derivando respecto de $\mathbf{w}$ e igualando a cero se obtiene:

$$\frac{\partial \mathcal{L}_1}{\partial \mathbf{w}} = 2\mathbf{X}'\mathbf{Y}\mathbf{Y}'\mathbf{X}\mathbf{w} - 2\lambda\mathbf{w} = 0$$

De lo anterior se deduce que:

$$\mathbf{X}'\mathbf{Y}\mathbf{Y}'\mathbf{X}\mathbf{w} = \lambda\mathbf{w}$$

Ello significa que la solución óptima es un autovector de la matriz $\mathbf{X}'\mathbf{Y}\mathbf{Y}'\mathbf{X}$. Es por otra parte, inmediato comprobar que $\lambda = \mathbf{w}'\mathbf{X}'\mathbf{Y}\mathbf{Y}'\mathbf{X}\mathbf{w}$. Simples cálculos algebráicos permiten concluir que:

$$\mathbf{w_1} = \frac{\mathbf{X}'\mathbf{Y}}{\| \mathbf{X}'\mathbf{Y} \|}$$

2. Para construir la segunda componente pls se ortogonalizan las variables X_j respecto de la primera dirección $Z_1 = \mathbf{X}\mathbf{w_1}$. Representaremos el vector correspopndiente por $\mathbf{X}^{(1)}$. La segunda componente tiene la forma; $Z_2 = \mathbf{X^{(1)}}\mathbf{w_2}$. Obviamente, Z_1 y Z_2 son ortogonales. El vector de pesos $\mathbf{w_2}$ se

obtiene como solución del problema:

$$\max_{\mathbf{w}} \operatorname{cov}(Z_2, \mathbf{Y})^2$$

sujeto a $\mathbf{w}'\mathbf{w} = \mathbf{1}$.

3. El procedimiento se itera hasta obtener p componentes *pls*. La matriz $\mathbf{Z}$, cuyas columnas son variables *latentes* $Z_1, \ldots, Z_p$ recibe el nombre de *matriz score*.

Resumimos ahora el algoritmo para la determinación de los mínimos cuadrados parciales, Hastie *et al*, 2008.

1. Tipificar las variables $X_j = (x_{1,j}, \ldots, x_{n,j})$. Inicializar $\hat{Y}^{(0)} = (\bar{y}, \ldots, \bar{y})$ y $X_j^{(0)} = X_j$ para $j = 1, \ldots, p$.

2. Para $m = 1, \ldots, p$, obtener:

   a) $Z_m = \sum_{j=1}^{p} w_{m,j} X_j^{(m-1)}$, donde $w_{m,j} =< X_j^{(m-1)}, Y >$.

   b) $\hat{Y}^{(m)} = \hat{Y}^{(m-1)} + \theta_m Z_M$, siendo $\theta_m =< Z_m, Y > / < Z_m, Z_m >$.

   c) Ortogonalizar cada $X_j^{(m)}$ respecto de Z_m ; esto es: $X_j^{(m)} = X_j^{(m-1)} - \left[< Z_m, X_j^{(m-1)} > / < Z_m, Z_m > \right] Z_m$, $j = 1, \ldots, p$.

3. Nótese finalmente que $\hat{Y}^{(m)} = \hat{Y}^{(0)} + \sum_{k=1}^{m} \theta_k Z_k$ (combinación lineal de las variables ortogonales Z_k), y de esta forma, $\hat{Y}^{(m)} = \hat{Y}^{(0)} + \sum_{j=1}^{p} \hat{\beta}_j^{\text{pls}} X_j$. El estimador $\hat{\boldsymbol{\beta}}^{\text{pls}} = \left(\hat{\beta}_1^{\text{pls}}, \ldots, \hat{\beta}_p^{\text{pls}} \right)$ es el estimador de mínimos cuadrados parciales (*pls*).

