8

Regresión lineal múltiple

8.1. El modelo

De forma análoga al caso de la regresión lineal simple, podemos considerar el modelo lineal entre una variable aleatoria respuesta Y y un grupo de k variables no aleatorias $x_1, \ldots, x_k$ explicativas o regresoras.

Si $y_1, \ldots, y_n$ son n observaciones independientes de Y, el modelo lineal de la regresión múltiple se define como

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + \epsilon_i \qquad i = 1, \ldots, n \tag{8.1}$$

donde $(x_{i1}, \ldots, x_{ik})$ son los valores observados correspondientes a y_i y se asumen las consabidas hipótesis de Gauss-Markov sobre los errores.

En notación matricial, el modelo se escribe

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

donde $\mathbf{Y} = (y_1, \ldots, y_n)'$, $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_k)'$, $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)'$ y la matriz de diseño es

$$\mathbf{X} = \begin{pmatrix} 1 & x_{11} & \ldots & x_{1k} \\ 1 & x_{21} & \ldots & x_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \ldots & x_{nk} \end{pmatrix}$$

Se supone además que $\operatorname{rg}(\mathbf{X}) = k + 1 = m$ coincide con el número de parámetros.

Se trata de calcular el ajuste MC a un hiperplano k dimensional, donde β_0 es el punto de intersección del hiperplano con el eje y cuando $x_1 = x_2 = \cdots = x_k = 0$.

Las ecuaciones normales son $\mathbf{X}'\mathbf{X}\boldsymbol{\beta} = \mathbf{X}'\mathbf{Y}$ donde

$$\mathbf{X}'\mathbf{X} = \begin{pmatrix} n & \sum x_{i1} & \sum x_{i2} & \ldots & \sum x_{ik} \\ & \sum x_{i1}^2 & \sum x_{i1}x_{i2} & \ldots & \sum x_{i1}x_{ik} \\ & & \sum x_{i2}^2 & \ldots & \sum x_{i2}x_{ik} \\ & & & \ddots & \vdots \\ & & & & \sum x_{ik}^2 \end{pmatrix} \qquad \mathbf{X}'\mathbf{Y} = \begin{pmatrix} \sum y_i \\ \sum x_{i1}y_i \\ \vdots \\ \sum x_{ik}y_i \end{pmatrix}$$

y cuya solución son las estimaciones $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$, sin ningún problema de estimabilidad ya que el modelo es de rango máximo. Además, estas estimaciones son insesgadas y de varianza mínima.

Las predicciones de los valores de Y dadas las observaciones de las variables regresoras $x_1, \ldots, x_k$ son

$$\widehat{\mathbf{Y}} = \mathbf{X}\widehat{\boldsymbol{\beta}} = \mathbf{P}\mathbf{Y}$$

es decir

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_k x_{ik} \qquad i = 1, \ldots, n \tag{8.2}$$

También podemos considerar el modelo con las variables regresoras centradas

$$\mathbf{Y} = (\mathbf{1}, \mathbf{Z}) \begin{pmatrix} \gamma \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix} + \boldsymbol{\epsilon}$$

donde las columnas de $\mathbf{Z}$ tienen media cero, es decir, $\mathbf{z}_{(j)} = \mathbf{x}_{(j)} - \bar{x}_j \mathbf{1}$ o

$$z_{ij} = x_{ij} - \bar{x}_j \qquad i = 1, \ldots, n \quad j = 1, \ldots, k$$

Este modelo es equivalente al anterior con $\gamma = \beta_0 + \sum_j \bar{x}_j \beta_j$, pero su estimación es más sencilla porque

$$[(\mathbf{1}, \mathbf{Z})'(\mathbf{1}, \mathbf{Z})]^{-1} = \begin{pmatrix} 1/n & \mathbf{0} \\ \mathbf{0} & (\mathbf{Z}'\mathbf{Z})^{-1} \end{pmatrix}$$

ya que $\mathbf{Z}'\mathbf{1} = \mathbf{0}$.
Entonces

$$\hat{\gamma} = \bar{y} \qquad (\hat{\beta}_1, \ldots, \hat{\beta}_k)' = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'(\mathbf{Y} - \mathbf{1}\bar{y})$$

Si definimos la matriz simétrica de varianzas-covarianzas, aunque de forma convencional, entre las variables $Y, x_1, \ldots, x_k$

$$\mathbf{S} = \begin{pmatrix} s_y^2 & \mathbf{S}_{\mathbf{x}y} \\ \mathbf{S}_{y\mathbf{x}} & \mathbf{S}_{\mathbf{xx}} \end{pmatrix} = n^{-1}(\mathbf{Y} - \mathbf{1}\bar{y}, \mathbf{Z})'(\mathbf{Y} - \mathbf{1}\bar{y}, \mathbf{Z})$$

resulta

$$(\hat{\beta}_1, \ldots, \hat{\beta}_k)' = \mathbf{S}_{\mathbf{xx}}^{-1}\mathbf{S}_{y\mathbf{x}}$$

Por todo ello, si consideramos las medias de los datos

$$\bar{y} = (1/n)\sum_i y_i \qquad \bar{x}_j = (1/n)\sum_i x_{ij} \qquad j = 1, \ldots, k$$

8.2 se expresa también en la forma

$$\hat{y}_i - \bar{y} = \hat{\beta}_1(x_{i1} - \bar{x}_1) + \cdots + \hat{\beta}_k(x_{ik} - \bar{x}_k)$$

Finalmente, observemos que el parámetro β_j, $j = 1, \ldots, k$, indica el incremento en Y cuando x_j aumenta en una unidad manteniéndose constantes el resto de variables regresoras. A veces se les llama coeficientes de regresión parcial porque reflejan el efecto de una variable regresora dada la presencia del resto que permanece constante.
Los residuos de la regresión son

$$\mathbf{e} = \mathbf{Y} - \widehat{\mathbf{Y}} = (\mathbf{I} - \mathbf{P})\mathbf{Y}$$

que verifican las propiedades que se han explicado para la regresión simple en la página 93 (ver ejercicio 6.4).

8.2. Medidas de ajuste

Como en la regresión simple, la evaluación del ajuste del hiperplano de regresión a los datos se puede hacer con la *varianza residual* o estimación MC de σ^2.
La suma de cuadrados residual es

$$\text{SCR} = \mathbf{e}'\mathbf{e} = \sum (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \cdots - \hat{\beta}_k x_{ik})^2 = \mathbf{Y}'\mathbf{Y} - \mathbf{Y}'\mathbf{X}\hat{\boldsymbol{\beta}}$$

que tiene $n - m$ grados de libertad. Así, la estimación centrada de la varianza del diseño es el llamado *error cuadrático medio*

$$\hat{\sigma}^2 = \text{SCR}/(n-m) = \text{ECM}$$

Su raíz cuadrada $\hat{\sigma}$, que tiene las mismas unidades que Y, es el *error estándar de la regresión múltiple*. También aquí, la varianza residual y el error estándar dependen de las unidades de la variable respuesta y no son útiles para comparar diversas regresiones.

En primer lugar, vamos a introducir el *coeficiente de correlación múltiple* de Y sobre $x_1, \ldots, x_k$. El uso del término *correlación* es convencional puesto que las variables regresoras no son aleatorias. El coeficiente se define como la correlación muestral entre Y e $\hat{Y}$

$$r_{y\mathbf{x}} = \text{corr}(Y, \hat{Y}) = \frac{\sum (y_i - \overline{y})(\hat{y}_i - \overline{y})}{[\sum (y_i - \overline{y})^2 \sum (\hat{y}_i - \overline{y})^2]^{1/2}}$$

ya que $(1/n)\sum \hat{y}_i = \overline{y}$.

El coeficiente de correlación múltiple $r_{y\mathbf{x}}$ verifica $0 \leq r_{y\mathbf{x}} \leq 1$ y es una buena medida del ajuste de Y al modelo $\mathbf{X}\boldsymbol{\beta}$, pues

$$r_{y\mathbf{x}} = 1 \quad \Rightarrow \quad \|\mathbf{Y} - \widehat{\mathbf{Y}}\| = 0$$

El siguiente teorema, idéntico al teorema 6.2.1, justifica la definición del coeficiente de determinación como medida de ajuste.

Teorema 8.2.1

Las sumas de cuadrados asociadas a la regresión múltiple verifican:

(i) $\sum (y_i - \overline{y})^2 = \sum (y_i - \hat{y}_i)^2 + \sum (\hat{y}_i - \overline{y})^2$

(ii) $r^2_{y\mathbf{x}} = \dfrac{\sum (\hat{y}_i - \overline{y})^2}{\sum (y_i - \overline{y})^2}$

(iii) $\text{SCR} = \sum (y_i - \hat{y}_i)^2 = (1 - r^2_{y\mathbf{x}})S_y$

Demostración:

La descomposición en suma de cuadrados *(i)* se justifica de la misma forma que se ha visto en el teorema 6.2.1. También se puede ver el ejercicio 5.9.

El hecho fundamental es la ortogonalidad

$$(\mathbf{Y} - \widehat{\mathbf{Y}})'\widehat{\mathbf{Y}} = 0$$

pues el vector $\mathbf{e} = \mathbf{Y} - \widehat{\mathbf{Y}} = \mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}}$ es ortogonal a $\Omega = \langle \mathbf{X} \rangle$, mientras que $\widehat{\mathbf{Y}} = \mathbf{X}\widehat{\boldsymbol{\beta}} \in \Omega$ (ver teorema 2.4.2 y su interpretación geométrica).

Luego

$$\begin{aligned}\sum (y_i - \overline{y})(\hat{y}_i - \overline{y}) &= \sum (y_i - \hat{y}_i + \hat{y}_i - \overline{y})(\hat{y}_i - \overline{y}) \\ &= \sum (y_i - \hat{y}_i)(\hat{y}_i - \overline{y}) + \sum (\hat{y}_i - \overline{y})^2 \\ &= \sum (\hat{y}_i - \overline{y})^2\end{aligned}$$

puesto que el primer sumando es nulo. Teniendo en cuenta la definición de $r_{y\mathbf{x}}$, es fácil deducir *(ii)*. Finalmente, combinando *(i)* y *(ii)* obtenemos *(iii)*. ■

Como en 6.7, la descomposición *(i)* del teorema anterior justifica la definición del coeficiente de determinación

$$R^2 = \frac{\text{VE}}{\text{VT}} = 1 - \frac{\text{SCR}}{S_y}$$

También aquí, esta medida del ajuste verifica $0 \leq R^2 \leq 1$ y coincide con el cuadrado del coeficiente de correlación múltiple

$$R^2 = 1 - \frac{(1 - r^2_{y\mathbf{x}})S_y}{S_y} = r^2_{y\mathbf{x}}$$

Sin embargo, el coeficiente de correlación múltiple $r_{y\mathbf{x}}$ es una medida de la asociación lineal entre la variable respuesta Y y las regresoras $\mathbf{x} = (x_1, \ldots, x_k)$ que, en este caso, es convencional.

Como R^2 es la proporción de variabilidad explicada por las variables regresoras, resulta que si $R^2 \approx 1$, entonces la mayor parte de la variabilidad es explicada por dichas variables. Pero R^2 es la proporción de la variabilidad total explicada por el modelo con todas las variables frente al modelo $y = \beta_0$, de manera que un R^2 alto muestra que el modelo mejora el modelo nulo y por tanto sólo tiene sentido comparar coeficientes de determinación entre modelos anidados (casos particulares).

Además un valor grande de R^2 no necesariamente implica que el modelo lineal es bueno. El coeficiente R^2 no mide si el modelo lineal es apropiado. Es posible que un modelo con un valor alto de R^2 proporcione estimaciones y predicciones pobres, poco precisas. El análisis de los residuos es imprescindible.

Tampoco está claro lo que significa un valor "grande", ya que problemas en diversas ciencias (física, ingeniería, sociología,. . .) tienen razonablemente criterios diferentes.

Por otra parte, cuando se añaden variables regresoras R^2 crece, pero eso no significa que el nuevo modelo sea superior:

$$R^2_{\text{nuevo}} = 1 - \frac{\text{SCR}_{\text{nuevo}}}{S_y} \geq R^2 = 1 - \frac{\text{SCR}}{S_y} \qquad \Rightarrow \qquad \text{SCR}_{\text{nuevo}} \leq \text{SCR}$$

pero es posible que

$$\text{ECM}_{\text{nuevo}} = \frac{\text{SCR}_{\text{nuevo}}}{n - (m + p)} \geq \text{ECM} = \frac{\text{SCR}}{n - m}$$

luego, en esta situación, el nuevo modelo será peor. Así, como R^2 crece al añadir nuevas variables regresoras, se corre el peligro de sobreajustar el modelo añadiendo términos innecesarios. El coeficiente de determinación ajustado penaliza esto.

Definición 8.2.1

Una medida del ajuste de la regresión múltiple a los datos es el coeficiente de determinación *o proporción de variabilidad explicada*

$$R^2 = \frac{\text{VE}}{\text{VT}} = 1 - \frac{\text{SCR}}{S_y}$$

Sin embargo, para corregir el peligro de sobreajuste se define el coeficiente de determinación ajustado *como*

$$\bar{R}^2 = 1 - \frac{\text{SCR}/(n - m)}{S_y/(n - 1)} = 1 - \frac{n - 1}{n - m}(1 - R^2)$$

Cuando $\bar{R}^2$ y R^2 son muy distintos, el modelo ha sido sobreajustado y debemos eliminar variables o términos.

8.3. Inferencia sobre los coeficientes de regresión

Cuando asumimos la hipótesis de normalidad sobre la distribución de los errores $\boldsymbol{\epsilon} \sim N_n(\mathbf{0}, \sigma^2\mathbf{I})$, se deduce la normalidad de la variable respuesta

$$\mathbf{Y} \sim N_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$$

lo que nos permite utilizar las distribuciones asociadas a los estimadores de los parámetros que hemos estudiado.

En el capítulo de contraste de hipótesis se ha visto de varias formas (ver 5.10) que para una función paramétrica estimable $\mathbf{a}'\boldsymbol{\beta}$

$$\frac{\mathbf{a}'\widehat{\boldsymbol{\beta}} - \mathbf{a}'\boldsymbol{\beta}}{(\hat{\sigma}^2 \cdot \mathbf{a}'(\mathbf{X}'\mathbf{X})^{-}\mathbf{a})^{1/2}} \sim t_{n-r}$$

En nuestro caso, todas las funciones paramétricas son estimables ya que $r = k + 1 = m$. De modo que el estimador $\hat{\beta}_j$ verifica

$$\frac{\hat{\beta}_j - \beta_j}{\sqrt{\text{ECM}\, c_{jj}}} \sim t_{n-m} \tag{8.3}$$

donde c_{jj} es el j-ésimo elemento de la diagonal de $(\mathbf{X}'\mathbf{X})^{-1}$ y $\hat{\sigma}^2 = \text{SCR}/(n-m) = \text{ECM}$.

En consecuencia, los intervalos de confianza de los coeficientes de regresión β_j con un nivel de confianza $100(1-\alpha)\,\%$ son

$$\hat{\beta}_j \pm t_{n-m}(\alpha) \cdot \text{ee}(\hat{\beta}_j)$$

donde $\text{ee}(\hat{\beta}_j) = \sqrt{\text{ECM}\, c_{jj}}$.

En cuanto a los intervalos de confianza para la respuesta media o los intervalos de predicción para una respuesta concreta, su deducción es similar al caso de la regresión simple.

Si $\mathbf{x}_0 = (1, x_{01}, \ldots, x_{0k})'$ recoge una observación particular del conjunto de variables regresoras, el intervalo de confianza con nivel $100(1-\alpha)\,\%$ para la respuesta media $E[Y|\mathbf{x}_0]$ está centrado en su estimación $\hat{y}_0 = \mathbf{x}_0'\widehat{\boldsymbol{\beta}}$

$$\hat{y}_0 \pm t_{n-m}(\alpha) \cdot (\text{ECM}\, \mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0)^{1/2}$$

ya que $E(\hat{y}_0) = \mathbf{x}_0'\boldsymbol{\beta} = E[Y|\mathbf{x}_0]$ y $\text{var}(\hat{y}_0) = \sigma^2\mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0$.

Extrapolación oculta

En la estimación de la respuesta media o la predicción de nuevas respuestas en un punto $(x_{01}, \ldots, x_{0k})$ debemos ser muy cuidadosos con la extrapolación. Si únicamente tenemos en cuenta el producto cartesiano de los recorridos de las variables regresoras, es fácil considerar la predicción para un punto que puede estar fuera de la nube de puntos con la que hemos calculado la regresión. Para evitar este problema deberemos ceñirnos al menor conjunto convexo que contiene los n puntos originales y que recibe el nombre de casco (*hull*) de las variables regresoras (ver figura 8.1).

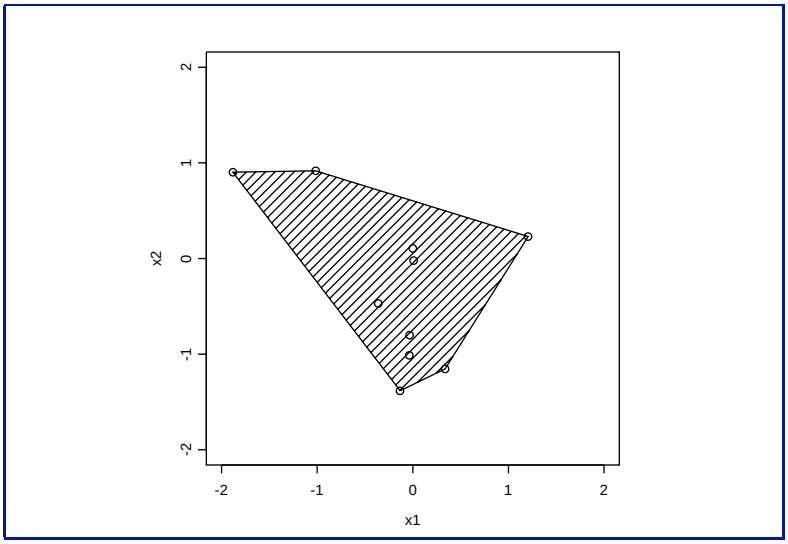

Figura 8.1: Conjunto convexo para los puntos de dos variables regresoras

Si consideramos los elementos h_{ii} de la diagonal de la matriz proyección $\mathbf{P} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$, podemos definir $h_{\text{máx}} = \text{máx}\{h_{11}, \ldots, h_{nn}\}$ y se puede comprobar que

$$\mathbf{x}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x} \leq h_{\text{máx}}$$

es un elipsoide que contiene al casco. No es el menor elipsoide, pero es el más fácil de calcular.

Así pues, para evitar en lo posible la extrapolación, podemos comprobar en el punto $\mathbf{x}_0 = (1, x_{01}, \ldots, x_{0k})'$ si

$$\mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0 < h_{\text{máx}}$$

Contraste de significación de la regresión

La hipótesis de mayor interés es la afirmación de que Y es independiente de las variables $x_1, \ldots, x_k$, es decir

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0 \tag{8.4}$$

El Análisis de la Varianza del teorema 5.3.1 se puede aplicar al contraste de la significación conjunta de los coeficientes de regresión puesto que se trata de una hipótesis contrastable del tipo $H_0 : \mathbf{A}\boldsymbol{\beta} = 0$, donde

$$\mathbf{A} = \begin{pmatrix} 0 & 1 & 0 & \ldots & 0 \\ 0 & 0 & 1 & \ldots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \ldots & 1 \end{pmatrix} \qquad \text{rango } \mathbf{A} = k$$

Si H_0 es cierta, al igual que en 6.9, la estimación del único parámetro que queda en el modelo es $\hat{\beta}_{0|H} = \overline{y}$ y la suma de cuadrados residual es

$$\text{SCR}_H = \sum (y_i - \overline{y})^2 = S_y$$

que tiene $n - 1$ grados de libertad.

La descomposición en suma de cuadrados es

$$S_y = \text{SCR} + (\text{SCR}_H - \text{SCR})$$

es decir

$$\sum (y_i - \overline{y})^2 = \sum (y_i - \hat{y}_i)^2 + \sum (\hat{y}_i - \overline{y})^2$$

La tabla siguiente recoge esta descomposición y realiza el contraste de la hipótesis. La hipótesis se rechaza si $F > F_{k,n-k-1}(\alpha)$.

Fuente de variación	grados de libertad	suma de cuadrados	cuadrados medios	F
Regresión	k	$\text{SC}_R = \text{SCR}_H - \text{SCR}$	CM_R	CM_R/ECM
Error	$n - k - 1$	SCR	ECM	
Total	$n - 1$	S_y		

Tabla 8.1: Análisis de la varianza para contrastar la significación de la regresión múltiple

Teniendo en cuenta las fórmulas del teorema 8.2.1

$$\text{SCR}_H - \text{SCR} = r^2_{y\mathbf{x}} S_y$$

y deducimos una expresión equivalente al estadístico F

$$F = \frac{r^2_{y\mathbf{x}}}{1 - r^2_{y\mathbf{x}}} \cdot \frac{n-k-1}{k}$$

que también se presenta en forma de tabla.