De forma análoga a las componentes principales, la variabilidad total del vector de observaciones $\mathbf{X}$ puede descomponerse en una serie de términos relacionados con los cuadrados de las covarianzas de la componente con las variables X_j. Más concretamente, sea $\tilde{\mathbf{Z}}$ es la matriz cuyas columnas son las componentes *pls* normalizadas ($\tilde{Z}_m = Z_m / \parallel Z_m \parallel$). Ello significa que las matrices $\mathbf{X}'\mathbf{X}$ y $\tilde{\mathbf{Z}}'\mathbf{X}\mathbf{X}'\tilde{\mathbf{Z}}$ son semejantes, lo que implica:

$$\operatorname{traza}(\mathbf{X}'\mathbf{X}) = \operatorname{traza}\left(\tilde{\mathbf{Z}}'\mathbf{X}\mathbf{X}'\tilde{\mathbf{Z}} \right)$$

Simples cálculos algebráicos llevan a la relación:

$$p = \sum_{j=1}^{p} \text{var}\left(X_j\right) = \sum_{m=1}^{p} \left\{ \sum_{j=1}^{p} \text{cor}\left(Z_m, X_j\right)^2 \right\}$$

En esta expresión, $\sum_{j=1}^{p} \text{cor}\left(Z_m, X_j\right)^2$ representa la contribución de la m-ésima componente pls a la variabilidad total del predictor $\mathbf{X}$. De esta forma, el porcentaje de contribución de Z_m a la variabilidad total de $\mathbf{X}$ de define como:

$$\frac{100}{p} \sum_{j=1}^{p} \text{cor}\left(Z_m, X_j\right)^2$$

Por analogía a la expresión anterior, obtenemos ahora los porcentajes de la variabilidad de $\mathbf{Y}$ explicada por las sucesivas componentes pls en la forma :

$$100 \times \frac{n \cdot \text{cov}\left(\tilde{Z}_m, \mathbf{Y}\right)^2}{\text{var}\left(\mathbf{Y}\right)} = 100 \times \text{cor}\left(Z_m, \mathbf{Y}\right)^2 \quad : \; m = 1, \ldots, p$$

En la siguiente salida de R se muestra la sintaxis del procedimiento plsr para la estimación del modelo de regresión para el $\log\left(psa\right)$ correspondiente a los datos de cáncer de próstata utilizando el procedimiento pls. En los argumentos del procedimiento se indica el número de componentes a calcular (8) y el método de estimación del error estándar de la predicción (validación cruzada). Nótese que éste es prácticamente constante a partir de tres componentes.

```
mpls <- plsr(lpsa_l ~ X_l, 8, scale = TRUE, validation = "CV")
summary(mpls)

## Data:  X dimension: 67 8
##  Y dimension: 67 1
## Fit method: kernelpls
## Number of components considered: 8
##
## VALIDATION: RMSEP
## Cross-validated using 10 random segments.
```

```
##        (Intercept)  1 comps  2 comps  3 comps  4 comps  5 comps  6 comps
## CV           1.217   0.8557   0.8306   0.8213   0.8077   0.8010   0.7952
## adjCV        1.217   0.8530   0.8251   0.8143   0.7998   0.7936   0.7883
##        7 comps  8 comps
## CV      0.7968   0.7969
## adjCV   0.7899   0.7899
##
## TRAINING: % variance explained
##         1 comps  2 comps  3 comps  4 comps  5 comps  6 comps  7 comps
## X         41.64    58.29    71.13    79.75    86.08    90.21    94.70
## lpsa_l    55.79    64.60    67.51    69.12    69.37    69.43    69.44
##         8 comps
## X        100.00
## lpsa_l    69.44
```

Tres componentes explican el 71.13 % de la variabilidad de $\mathbf{X}$ y el 67.51 % de la variabilidad de $\mathbf{Y}$ (las dos primeras componentes explicaban el 64.6 %, por lo que este número podría se preferible). En la figura 3.3 se muestran las predicciones del $\log(psa)$ para 8 y 2 componentes, frente a los valores observados.