Fuente de variación	Grados de libertad	Suma de cuadrados	F
Regresión	k	$r^2_{y\mathbf{x}} S_y$	$\frac{r^2_{y\mathbf{x}}}{1 - r^2_{y\mathbf{x}}} \cdot \frac{n-k-1}{k}$
Residuo	$n-k-1$	$(1 - r^2_{y\mathbf{x}})S_y$	
Total	$n-1$	S_y	

Tabla 8.2: Análisis de la varianza en regresión múltiple

Del mismo modo que en la sección 6.5 la hipótesis 8.4 equivale a afirmar que el coeficiente de correlación múltiple poblacional es cero y se resuelve con el contraste asociado a la tabla anterior.

Significación parcial

El contraste de significación de un coeficiente de regresión particular $H_0 : \beta_j = 0$, para un j fijo, se resuelve con el estadístico 8.3 y la región crítica

$$\left| \frac{\hat{\beta}_j}{(\mathrm{ECM}\, c_{jj})^{1/2}} \right| > t_{n-k-1}(\alpha) \tag{8.5}$$

donde c_{jj} es el j-ésimo elemento de la diagonal de $(\mathbf{X}'\mathbf{X})^{-1}$.

Aceptar esta hipótesis significa que la variable regresora x_j se puede eliminar del modelo. Sin embargo, es preciso actuar con cuidado ya que se trata de un contraste *parcial* porque el coeficiente $\hat{\beta}_j$ depende de todas las otras variables regresoras x_i $(i \neq j)$. Es un contraste de la contribución de x_j dada la presencia de las otras variables regresoras en el modelo.

De forma general podemos estudiar la contribución al modelo de un subconjunto de las variables regresoras. Esto se puede hacer mediante la descomposición de la suma de cuadrados asociada a un contraste de modelos.

Consideremos el modelo lineal completo, dividido en dos grupos de variables regresoras,

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} = \begin{pmatrix} \mathbf{X}_1 & \mathbf{X}_2 \end{pmatrix} \begin{pmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{pmatrix} + \boldsymbol{\epsilon}$$

donde $\mathbf{X}_1$ es $n \times (m-p)$ y $\mathbf{X}_2$ es $n \times p$.

Para este modelo, la estimación de los parámetros es $\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$ y la suma de cuadrados de la regresión es

$$\mathrm{SC}_R(\boldsymbol{\beta}) = \mathrm{SCR}_H - \mathrm{SCR} = \mathbf{Y}'\mathbf{Y} - (\mathbf{Y}'\mathbf{Y} - \widehat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{Y}) = \widehat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{Y}$$

con m grados de libertad. Esto es así porque la hipótesis considerada es $H_0 : \boldsymbol{\beta} = \mathbf{0}$ y, bajo esta hipótesis, $\mathrm{SCR}_H = \mathbf{Y}'\mathbf{Y}$.

Para hallar la contribución de los términos de $\boldsymbol{\beta}_2$ en la regresión, podemos considerar la hipótesis $H_0 : \boldsymbol{\beta}_2 = \mathbf{0}$ que es equivalente al modelo reducido $\mathbf{Y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \boldsymbol{\epsilon}$. Bajo esta hipótesis, la estimación de los parámetros es $\widehat{\boldsymbol{\beta}}_1 = (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{Y}$ y la suma de cuadrados de la regresión

$$\mathrm{SC}_R(\boldsymbol{\beta}_1) = \widehat{\boldsymbol{\beta}}_1'\mathbf{X}_1'\mathbf{Y}$$

con $m-p$ grados de libertad.

Luego la suma de cuadrados de la regresión debida a $\boldsymbol{\beta}_2$, dado que $\boldsymbol{\beta}_1$ está ya en el modelo, es

$$\mathrm{SC}_R(\boldsymbol{\beta}_2|\boldsymbol{\beta}_1) = \mathrm{SC}_R(\boldsymbol{\beta}) - \mathrm{SC}_R(\boldsymbol{\beta}_1)$$

con $m - (m - p) = p$ grados de libertad.

Como $\mathrm{SC}_R(\boldsymbol{\beta}_2|\boldsymbol{\beta}_1)$ es independiente de SCR, la hipótesis $H_0 : \boldsymbol{\beta}_2 = \mathbf{0}$ se puede contrastar con el estadístico

$$\frac{\mathrm{SC}_R(\boldsymbol{\beta}_2|\boldsymbol{\beta}_1)/p}{\mathrm{ECM}} \sim F_{p,n-m}$$

que se puede llamar una F parcial, pues mide la contribución de $\mathbf{X}_2$ considerando que $\mathbf{X}_1$ está en el modelo.

Por ejemplo, la suma de cuadrados de la regresión $\mathrm{SC}_R(\beta_j|\beta_0,\beta_1,\ldots,\beta_{j-1},\beta_{j+1},\ldots,\beta_k)$ para $1 \le j \le k$ es el crecimiento en la suma de cuadrados debido a añadir x_j al modelo que ya contiene todas las otras variables, como si fuera la última variable añadida al modelo. El contraste es equivalente al contraste 8.5.

Estos contrastes F parciales juegan un papel muy importante en la búsqueda del mejor conjunto de variables regresoras a utilizar en un modelo. Por ejemplo, en el modelo parabólico $Y = \beta_0 + \beta_1 x + \beta_2 x^2 + \epsilon$ estaremos interesados en $\mathrm{SC}_R(\beta_1|\beta_0)$ y luego en $\mathrm{SC}_R(\beta_2|\beta_0,\beta_1)$ que es la contribución cuadrática al modelo lineal simple.

En el modelo $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon$, la descomposición en suma de cuadrados es

$$S_y = \mathrm{SC}_R(\beta_1,\beta_2,\beta_3|\beta_0) + \mathrm{SCR}$$

pero

$$\begin{aligned}\mathrm{SC}_R(\beta_1,\beta_2,\beta_3|\beta_0) &= \mathrm{SC}_R(\beta_1|\beta_0) + \mathrm{SC}_R(\beta_2|\beta_0,\beta_1) + \mathrm{SC}_R(\beta_3|\beta_0,\beta_1,\beta_2)\\ &= \mathrm{SC}_R(\beta_2|\beta_0) + \mathrm{SC}_R(\beta_1|\beta_0,\beta_2) + \mathrm{SC}_R(\beta_3|\beta_0,\beta_1,\beta_2)\\ &= \ldots\end{aligned}$$

Sin embargo, hay que ir con cuidado porque este método no siempre produce una partición de la suma de cuadrados de la regresión y, por ejemplo,

$$\mathrm{SC}_R(\beta_1,\beta_2,\beta_3|\beta_0) \neq \mathrm{SC}_R(\beta_1|\beta_2,\beta_3,\beta_0) + \mathrm{SC}_R(\beta_2|\beta_1,\beta_3,\beta_0) + \mathrm{SC}_R(\beta_3|\beta_1,\beta_2,\beta_0)$$

Un resultado interesante se tiene cuando las columnas de $\mathbf{X}_1$ y $\mathbf{X}_2$ son ortogonales, ya que entonces

$$\mathrm{SC}_R(\boldsymbol{\beta}_2|\boldsymbol{\beta}_1) = \mathrm{SC}_R(\boldsymbol{\beta}_2) \qquad \mathrm{SC}_R(\boldsymbol{\beta}_1|\boldsymbol{\beta}_2) = \mathrm{SC}_R(\boldsymbol{\beta}_1)$$

Región de confianza y intervalos simultáneos

Del mismo modo que hemos explicado en 6.3.6, en regresión múltiple la región con una confianza conjunta del $100(1-\alpha)\,\%$ es

$$\frac{(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})'\mathbf{X}'\mathbf{X}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})}{m\mathrm{ECM}} \le F_{m,n-m}(\alpha)$$

Los intervalos simultáneos para los coeficientes de la regresión son del tipo

$$\hat{\beta}_j \pm \Delta \cdot \mathrm{ee}(\hat{\beta}_j)$$

para un conjunto de s coeficientes entre los $k+1$. Por ejemplo, el método de Scheffé proporciona los intervalos simultáneos

$$\hat{\beta}_j \pm (sF_{s,n-k-1}(\alpha))^{1/2} \cdot \mathrm{ee}(\hat{\beta}_j)$$

Los intervalos simultáneos para un conjunto de s respuestas medias, una para cada uno de los puntos $\mathbf{x}_{01},\ldots,\mathbf{x}_{0s}$, son

$$\hat{y}_{\mathbf{x}_{0j}} \pm \Delta(\mathrm{ECM}\,\mathbf{x}_{0j}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_{0j})^{1/2}$$

donde $\Delta = (sF_{s,n-k-1}(\alpha))^{1/2}$ por el método de Scheffé.

8.4. Coeficientes de regresión estandarizados

Es difícil comparar coeficientes de regresión porque la magnitud de $\hat{\beta}_j$ refleja las unidades de medida de la variable regresora. Por ejemplo, en el modelo

$$Y = 5 + x_1 + 1000x_2$$

donde x_1 se mide en litros y x_2 en mililitros, aunque $\hat{\beta}_2 = 1000$ es mucho mayor que $\hat{\beta}_1 = 1$, el efecto sobre Y es el mismo.

Generalmente, las unidades de los coeficientes de regresión son

$$\text{unidades } \hat{\beta}_j = \frac{\text{unidades } Y}{\text{unidades } x_j}$$

Por todo ello, frecuentemente es de gran ayuda trabajar con variables estandarizadas que producen coeficientes de regresión sin dimensión. Básicamente hay dos técnicas:

Escala normal unidad

$$z_{ij} = \frac{x_{ij} - \bar{x}_j}{\hat{s}_j} \qquad i = 1, \ldots, n; j = 1, \ldots, k$$
$$y_i^* = \frac{y_i - \bar{y}}{\hat{s}_y} \qquad i = 1, \ldots, n$$

donde

$$\bar{x}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij} \qquad \hat{s}_j^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2 \qquad \hat{s}_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$$

El modelo es

$$y_i^* = b_0 + b_1 z_{i1} + b_2 z_{i2} + \cdots + b_k z_{ik} + \eta_i \qquad i = 1, \ldots, n$$

donde las variables regresoras y la variable respuesta tienen media cero y varianza muestral uno. La estimación del modelo es $\widehat{\mathbf{b}} = (\hat{b}_1, \ldots, \hat{b}_k)' = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{Y}^*$ y $\hat{b}_0 = \bar{y}^* = 0$.

Escala longitud unidad

$$w_{ij} = \frac{x_{ij} - \bar{x}_j}{S_j^{1/2}} \qquad i = 1, \ldots, n; j = 1, \ldots, k$$
$$y_i^0 = \frac{y_i - \bar{y}}{S_y^{1/2}} \qquad i = 1, \ldots, n$$

donde

$$S_j = \sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2 \qquad S_y = \sum_{i=1}^{n}(y_i - \bar{y})^2$$

El modelo es

$$y_i^0 = b_1 w_{i1} + b_2 w_{i2} + \cdots + b_k w_{ik} + \eta_i \qquad i = 1, \ldots, n$$

donde las variables regresoras y la variable respuesta tienen media cero y longitud

$$\sqrt{\sum_{i=1}^{n}(w_{ij} - \bar{w}_j)^2} = 1$$

y la estimación de los parámetros es $\widehat{\mathbf{b}} = (\mathbf{W}'\mathbf{W})^{-1}\mathbf{W}'\mathbf{Y}^0$.

Pero en este modelo tenemos

$$\mathbf{W}'\mathbf{W} = \mathbf{R}_{\mathbf{xx}} = \begin{pmatrix} 1 & r_{12} & \dots & r_{1k} \\ r_{21} & 1 & \dots & r_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ r_{k1} & r_{k2} & \dots & 1 \end{pmatrix}$$

donde $\mathbf{R}_{\mathbf{xx}}$ es la matriz de correlaciones de las variables regresoras ya que

$$r_{ij} = \frac{\sum_{s=1}^{n}(x_{si} - \bar{x}_i)(x_{sj} - \bar{x}_j)}{(S_i S_j)^{1/2}}$$

También podemos considerar que $\mathbf{W}'\mathbf{Y}^0 = \mathbf{R}_{\mathbf{x}y}$ es el vector de correlaciones de las variables regresoras con la variable respuesta. También aquí el término correlación es convencional.

En todo caso, como

$$\mathbf{Z}'\mathbf{Z} = (n-1)\mathbf{W}'\mathbf{W}$$
$$\mathbf{Z}'\mathbf{Y}^* = (n-1)\mathbf{W}'\mathbf{Y}^0$$

las estimaciones de $\mathbf{b} = (b_1, \dots, b_k)'$ por ambos métodos son idénticas.

Definición 8.4.1

Se llaman coeficientes de regresión estandarizados *los que se obtienen como solución del sistema de ecuaciones*

$$\begin{aligned} b_1 &+ r_{12}b_2 + \cdots + r_{1k}b_k = r_{1y} \\ r_{21}b_1 &+ b_2 + \cdots + r_{2k}b_k = r_{2y} \\ &\vdots \qquad\quad \vdots \qquad\quad \vdots \\ r_{k1}b_1 &+ r_{k2}b_2 + \cdots + b_k = r_{ky} \end{aligned}$$

es decir

$$\mathbf{R}_{\mathbf{xx}}\mathbf{b} = \mathbf{R}_{\mathbf{x}y}$$

donde $\mathbf{R}_{\mathbf{xx}}$ *es la matriz de coeficientes de correlación entre las variables regresoras y* $\mathbf{R}_{\mathbf{x}y} = (r_{1y}, \dots, r_{ky})'$ *el vector columna con los coeficientes de correlación entre las variables regresoras y la respuesta.*

Los coeficientes de regresión ordinarios se deducen de las ecuaciones

$$\hat{\beta}_j = \hat{b}_j \left(\frac{S_y}{S_j}\right)^{1/2} = \hat{b}_j \frac{s_y}{s_j} \qquad j = 1, \dots, k$$

$$\hat{\beta}_0 = \bar{y} - \sum_{j=1}^{k} \hat{\beta}_j \bar{x}_j$$

Además, el coeficiente de determinación es

$$R^2 = r^2_{y\mathbf{x}} = \hat{b}_1 r_{1y} + \hat{b}_2 r_{2y} + \cdots + \hat{b}_k r_{ky}$$

Algunos paquetes estadísticos calculan ambos conjuntos de coeficientes de regresión. En algún caso, a los coeficientes de regresión estandarizados les llaman "beta coeficientes" lo que para nosotros es confuso.

Finalmente señalaremos que debemos cuidar las interpretaciones puesto que los coeficientes estandarizados todavía son parciales, es decir, miden el efecto de x_j dada la presencia de las otras variables regresoras. También $\hat{b}_j$ está afectado por el recorrido de los valores de las variables regresoras, de modo que es peligroso utilizar $\hat{b}_j$ para medir la importancia relativa de la variable regresora x_j.

Ejemplo 8.4.1

En un estudio sobre la incidencia que puede tener sobre el rendimiento en lenguaje Y, la comprensión lectora x_1 y la capacidad intelectual x_2, se obtuvieron datos sobre 10 estudiantes tomados al azar de un curso de básica (ver tabla 8.3).

Y	x_1	x_2
3	1	3
2	1	4
4	3	7
9	7	9
6	8	7
7	7	6
2	4	5
6	6	8
5	6	5
8	9	7

Tabla 8.3: Datos del rendimiento en lenguaje

La matriz de correlaciones, las medias y las desviaciones típicas son:

$$\begin{array}{cccccc} & x_1 & x_2 & Y & & \\ x_1 & 1 & 0.6973 & 0.8491 & \overline{x_1} = 5.2 & s_1 = 2.82 \\ x_2 & & 1 & 0.7814 & \overline{x_2} = 6.1 & s_2 = 1.86 \\ Y & & & 1 & \overline{y} = 5.2 & s_y = 2.44 \end{array}$$

Empezaremos planteando el sistema

$$\begin{aligned} b_1 + 0.6973 \cdot b_2 &= 0.8491 \\ 0.6973 \cdot b_1 + b_2 &= 0.7814 \end{aligned}$$

cuya solución es

$$\hat{b}_1 = 0.592 \qquad \hat{b}_2 = 0.368$$

Entonces

$$\hat{\beta}_1 = \hat{b}_1 \frac{s_y}{s_1} = 0.512 \qquad \hat{\beta}_2 = \hat{b}_2 \frac{s_y}{s_2} = 0.485$$

$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x_1} - \hat{\beta}_2 \overline{x_2} = -0.424$$

La ecuación de regresión es

$$y = -0.424 + 0.512x_1 + 0.485x_2$$

El coeficiente de determinación es

$$R^2 = r^2_{y\mathbf{x}} = \hat{b}_1 \cdot 0.849 + \hat{b}_2 \cdot 0.781 = 0.791$$

y puede afirmarse que hay una buena relación entre el rendimiento en lenguaje y la comprensión lectora y la capacidad intelectual.

Finalmente, para decidir sobre la hipótesis $H_0 : \beta_1 = \beta_2 = 0$ calcularemos

$$F = \frac{r^2_{y\mathbf{x}}}{1 - r^2_{y\mathbf{x}}} \cdot \frac{10 - 3}{3 - 1} = 13.22$$

con 2 y 7 grados de libertad. Así H_0 puede ser rechazada, es decir, la relación anterior es significativa.

8.5. Multicolinealidad

Cuando la matriz $\mathbf{X}$ no es de rango máximo, sabemos que $\mathbf{X}'\mathbf{X}$ es singular y no podemos calcular su inversa. Ya sabemos que la solución puede ser la utilización de alguna g-inversa, aunque ello implica que la solución de las ecuaciones normales no es única. En el caso de la regresión múltiple es difícil, aunque no imposible, que alguna columna sea linealmente dependiente de las demás. Si ocurriera esto diríamos que existe colinealidad entre las columnas de $\mathbf{X}$. Sin embargo, el término colinealidad o multicolinealidad se refiere al caso, mucho más frecuente, de que la dependencia entre las columnas no es exacta sino aproximada, es decir, a la quasi-dependencia lineal entre las variables regresoras. Esto puede provocar problemas de computación de los parámetros y en el cálculo de la precisión de los mismos (ver Apéndice A.4).

Entre las múltiples formas de detección de la multicolinealidad vamos a destacar el cálculo de los factores de inflación de la varianza. Nosotros hemos visto que la matriz de varianzas-covarianzas de los estimadores de los parámetros de un modelo lineal es

$$\text{var}(\widehat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$$

Si consideramos el modelo de regresión estandarizado por la escala de longitud unidad, la matriz de varianzas-covarianzas de los coeficientes de regresión estandarizados es

$$\text{var}(\widehat{\mathbf{b}}) = \tilde{\sigma}^2\mathbf{R}_{\mathbf{xx}}^{-1}$$

donde $\tilde{\sigma}^2$ es la varianza del error del modelo transformado. En particular, la varianza de uno de los coeficientes es

$$\text{var}(\hat{b}_j) = \tilde{\sigma}^2[\mathbf{R}_{\mathbf{xx}}^{-1}]_{jj}$$

donde $[\mathbf{R}_{\mathbf{xx}}^{-1}]_{jj}$ es el j-ésimo elemento de la diagonal de la matriz. Estas varianzas pueden estar "infladas" a causa de la multicolinealidad que puede ser evidente a partir de la observación de los elementos no nulos fuera de la diagonal de $\mathbf{R}_{\mathbf{xx}}$, es decir, de las correlaciones simples entre las variables regresoras.

Definición 8.5.1

Los elementos de la diagonal de la matriz $\mathbf{R}_{\mathbf{xx}}^{-1}$ *se llaman* FIV *o factores de inflación de la varianza*[1] *ya que*

$$\text{var}(\hat{b}_j) = \tilde{\sigma}^2\text{FIV}_j$$

Se demuestra que

$$\text{FIV}_j = (1 - R_j^2)^{-1}$$

donde R_j^2 es el coeficiente de determinación múltiple de la variable regresora x_j con todas las demás variables regresoras.

El factor de inflación de la varianza $\text{FIV}_j = 1$ cuando $R_j^2 = 0$, es decir, cuando x_j no depende linealmente del resto de las variables. Cuando $R_j^2 \neq 0$, entonces $\text{FIV}_j > 1$ y si $R_j^2 \approx 1$, entonces FIV_j es grande. Así pues, el factor de inflación de la varianza mide el incremento que se produce en la varianza de los estimadores de los coeficientes de regresión al comparar dicha varianza con la que deberían tener si las variables regresoras fuesen incorrelacionadas.

Cuando $\text{FIV}_j > 10$ tenemos un grave problema de multicolinealidad. Algunos autores prefieren calcular la media de los FIV_j y alertar sobre la multicolinealidad cuando dicha media supera el número 10.

Una de las posibles soluciones tras la detección de multicolinealidad es la estimación por la regresión *ridge* (ver 4.3.1).

1. VIF en inglés

Ejemplo 8.5.1

Con los datos del ejemplo 8.4.1, la matriz de correlaciones $\mathbf{R_{xx}}$ *y su inversa son*

$$\mathbf{R_{xx}} = \begin{pmatrix} 1.0000 & 0.6973 \\ 0.6973 & 1.0000 \end{pmatrix} \qquad \mathbf{R_{xx}^{-1}} = \begin{pmatrix} 1.9465 & -1.3574 \\ -1.3574 & 1.9465 \end{pmatrix}$$

y los factores de inflación de la varianza son $\mathrm{FIV}_1 = 1.9465, \mathrm{FIV}_2 = 1.9465$*, que coinciden naturalmente cuando* $k = 2$.