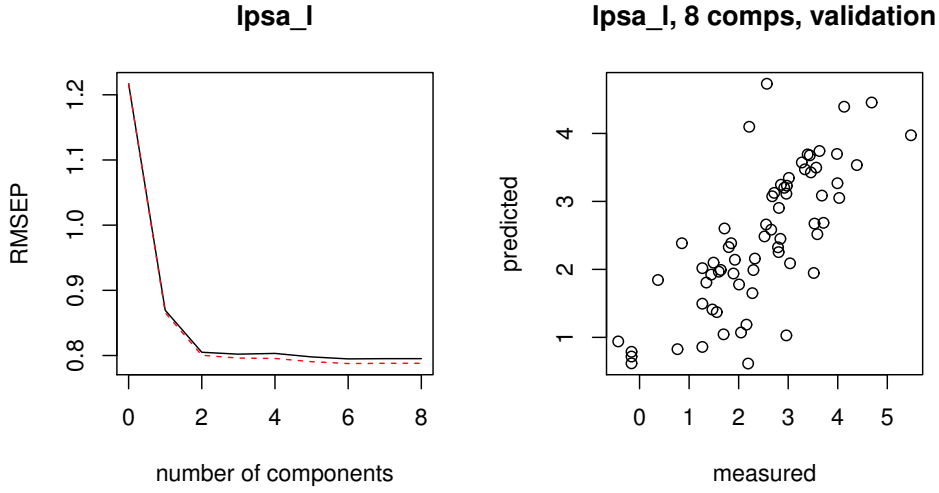


Figura 3.3: Para la estimación pls basada en 8 componentes, el error estándar de predicción es 0.7854, mientras que para dos componentes es de 0.8092

En la figura 3.4 se muestran los predictores y la respuesta ($\log(psa)$) representados frente a las componentes pls (Z_1 y Z_2). Nótese que aquellas variables que están fuertemente correladas tienen representaciones próximas.

```
##
## Loadings:
##          Comp 1 Comp 2
## lcavol    0.475  0.128
## lweight   0.265  0.617
## age       0.256 -0.130
## lbph      0.111  0.527
## svi       0.411
## lcp       0.459 -0.222
## gleason   0.358 -0.502
## pgg45     0.416 -0.458
##
##                Comp 1 Comp 2
## SS loadings     1.055  1.204
## Proportion Var  0.132  0.151
## Cumulative Var  0.132  0.282
```

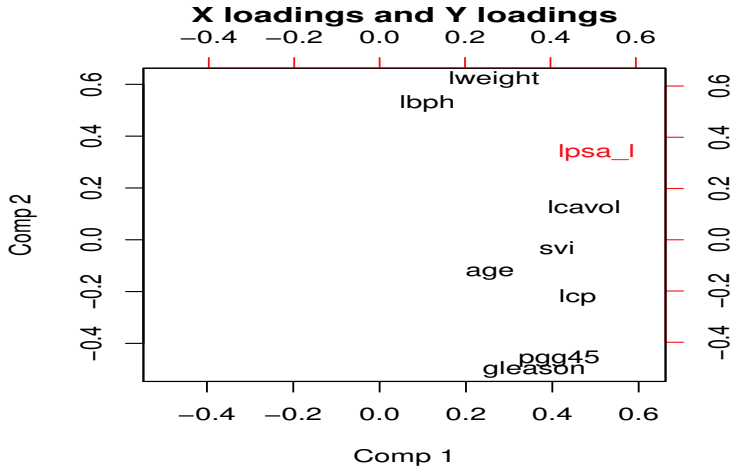


Figura 3.4: Variables predictivas y variable de respuesta en función de las componentes pls

En la tabla 3.1 se muestra finalmente la comparación de los diversos métodos de estimación del modelo de regresión lineal para los datos del cáncer de próstata. Los métodos ridge y laso se han resumido en el elasticnet. Nótese que el test de error es similar entre éste y el PLS. Éste último tiene la ventaja de explicar la regresión en un conjunto más reducido de variables latentes.