8.6. Regresión polinómica

Supongamos que una variable aleatoria Y se ajusta a una variable de control x según un modelo polinómico de grado m

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \cdots + \beta_m x_i^m + \epsilon_i \tag{8.6}$$

Observemos que se trata de un modelo de regresión lineal múltiple de Y sobre las variables $x_1 = x, x_2 = x^2, \ldots, x_m = x^m$. Para una regresión polinómica de grado m, la matriz de diseño es

$$\mathbf{X} = \begin{pmatrix} 1 & x_1 & x_1^2 & \ldots & x_1^m \\ 1 & x_2 & x_2^2 & \ldots & x_2^m \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_n & x_n^2 & \ldots & x_n^m \end{pmatrix}$$

Estos modelos se pueden aplicar cuando el analista sabe que efectos curvilíneos están presentes en la función respuesta. También se pueden utilizar como aproximaciones a desconocidas, y posiblemente muy complejas, relaciones no lineales. Así, los polinomios se pueden considerar los desarrollos de Taylor de la función desconocida.

La regresión polinómica se justifica por el teorema de Weierstrass, el cual dice que toda función continua $f(x)$ se puede aproximar por un polinomio $P_m(x)$ de grado m adecuado. Se puede probar esta propiedad desde el punto de vista probabilístico:

Sea $f(x)$ una función continua en el intervalo $(0, 1)$ y consideremos

$$P_n(x) = \sum_{k=0}^{n} f(k/n) x^k (1-x)^{n-k}$$

llamados polinomios de Bernstein. Entonces $P_n(x)$ converge a $f(x)$ cuando $n \to \infty$, uniformemente en x.

Como en cualquier modelo lineal, la estimación de los parámetros de regresión se hace con las ecuaciones normales. Sin embargo, hay varios problemas especiales que se presentan en este caso.

1) Es muy importante que el orden del polinomio sea tan bajo como sea posible. Para utilizar polinomio de grado $m > 2$ se debe justificar con razones externas a los datos. Existen transformaciones de las variables, en particular de la respuesta, que hacen que el modelo sea de primer orden. Un modelo de orden bajo con una variable transformada es casi siempre preferible a un modelo de orden superior con la métrica original. Se trata de mantener el principio de parsimonia o simplicidad de los modelos.

2) Hay varias estrategias para elegir el grado del polinomio.

   Selección hacia adelante (forward selection): Se trata de ir ajustando modelos en orden creciente hasta que el test t para el término de mayor orden es no significativo ($\alpha = 0.1$).

Selección hacia atrás (backward selection): Se trata de ajustar un modelo de alto orden e ir eliminando términos si no son significativos para el test t ($\alpha = 0.1$).

Ambos métodos no necesariamente conducen al mismo modelo. En todo caso, hay que recordar el consejo anterior y tratar con modelos de orden dos o muy bajo.

3) Debemos ser muy cuidadosos con la extrapolación (ver página 137), ya que las consecuencias pueden ser ruinosas.

4) Cuando el orden del polinomio es alto, la matriz $\mathbf{X}'\mathbf{X}$ está mal condicionada (ver apéndice A.4 y sección 8.5). Esto provoca problemas graves para el cálculo de los coeficientes de regresión y deficiencias en la precisión de los mismos. En Seber [66] pág. 214 se ve un ejemplo en el que variaciones del orden de 10^{-10} en $\mathbf{X}'\mathbf{Y}$ producen variaciones del orden de 3 en los elementos de $\widehat{\boldsymbol{\beta}}$.

   De hecho, los modelos de regresión polinómicos están notablemente mal condicionados cuando el grado es mayor que 5 o 6, particularmente si los valores de x están igualmente espaciados.

5) Si los valores de x tienen un recorrido muy estrecho, esto puede conducir a la multicolinealidad entre las columnas de $\mathbf{X}$. Por ejemplo, si x varía entre 1 y 2, x^2 varía entre 1 y 4, lo que puede provocar una fuerte dependencia entre los datos de x y x^2.

Para reducir el efecto no esencial de la mala condición de los modelos de regresión polinómicos se deben centrar las variables regresoras. Además se pueden utilizar polinomios de Tchebychev o, mejor, polinomios ortogonales.

La utilización de polinomios de Tchebychev consiste en considerar el modelo

$$y_i = \gamma_0 T_0(x_i) + \gamma_1 T_1(x_i) + \cdots + \gamma_m T_m(x_i) + \epsilon_i$$

donde $T_j(x)$ es un polinomio de Tchebychev de grado j. Estos polinomios se generan mediante la relación de recurrencia

$$T_{j+1}(x) = 2xT_j(x) - T_{j-1}(x)$$

Tomando inicialmente

$$T_0(x) = 1 \qquad T_1(x) = x$$

se obtienen

$$\begin{aligned} T_2(x) &= 2x^2 - 1 \\ T_3(x) &= 4x^3 - 3x \\ T_4(x) &= 8x^4 - 8x^2 + 1 \\ &\vdots \end{aligned}$$

El campo de variación de x debe "normalizarse" adecuadamente entre -1 y 1 mediante un cambio de variable. Esto se hace en favor de la estabilidad numérica.

Los polinomios de Tchebychev tienen propiedades muy interesantes que sugieren que, para valores de x razonablemente espaciados, la matriz del modelo $\widetilde{\mathbf{X}}$ tiene columnas que son aproximadamente ortogonales, de forma que la matriz $\widetilde{\mathbf{X}}'\widetilde{\mathbf{X}}$ tiene los elementos de fuera de la diagonal bastante pequeños y generalmente está bien condicionada. Así pues, un procedimiento de cálculo de regresión polinómica consiste en usar polinomios de Tchebychev junto con un método de descomposición ortogonal de la matriz de diseño, como el algoritmo QR.

8.6.1. Polinomios ortogonales

El replanteamiento del modelo 8.6 mediante polinomios ortogonales permite una solución sencilla de los problemas numéricos mencionados.

Consideremos ahora el modelo

$$y_i = \gamma_0\phi_0(x_i) + \gamma_1\phi_1(x_i) + \cdots + \gamma_m\phi_m(x_i) + \epsilon_i \tag{8.7}$$

donde $\phi_j(x_i)$ es un polinomio de grado j en x_i $(j = 0, 1, \ldots, m)$. Supongamos que los m polinomios son ortogonales, es decir,

$$\sum_{i=1}^{n} \phi_j(x_i)\phi_{j'}(x_i) = 0 \qquad \forall j \neq j' \tag{8.8}$$

El modelo lineal es entonces

$$\mathbf{Y} = \widetilde{\mathbf{X}}\boldsymbol{\gamma} + \boldsymbol{\epsilon}$$

donde

$$\widetilde{\mathbf{X}} = \begin{pmatrix} \phi_0(x_1) & \phi_1(x_1) & \ldots & \phi_m(x_1) \\ \phi_0(x_2) & \phi_1(x_2) & \ldots & \phi_m(x_2) \\ \vdots & \vdots & \ddots & \vdots \\ \phi_0(x_n) & \phi_1(x_n) & \ldots & \phi_m(x_n) \end{pmatrix}$$

Entonces, debido a la ortogonalidad, tenemos que

$$\widetilde{\mathbf{X}}'\widetilde{\mathbf{X}} = \begin{pmatrix} \sum \phi_0^2(x_i) & 0 & \ldots & 0 \\ 0 & \sum \phi_1^2(x_i) & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \sum \phi_m^2(x_i) \end{pmatrix}$$

y la solución de las ecuaciones normales es

$$\hat{\gamma}_j = \frac{\sum_i \phi_j(x_i)y_i}{\sum_i \phi_j^2(x_i)} \qquad j = 0, 1, \ldots, m$$

lo que es cierto para toda m. La estructura ortogonal de $\widetilde{\mathbf{X}}$ implica que el estimador MC de γ_j $(j \leq m)$ es independiente del grado m del polinomio, lo que es una propiedad muy deseable.

Como $\phi_0(x)$ es un polinomio de grado cero, si tomamos $\phi_0(x) = 1$ tendremos $\hat{\gamma}_0 = \overline{y}$.

La suma de cuadrados residual es entonces

$$\mathrm{SCR}(m) = \sum (y_i - \overline{y})^2 - \sum_{j=1}^{m} \left(\sum_i \phi_j^2(x_i) \right) \hat{\gamma}_j^2 \tag{8.9}$$

cantidad que indicaremos por $Q(m)$.

En efecto:

$$\hat{y}_i = \sum_{j=0}^{m} \phi_j(x_i)\hat{\gamma}_j \qquad \text{siendo} \qquad \overline{y} = \phi_0(x_i)\hat{\gamma}_0$$

Aplicando *(i)* de 8.2.1 tenemos

$$\mathrm{SCR}(m) = \sum_i (y_i - \hat{y}_i)^2 = \sum_i (y_i - \overline{y})^2 - \sum_i (\hat{y}_i - \overline{y})^2$$

siendo ahora

$$\sum_i (\hat{y}_i - \overline{y})^2 = \sum_i \left(\sum_{j=1}^{m} \phi_j(x_i)\hat{\gamma}_j \right)^2$$

Por otra parte

$$(\sum_{j=1}^{m} \phi_j(x_i)\hat{\gamma}_j)^2 = \sum_j \sum_{j'} \phi_j(x_i)\hat{\gamma}_j \cdot \phi_{j'}(x_i)\hat{\gamma}_{j'}$$

y sumando respecto de i tenemos, considerando 8.8,

$$\begin{aligned} \sum_i (\hat{y}_i - \overline{y})^2 &= \sum_j \sum_{j'} \hat{\gamma}_j \hat{\gamma}_{j'} (\sum_i \phi_j(x_i)\phi_{j'}(x_i)) \\ &= \sum_{j=1}^{m} \hat{\gamma}_j^2 (\sum_{i=1}^{n} \phi_j^2(x_i)) \end{aligned}$$

lo que demuestra 8.9.

Existen diversos procedimientos para generar polinomios ortogonales (Fisher, Forsythe, Hayes, etc.).
En el caso particular que los valores de x sean igualmente espaciados podemos transformarlos de manera que

$$x_i = i - \tfrac{1}{2}(n+1) \qquad i = 1, 2, \ldots, n$$

Entonces se puede considerar el siguiente sistema de polinomios ortogonales

$$\begin{aligned} \phi_0(x) &= 1 \\ \phi_1(x) &= \lambda_1 x \\ \phi_2(x) &= \lambda_2 (x^2 - \tfrac{1}{12}(n^2 - 1)) \\ \phi_3(x) &= \lambda_3 (x^3 - \tfrac{1}{20}(3n^2 - 7)x) \\ &\vdots \end{aligned}$$

donde las λ_j se eligen de forma que los valores de $\phi_j(x_i)$ sean enteros. Estos polinomios se encuentran tabulados para varios valores de n.

8.6.2. Elección del grado

Un aspecto importante de la regresión polinómica es la elección del grado m adecuado. El contraste de hipótesis

$$\begin{aligned} H_0 &: m = m_0 \\ H_1 &: m = m_1 > m_0 \end{aligned} \tag{8.10}$$

equivale a plantear una regresión polinómica de grado m y entonces establecer la hipótesis lineal

$$H_0 : \beta_{m_0+1} = \cdots = \beta_{m_1} = 0$$

sobre el modelo 8.6, o bien, utilizando el modelo equivalente 8.7 en términos de polinomios ortogonales

$$H_0 : \gamma_{m_0+1} = \cdots = \gamma_{m_1} = 0$$

Las sumas de cuadrados residuales son

$$\text{SCR} = Q(m_1) \qquad \text{SCR}_H = Q(m_0)$$

Teniendo en cuenta 8.9 resulta

$$\text{SCR}_H - \text{SCR} = Q(m_0) - Q(m_1) = \sum_{j=m_0+1}^{m_1} (\sum_{i=1}^{n} \phi_j^2(x_i))\hat{\gamma}_j^2$$

Entonces, para contrastar $H_0 : m = m_0$ frente $H_1 : m = m_1$, calcularemos el estadístico

$$F = \frac{(Q(m_0) - Q(m_1))/(m_1 - m_0)}{Q(m_1)/(n - m_1 - 1)} \tag{8.11}$$

cuya distribución, bajo H_0, es una F con $m_1 - m_0$ y $n - m_1 - 1$ grados de libertad.

La estrategia para elegir el grado puede ser mediante elección descendente o elección ascendente. En el primer caso empezamos por el grado que se supone máximo. Supongamos, por ejemplo, que $m = 5$. Entonces se contrasta $m = 4$ frente a $m = 5$. Si el test F no es significativo, se contrasta $m = 3$ con $m = 4$, y así sucesivamente. El proceso es el inverso en el caso de elección ascendente.

También es útil tener en cuenta que un descenso importante de la suma de cuadrados residual $Q(m)$ al pasar de grado k a grado m, es un indicio de que el grado es m.

Finalmente, si disponemos de n_i observaciones $y_{i1}, \ldots, y_{in_i}$ para cada valor de la variable de control x_i $i = 1, \ldots, p$, una vez elegido el grado m, podemos analizar la validez del modelo planteando el contraste

$$H_0 : y_{ih} = P_m(x_i) + \epsilon_{ih}$$
$$H_1 : y_{ih} = g(x_i) + \epsilon_{ih}$$

donde $g(x)$ es una función desconocida de x. La hipótesis nula significa afirmar que $g(x) = P_m(x)$ es un polinomio de grado m en x. Tenemos entonces (véase 6.12):

$$\text{SCR} = \sum_{i,h} (y_{ih} - \overline{y}_i)^2 = n s_y^2 (1 - \hat{\eta}^2) \qquad n - p \quad \text{g.l.}$$
$$\text{SCR}_H = Q(m) = n s_y^2 (1 - r_{y\mathbf{x}}^2) \qquad n - m - 1 \quad \text{g.l.}$$

donde $r_{y\mathbf{x}}$ es la correlación múltiple de Y sobre $x, x^2, \ldots, x^m$ (ver teorema 8.2.1). Calcularemos entonces el estadístico

$$F = \frac{(\hat{\eta}^2 - r_{y\mathbf{x}}^2)/(p - m - 1)}{(1 - \hat{\eta}^2)/(n - p)}$$

y aceptaremos el ajuste polinómico de grado m si esta F no es significativa.

Ejemplo 8.6.1

Se dispone de la respuesta a un test de conducta de dos grupos de ratas, uno control y otro experimental, para diez observaciones realizadas cada tres días desde el día 47 al día 74 de vida (ver tabla 8.4).

día	grupo control	grupo experimental
47	25.7	34.1
50	20.1	24.9
53	16.2	21.2
56	14.0	23.3
59	21.3	22.0
62	20.3	30.9
65	28.4	31.4
68	23.5	26.5
71	16.8	23.0
74	9.9	17.2

Tabla 8.4: Datos del test de conducta a dos grupos de ratas

El modelo considerado hace depender la variable conducta (medida mediante el test) del tiempo t según una función polinómica

$$\textit{var. obs.} = \textit{polinomio de grado m en t} + \textit{error} \qquad \Leftrightarrow \qquad y = P_m(t) + \epsilon$$

Para determinar el grado del polinomio al cual se ajustan los valores experimentales se plantea la hipótesis 8.10 que se resuelve mediante el test F 8.11.

Los resultados, obtenidos según el método de los polinomios ortogonales, son los siguientes

grupo control	g.l.	grupo experimental	g.l.
$Q(0) = 273.87$	9	$Q(0) = 249.99$	9
$Q(1) = 249.22$	8	$Q(1) = 216.12$	8
$Q(2) = 233.52$	7	$Q(2) = 213.15$	7
$Q(3) = 41.61$	6	$Q(3) = 37.80$	6
$Q(4) = 41.52$	5	$Q(4) = 27.10$	5

Observemos que hay un fuerte descenso de la suma de cuadrados residual $Q(m)$ al pasar de grado 2 a grado 3, indicio de que los datos experimentales se ajustan a un polinomio de grado 3.

Las F obtenidas son:

contraste	grupo control	grupo experimental
0 v.s. 1	$F = 0.79$ *(n.s.)*	$F = 1.25$ *(n.s.)*
0 v.s. 2	$F = 0.60$ *(n.s.)*	$F = 0.60$ *(n.s.)*
0 v.s. 3	$F = 11.16$ $(p < 0.01)$	$F = 11.23$ $(p < 0.01)$
1 v.s. 3	$F = 14.97$ $(p < 0.01)$	$F = 14.25$ $(p < 0.01)$
2 v.s. 3	$F = 27.67$ $(p < 0.01)$	$F = 27.83$ $(p < 0.01)$
3 v.s. 4	$F = 0.01$ *(n.s.)*	$F = 1.98$ *(n.s.)*

Efectivamente, tanto los datos del grupo control como los del grupo experimental se ajustan a un polinomio de grado 3 (ver Figura 8.2).

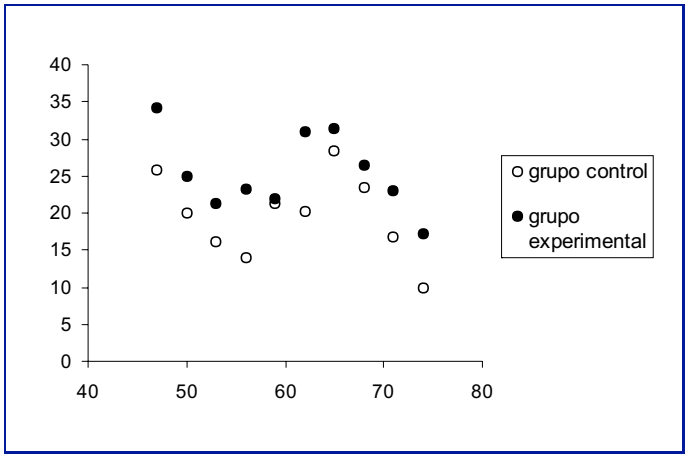

Figura 8.2: Gráfico de los dos grupos de ratas

El modelo es:

grupo control (∘)

$$y_i = 1929.24 - 97.86t_i + 1.654t_i^2 - 0.0092t_i^3 + \epsilon_i$$

grupo experimental (•)

$$y_i = 1892.28 - 94.94t_i + 1.593t_i^2 - 0.0088t_i^3 + \epsilon_i$$

8.7. Comparación de curvas experimentales

8.7.1. Comparación global

Si dos curvas experimentales se ajustan bien a modelos de formulación matemática diferente (por ejemplo, dos polinomios de distinto grado) hay que aceptar que las curvas experimentales son distintas.

Si las dos curvas son polinomios del mismo grado

$$\begin{aligned} y_1 &= P_m(x) + \epsilon \\ y_2 &= \overline{P}_m(x) + \epsilon \end{aligned}$$

la comparación se expresa planteando el siguiente contraste de hipótesis

$$\begin{aligned} H_0 &: P_m(x) = \overline{P}_m(x) \\ H_1 &: P_m(x) \neq \overline{P}_m(x) \end{aligned} \tag{8.12}$$

que implica la hipótesis lineal

$$H_0 : \beta_i = \overline{\beta}_i \qquad i = 0, 1, \ldots, m$$

análoga a

$$H_0 : \gamma_i = \overline{\gamma}_i \qquad i = 0, 1, \ldots, m \tag{8.13}$$

si utilizamos el modelo planteado mediante polinomios ortogonales (ver 8.7).

Sean $\mathrm{SCR}_1 = Q_1(m)$, $\mathrm{SCR}_2 = Q_2(m)$ las sumas de cuadrados residuales para cada curva y $\mathrm{SCR} = \mathrm{SCR}_1 + \mathrm{SCR}_2$ la suma de cuadrados residual del modelo conjunto construido mediante la unión de los dos modelos.

La construcción del modelo conjunto es sólo posible si los dos modelos poseen varianzas iguales. Por este motivo, es necesario plantear previamente el test de homogeneidad de varianzas

$$\begin{aligned} H_0 &: \sigma_1^2 = \sigma_2^2 \\ H_1 &: \sigma_1^2 \neq \sigma_2^2 \end{aligned}$$

que se resuelve mediante el estadístico

$$F = \frac{\mathrm{SCR}_1/(n_1 - m - 1)}{\mathrm{SCR}_2/(n_2 - m - 1)} \tag{8.14}$$

cuya distribución si H_0 es cierta es una F con $n_1 - m - 1$ y $n_2 - m - 1$ g.l..

Si aceptamos la igualdad de varianzas, podemos resolver 8.13 mediante el estadístico

$$F = \frac{(\mathrm{SCR}_H - \mathrm{SCR}_1 - \mathrm{SCR}_2)/(m+1)}{(\mathrm{SCR}_1 + \mathrm{SCR}_2)/(n_1 + n_2 - 2m - 2)} \tag{8.15}$$

que bajo H_0 sigue una F con $m + 1$ y $n_1 + n_2 - 2m - 2$ g.l.. La suma de cuadrados $\mathrm{SCR}_H = Q_{12}(m)$ es la suma de cuadrados residual bajo H_0, es decir, considerando que las dos curvas son iguales y que en consecuencia todos los datos se ajustan a un mismo polinomio de grado m.

8.7.2. Test de paralelismo

La hipótesis lineal de que las curvas son paralelas se plantea de la siguiente forma

$$H_0 : \beta_i = \overline{\beta}_i \qquad i = 1, \ldots, m$$

o bien, si nos referimos a 8.7

$$H_0 : \gamma_i = \overline{\gamma}_i \qquad i = 1, \ldots, m \tag{8.16}$$

Es decir, las curvas difieren únicamente respecto a la ordenada en el origen.