Variable	OLS	Mejor sub-conjunto	Elastic-net	PCR	PLS
lcavol	0.716	0.780	0.573	0.571	0.598
lweight	0.293	0.352	0.225	0.323	0.313
age	-0.143		0	-0.154	-0.184
lbph	0.212		0.093	0.216	0.205
svi	0.310		0.163	0.322	0.312
lcp	-0.289		0	-0.050	-0.037
gleason	-0.021		0	0.229	0.006
pgg45	0.277		0.058	-0.064	0.123
Test error	0.521	0.492	0.452	0.474	0.456

Tabla 3.1: Comparación de los estimadores OLS, mejor subconjunto, ridge, Lasso y elastic net para los datos del estudio del cáncer de próstata. Los coeficientes son los estandarizados

Ilustración para los datos de cáncer de próstata

En la siguiente imagen puede apreciarse un cáncer de próstata. Nótese la invasión del tumor en la vesícula seminal.

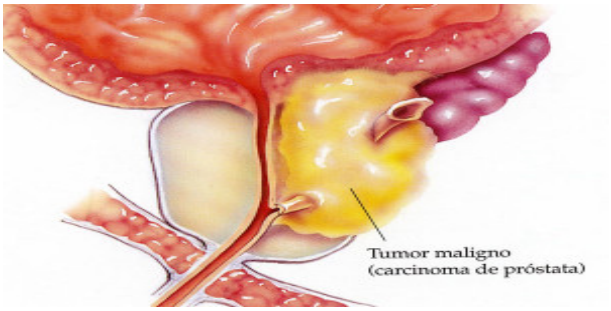


Figura 3.5:

Bibliografía

[1] Arnold, S. (1981). *The Theory of Linear Models and Multivariate Analysis*. Wiley.

[2] Delaney, N.J. and Chatterjee, S. (1986). Use of the Bootstrap and Cross-Validation in Ridge Regression. *Journal of Business & Economic Statistics*, 4:255-262.

[3] Frank, I.E. and Friedman, J.H. (1993). A statistical view of some chemometrics regression tools. *Technometrics*, 35: 109-135.

[4] Efron, B. (1979). Bootstrap methods: another look at the jackknife. *The Annals of Statistics*, 7:1-26.

[5] Freedman, D.A. (1981). Bootstraping regression models. *The Annals of Statistics*, 9:1218-1228.

[6] Friedman, J., Hastie, T. and Tibshirani, R. (2008) Regularization Paths for Generalized Linear Models via Coordinate Descent, http://www.stanford.edu/~hastie/Papers/glmnet.pdf. *Journal of Statistical Software*, Vol. 33(1), 1-22 Feb 2010.

[7] Hastie, T., Tibshirani, R. and Friedman, J. (2008). *The elements of Statistical Learning: Data Mining. Inference and Prediction.* Springer-Verlag.

[8] Hoerl AE. (1962). Application of ridge analysis to regression problems. *Chemical Engineering Progress*. 1958, 54–59.

[9] Hoerl, A.E. and Kennard. R.W. (1970). Ridge regression:biased estimation for nonorthogonal problems. *Technometrics*. 12:53-63.

[10] Obsborne, M.R., Presnell, B. and Turlach, B.A. (2000). On the LASSO and its Dual. *Journal of Computational and Graphical Statistics*, 9:319-337.

[11] Stamey, T., Kabalin, J., McNeal, J., Johnstone, I., Freiha, F., Redwine, E. and Yang, N (1989). Prostate specific antigen in the diagnosis and treatment of adenocarcinoma of the prostate II. Radical prostatectomy treted patients. *Journall of Urology* 16: 1076–1083.

[12] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B.* 58:267–288.

[13] Wold, S., Sjöström and Eriksson, L. (2001). PLS-regression: a basic tool of chemomerics. *Chemometrics and intelligent laboratory systems*, 58:109-130.

[14] Zou, H. and T. Hastie (2005). Regularization and variable selection via the elastic net. *J. Roy. Statist. Soc. Ser. B.* 67: 301-320.