Esta hipótesis tiene generalmente interés cuando se rechaza H_0 de 8.12. Se resuelve mediante el estadístico

$$F = \frac{(\mathrm{SCR}_H^* - \mathrm{SCR}_1 - \mathrm{SCR}_2)/m}{(\mathrm{SCR}_1 + \mathrm{SCR}_2)/(n_1 + n_2 - 2m - 2)} \tag{8.17}$$

cuya distribución sigue una F con m y $n_1 + n_2 - 2m - 2$ g.l. cuando H_0 es cierta. La suma de cuadrados SCR_H^* es la suma de cuadrados residual bajo H_0 que supone aceptar la existencia de dos curvas distintas pero paralelas.

Ejemplo 8.7.1

En el ejemplo 8.6.1 hemos ajustado los datos del grupo control y del grupo experimental a dos polinomios de grado 3.

¿Podemos aceptar que en realidad los dos polinomios son iguales? Esta pregunta equivale a plantear la hipótesis lineal 8.13. Para resolverla es necesario realizar previamente el test de homogeneidad de varianzas utilizando 8.14

$$F = \frac{41.61/(10-3-1)}{37.80/(10-3-1)} = 1.10$$

con 6 y 6 g.l. (no significativa).

Pasamos pues a contrastar 8.13 mediante el estadístico 8.15. La suma de cuadrados residual bajo H_0 es $\mathrm{SCR}_H = Q_{12}(3) = 249.06$

$$F = \frac{(249.06 - 41.61 - 37.80)/(3+1)}{(41.61+37.80)/(10+10-6-2)} = 6.41$$

con 4 y 12 g.l. que es significativa ($p < 0.01$). Debemos aceptar en consecuencia que las dos curvas son diferentes (la conducta de los individuos del grupo control es diferente de la conducta de los individuos del grupo experimental).

No obstante, podemos preguntarnos si las dos curvas son paralelas y plantear la hipótesis lineal 8.16 que resolveremos utilizando el estadístico 8.17. La suma de cuadrados residual bajo H_0 es ahora $\mathrm{SCR}^*_H = Q^*_{12} = 82.59$

$$F = \frac{(82.59 - 41.61 - 37.80)/3}{(41.61+37.80)/(10+10-6-2)} = 0.16$$

con 3 y 12 g.l. (no significativa). Podemos entonces aceptar que las dos curvas experimentales son paralelas. La interpretación en términos de la conducta podría realizarse conociendo con más precisión el planteamiento del problema.

8.8. Ejemplos con R

Vamos a utilizar los datos del ejemplo 8.4.1 sobre el lenguaje. Las siguientes instrucciones permiten introducir los datos y dibujar los diagramas de dispersión dos a dos de las variables del ejemplo (ver figura 8.3).

```
> y<-c(3,2,4,9,6,7,2,6,5,8)
> x1<-c(1,1,3,7,8,7,4,6,6,9)
> x2<-c(3,4,7,9,7,6,5,8,5,7)
> exp<-cbind(x1,x2)
> lenguaje.datos<-data.frame(y,exp)
> par(pty="s")
> pairs(lenguaje.datos)
```

El siguiente paso es calcular el modelo de regresión lineal múltiple que permita predecir los valores de Y en función de las variables explicativas x_1 y x_2.

```
> regrem<-lm(y~x1+x2)
> summary(regrem)

Call: lm(formula = y ~ x1 + x2)
Residuals:
    Min      1Q   Median      3Q  Max
 -2.051 -0.5264 -0.05257 0.7989 1.47

Coefficients:
             Value Std. Error t value Pr(>|t|)
```

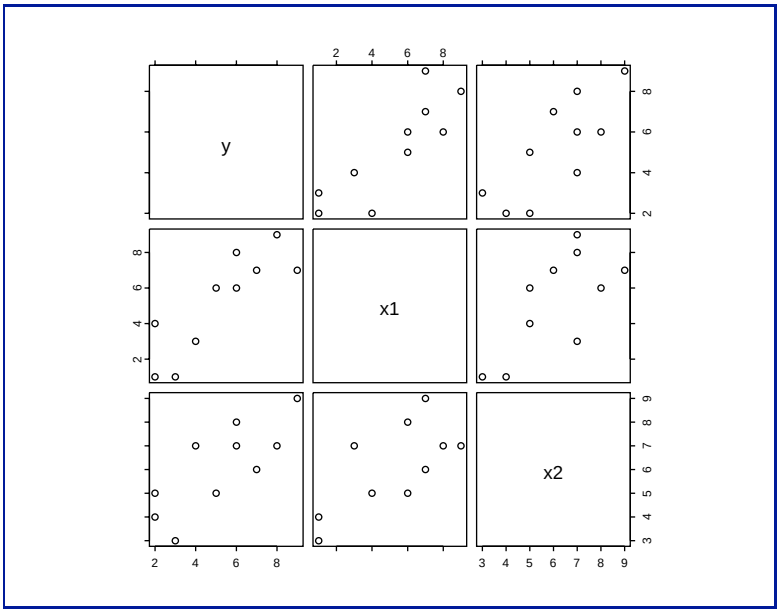

Figura 8.3: Diagramas de dispersión dos a dos entre la variable respuesta y las variables explicativas del ejemplo 8.4.1

```
(Intercept) -0.4244   1.4701    -0.2887   0.7812
         x1  0.5123   0.2087     2.4543   0.0438
         x2  0.4853   0.3178     1.5273   0.1705

Residual standard error: 1.266 on 7 degrees of freedom
Multiple R-Squared: 0.7907
F-statistic: 13.22 on 2 and 7 degrees of freedom, the p-value is 0.004196

Correlation of Coefficients:
   (Intercept)      x1
x1  0.1811
x2 -0.8036     -0.6973
```

El plano estimado es $\hat{y} = -0.4244 + 0.5123x_1 + 0.4853x_2$ con un coeficiente de determinación $R^2 = 0.7907$ y el estadístico F nos dice que el modelo es útil, si un estudio más profundo decide finalmente que es realmente válido.
Resulta curioso que en S-PLUS se puede obtener el coeficiente de determinación R^2 a partir de la función `summary.lm` en la forma

```
> summary(regrem)$r.squared
[1] 0.790684
```

pero no hay nombre para el coeficiente ajustado. Mientras que en R sí es posible.
También se pueden obtener los coeficientes a partir de la matriz $\mathbf{X}'\mathbf{X}$:

```
> XtX<-t(regrem$R)%*%regrem$R
> XtX
            (Intercept)  x1  x2
(Intercept)          10  52  61
         x1          52 342 350
         x2          61 350 403
> XtX.inv<-solve(XtX)
> XtX.inv
            (Intercept)          x1         x2
(Intercept)  1.34840753  0.03466479 -0.2342073
         x1  0.03466479  0.02718635 -0.0288580
```

```
         x2 -0.23420728 -0.02885800  0.0629949
> XtX.inv%*%t(cbind(1,exp))%*%y
                 [,1]
(Intercept) -0.4244237
         x1  0.5123174
         x2  0.4853071
```

La matriz `XtX.inv` se puede obtener de forma directa así:

```
> summary(regrem)$cov.unscaled
            (Intercept)          x1         x2
(Intercept)  1.34840753  0.03466479 -0.2342073
         x1  0.03466479  0.02718635 -0.0288580
         x2 -0.23420728 -0.02885800  0.0629949
```

También se obtiene más fácilmente con los elementos que proporciona la función `lsfit`:

```
> regrem.ls<-lsfit(exp,y)
> regrem.diag<-ls.diag(regre.ls)
> regrem.diag$cov.unscaled
```

La matriz $\hat{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1}$ de varianzas y covarianzas entre los estimadores MC de los coeficientes se obtiene de forma sencilla:

```
> summary(regrem)$sigma^2*summary(regrem)$cov.unscaled
            (Intercept)          x1          x2
(Intercept)  2.16117719  0.05555943 -0.37537868
         x1  0.05555943  0.04357326 -0.04625252
         x2 -0.37537868 -0.04625252  0.10096587
```

o también

```
> regrem.diag$std.dev^2*regrem.diag$cov.unscaled
```

Para calcular intervalos de confianza sobre los coeficientes de regresión hacemos

```
> beta.est<-cbind(regrem.ls$coef);beta.est
                [,1]
Intercept -0.4244237
       x1  0.5123174
       x2  0.4853071
> cbind(beta.est+qt(0.025,7)*regrem.diag$std.err,
+ beta.est+qt(0.975,7)*regrem.diag$std.err)
                   [,1]     [,2]
(Intercept) -3.90064431 3.051797
         x1  0.01872084 1.005914
         x2 -0.26605529 1.236669
```

Observamos que los intervalos correspondientes a β_0 y β_2 contienen al cero, en coherencia con los test t parciales. Pero también nos puede interesar reproducir la tabla ANOVA sobre la significación de la regresión, aunque el test F ya se ha obtenido con la función `summary(regrem)`. Las funciones `anova.lm` o `summary.aov` nos pueden ayudar.

```
> summary.aov(regrem)
           Df Sum of Sq  Mean Sq  F Value     Pr(F)
       x1   1  38.64190 38.64190 24.10956 0.0017330
       x2   1   3.73876  3.73876  2.33270 0.1705213
Residuals   7  11.21934  1.60276
```

Sin embargo, los resultados se refieren a contrastes F secuenciales y parciales. Exactamente $SC_R(\beta_0,\beta_1) = 38.64190$ y $SC_R(\beta_2|\beta_0,\beta_1) = 3.73876$, de manera que

$$SC_R = SC_R(\beta_1,\beta_0) + SC_R(\beta_2|\beta_0,\beta_1) = 42.38066$$

Por otra parte, se observa directamente que SCR = 11.21934. Con estos datos, completar la tabla 8.1 es relativamente sencillo. Sin embargo se puede conseguir dicha tabla, aunque con otra organización, mediante un contraste de modelos:

```
> regrem0<-lm(y~1)
> anova(regrem0,regrem)
Analysis of Variance Table

Response: y

    Terms Resid. Df      RSS Test Df Sum of Sq  F Value      Pr(F)
1       1         9 53.60000
2 x1 + x2         7 11.21934       2  42.38066 13.22113 0.00419574
```

Otro aspecto que también hemos visto ha sido el cálculo de los coeficientes de regresión estandarizados, que con R se obtienen así:

```
> cor(exp)
          x1        x2
x1 1.0000000 0.6973296
x2 0.6973296 1.0000000
> cor(exp,y)
        [,1]
x1 0.8490765
x2 0.7813857
> solve(cor(exp),cor(exp,y))
        [,1]
x1 0.5921248
x2 0.3684796
```

Si queremos más detalles sobre los coeficientes de regresión estandarizados, podemos utilizar el siguiente modelo sin coeficiente de intercepción:

```
> x1.est<-(x1-mean(x1))/stdev(x1)
> x2.est<-(x2-mean(x2))/stdev(x2)
> y.est<-(y-mean(y))/stdev(y)
> regrem.est<-lm(y.est~-1+x1.est+x2.est)
> summary(regrem.est)
```

Por último, podemos estudiar la multicolinealidad calculando los FIV

```
> diag(solve(cor(exp)))
[1] 1.946542 1.946542
```

que en este caso no existe.

El cálculo de predicciones puntuales o por intervalo se obtiene mediante la función `predict.lm` del modelo lineal.

8.9. Ejercicios

Ejercicio 8.1

Consideremos el modelo de la regresión lineal múltiple

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_m x_{im} \qquad i = 1, \ldots, n$$

Sean $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_m$ las estimaciones MC de los parámetros. Explicar en qué condiciones podemos afirmar que $E(\hat{\beta}_j) = \beta_j$, $j = 0, 1, \ldots, m$.

Por otra parte, ¿es siempre válido afirmar que

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_m x_{im}$$

es una estimación centrada de

$$\beta_0 + \beta_1 x_{i1} + \cdots + \beta_m x_{im} \ ?$$

Ejercicio 8.2

En la regresión múltiple de una variable Y sobre tres variables control x_1, x_2, x_3

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \epsilon_i \qquad i = 1, \ldots, n$$

donde $\epsilon_i \sim N(0, \sigma^2)$, se desea contrastar la hipótesis nula

$$H_0 : \beta_2 = \beta_3 = 0$$

Sea $r_{y\mathbf{x}}$ el coeficiente de correlación múltiple de Y sobre x_1, x_2, x_3 y sea r_{y1} el coeficiente de correlación simple entre Y y x_1. Deducir un test F para contrastar H_0 que sea función de $r_{y\mathbf{x}}$ y r_{y1}.

Ejercicio 8.3

En una gran ciudad, queremos relacionar el número de muertos diarios por enfermedades cardio-respiratorias con la media de humos (mg/m^3) i la media de dióxido de azufre (partes/millón) medidas por los equipos del Ayuntamiento en diversas zonas de la ciudad.

Consideremos un modelo de regresión lineal no centrado con los siguientes datos:

$$\mathbf{X'X} = \begin{pmatrix} 15 & 6.87 & 21.09 \\ & 5.6569 & 18.7243 \\ & & 63.2157 \end{pmatrix} \quad (\mathbf{X'X})^{-1} = \begin{pmatrix} 0.2243 & -1.2611 & 0.2987 \\ & 16.1158 & -4.3527 \\ & & 1.2054 \end{pmatrix}$$

$$\mathbf{X'Y} = \begin{pmatrix} 3922 \\ 2439.54 \\ 7654.35 \end{pmatrix} \qquad \mathbf{Y'Y} = 1264224$$

Se pide:

1) Calcular la estimación MC de todos los coeficientes de regresión del modelo.
2) Obtener una estimación insesgada de la varianza del modelo.
3) Contrastar la significación del modelo propuesto con $\alpha = 0.1$.
4) Calcular el intervalo de confianza al 95 % para la media del valor respuesta para una media de humos de 1 mg/m^3 y una media de SO_2 de 1.

Ejercicio 8.4

Se dispone de los siguientes datos sobre diez empresas fabricantes de productos de limpieza doméstica:

Empresa	V	IP	PU
1	60	100	1.8
2	48	110	2.4
3	42	130	3.6
4	36	100	0.6
5	78	80	1.8
6	36	80	0.6
7	72	90	3.6
8	42	120	1.2
9	54	120	2.4
10	90	90	4.2

En el cuadro anterior, V son las ventas anuales, expresadas en millones de euros, IP es un índice de precios relativos (Precios de la empresa/Precios de la competencia) y PU son los gastos anuales realizados en publicidad y campañas de promoción y difusión, expresados también en millones de euros.

Tomando como base la anterior información:

1) Estimar el vector de coeficientes $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)'$ del modelo

$$V_i = \beta_0 + \beta_1 IP_i + \beta_2 PU_i + \epsilon_i$$

2) Estimar la matriz de varianzas-covarianzas del vector $\widehat{\boldsymbol{\beta}}$.

3) Calcular el coeficiente de determinación.

Ejercicio 8.5

Dado el modelo

$$Y_t = \beta_0 + \beta_1 X_{1t} + \beta_2 X_{2t} + u_t$$

y los siguientes datos

Y_t	X_{1t}	X_{2t}
10	1	0
25	3	−1
32	4	0
43	5	1
58	7	−1
62	8	0
67	10	−1
71	10	2

obtener:

(a) La estimación MC de $\beta_0, \beta_1, \beta_2$ utilizando los valores originales.

(b) La estimación MC de $\beta_0, \beta_1, \beta_2$ utilizando los datos expresados en desviaciones respecto a la media.

(c) La estimación insesgada de σ^2.

(d) El coeficiente de determinación.

(e) El coeficiente de determinación corregido.

(f) El contraste de la hipótesis nula $H_0 : \beta_0 = \beta_1 = \beta_2 = 0$.

(g) El contraste de la hipótesis nula $H_0 : \beta_1 = \beta_2 = 0$ utilizando datos originales.

(h) El contraste de la hipótesis nula $H_0 : \beta_1 = \beta_2 = 0$ utilizando datos en desviaciones respecto a la media.

(i) La representación gráfica de una región de confianza del 95 % para β_1 y β_2.

(j) El contraste individual de los parámetros β_0, β_1 y β_2.

(k) El contraste de la hipótesis nula $H_0 : \beta_1 = 10\beta_2$.

(l) El contraste de la hipótesis nula $H_0 : 2\beta_0 + 2\beta_1 + 7\beta_2 = 50$.

(m) El contraste de la hipótesis nula conjunta $H_0 : \beta_1 = 10\beta_2, 2\beta_0 + 2\beta_1 + 7\beta_2 = 50$.

Ejercicio 8.6

Supongamos que hemos estimado la siguiente ecuación utilizando MC (con las variables medidas en logaritmos)

$$Y_t = \beta_0 + \beta_1 X_{1t} + \beta_2 X_{2t} \qquad t = 1, \ldots, 17$$

y las estimaciones de los parámetros son:

$$\hat{\beta}_0 = 1.37 \qquad \hat{\beta}_1 = 1.14 \qquad \hat{\beta}_2 = -0.83$$

También hemos obtenido la siguiente expresión escalar:

$$\mathbf{Y}'[\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']\mathbf{Y} = 0.0028$$

y los elementos triangulares de la matriz $(\mathbf{X}'\mathbf{X})^{-1}$ son:

$$\begin{pmatrix} 510.89 & -254.35 & 0.42 \\ & 132.70 & -6.82 \\ & & 7.11 \end{pmatrix}$$

Se pide:

1. Calcular las varianzas de los estimadores MC de $\beta_0, \beta_1, \beta_2$.
2. Si X_{1t} aumenta en un 1 por 100 y X_{2t} en un 2 por 100, ¿cuál sería el efecto estimado en Y_t?
3. Efectuar un test estadístico para verificar la hipótesis de que $\beta_1 = 1$ y $\beta_2 = -1$ y dar el valor de dicho estadístico. ¿Cuáles son las tablas que necesitaremos para realizar el test y cuántos son los grados de libertad?

Ejercicio 8.7

Una variable Y depende de otra variable control x que toma los valores $x_1 = 1, x_2 = 2, x_3 = 3, x_4 = 4$ de acuerdo con el modelo lineal normal

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \epsilon_i \qquad i = 1, 2, 3, 4$$

Estudiar la expresión del estadístico F para contrastar la hipótesis $H_0 : \beta_1 = \beta_2$.

Ejercicio 8.8

La puntuación del test *open-field* para un grupo de 10 ratas control (C) y otro grupo de 10 ratas experimentales (E) a lo largo de los días $47, 50, \ldots, 74$ contados desde el instante del nacimiento fue

Día	47	50	53	56	59	62	65	68	71	74
grupo C	34	24	21	23	23	30	31	26	23	17
grupo E	25	20	16	15	21	20	28	23	18	9

Se ajustaron al grupo control polinomios de grado $0, 1, 2$ y 3 respecto la variable "edad en días" y se obtuvieron las siguientes sumas de cuadrados residuales:

$$\begin{aligned} Q(0) &= 235.6 \\ Q(1) &= 202.8 \\ Q(2) &= 199.4 \\ Q(3) &= \ 29.7 \end{aligned}$$

Se pide:

(a) Comprobar que se puede aceptar como válido el polinomio de grado 3 como polinomio de regresión de Y (puntuación) sobre x (edad en días).

(b) El polinomio de grado 3 que ajusta Y a x es

$$y = 318.8 - 93.3x + 1.56x^2 - 0.0086x^3$$

El coeficiente de correlación múltiple de Y sobre x, x^2, x^3 es $r_{y\mathbf{x}} = 0.8734$. Estudiar si es significativo.

(c) Para el grupo experimental es también adecuado un ajuste polinómico de grado 3 con suma de cuadrados residual $Q(3) = 29.2$. Además, juntando todos los datos referentes a Y, es decir, juntando los dos grupos y en consecuencia las 20 observaciones y realizando un ajuste polinómico de grado 3, se obtiene

$$\mathrm{SCR}_H = 225.8$$

Contrastar las hipótesis

$$\begin{aligned} H_0 &: \text{los dos polinomios (C y E) son idénticos} \\ H_1 &: \text{hay diferencias significativas entre ambos polinomios} \end{aligned}$$

Ejercicio 8.9

En función de las variables

MaxAT, MinAT: Máxima y mínima temperatura del aire.

MaxST, MinST: Máxima y mínima temperatura del suelo.

MAT, MST: Temperaturas medias del aire y del suelo.

MaxH, MinH: Máxima y mínima humedades relativas.

MH: Humedad media durante el día.

V: Fuerza del viento.

se dispone de la siguiente información sobre la evaporación:

Fuente	SS	df	MS	F
Regresión	8159.83	10	815.98	19.27
Error	1482.27	35	42.35	
Total	9642.11	45		

Variable	$\hat{\beta}$	t	VIF
MaxAT	0.5011	0.88	8.83
MinAT	0.3041	0.39	8.89
MAT	0.0930	0.42	21.21
MaxST	2.2320	2.22	39.29
MinST	0.2050	0.19	14.08
MST	−0.7400	−2.12	52.36
MaxH	1.1100	0.98	1.98
MinH	0.7510	1.54	25.38
MH	−0.5560	−3.44	24.12
V	0.0089	0.97	1.98

Se pide:

(a) Calcular R^2 y verificar la significación de la regresión con $\alpha = 0.01$.

(b) Calcular R_j^2 para cada variable explicativa.

(c) ¿Hay variables implicadas en una posible multicolinealidad?

(d) ¿Qué pasa con los coeficientes negativos? Discutir los resultados obtenidos.