Estadística II, grado en Matemáticas
Resumen del modelo de regresión lineal múltiple
Curso 2019-2020

1. Ingredientes y datos

- $k \geq 1$ variables explicativas $X_1, \ldots, X_k$.
- Variable respuesta Y.
- Serie de n datos (cada uno de longitud $k+1$).
- $n \geq k+2$.
- No colinealidad (columnas X_1 a X_k).

X_1	X_2	$\cdots$	X_k	Y
$x_{1,1}$	$x_{1,2}$	$\cdots$	$x_{1,k}$	y_1
$x_{2,1}$	$x_{2,2}$	$\cdots$	$x_{2,k}$	y_2
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
$x_{n,1}$	$x_{n,2}$	$\cdots$	$x_{n,k}$	y_n

2. Modelo

El vector $\mathbf{y} = (y_1, \ldots, y_n)^\intercal$ es una realización del vector aleatorio $\mathbb{Y} = (Y_1, \ldots, Y_n)^\intercal$ dado por

$$\mathbb{Y} = X \cdot \boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

donde

$$X = \begin{pmatrix} 1 & x_{1,1} & x_{1,2} & \cdots & x_{1,k} \\ 1 & x_{2,1} & x_{2,2} & \cdots & x_{2,k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1,1} & x_{n,2} & \cdots & x_{n,k} \end{pmatrix}$$

es la matriz de diseño (de rango $k+1$), y donde

- $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_k)^\intercal$ es el vector de parámetros,
- $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^\intercal \sim \mathcal{N}(\mathbf{0}, \sigma^2 I_n)$, donde σ^2 es otro parámetro e I_n es la matriz identidad $n \times n$. Es decir, las variables ε_i son normales independientes de media 0 y varianza σ^2.

El vector $\mathbb{Y}$ se distribuye como una $\mathcal{N}(X \cdot \boldsymbol{\beta}, \sigma^2\, I_n)$.

3. Estimación de parámetros

a) Dada la muestra $\mathbf{y} = (y_1, \ldots, y_n)^\intercal$, la estimación (mínimo error cuadrático/máxima verosimilitud) de los parámetros es

$$\widehat{\boldsymbol{\beta}} = (X^\intercal X)^{-1} X^\intercal \mathbf{y}.$$

Para el caso de la regresión lineal simple ($k=1$), llamando $\mathbf{x} = (x_1, \ldots, x_n)^\intercal$ a la (única) columna de observaciones,

$$\widehat{\beta}_1 = \frac{\mathrm{cov}_{\mathbf{x},\mathbf{y}}}{V_{\mathbf{x}}}, \qquad \widehat{\beta}_0 = \overline{y} - \frac{\mathrm{cov}_{\mathbf{x},\mathbf{y}}}{V_{\mathbf{x}}}\, \overline{x},$$

donde

$$\overline{x} = \frac{1}{n}\sum_{i=1}^n x_i, \quad \overline{y} = \frac{1}{n}\sum_{i=1}^n y_i, \quad V_{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^n (x_i - \overline{x})^2, \quad \mathrm{cov}_{\mathbf{x},\mathbf{y}} = \frac{1}{n}\sum_{i=1}^n (x_i - \overline{x})(y_i - \overline{y}).$$

b) Valores pronosticados y residuos. Dada la muestra $\mathbf{y} = (y_1, \ldots, y_n)^\intercal$, los pronósticos $\widehat{\mathbf{y}} = (\widehat{y}_1, \ldots \widehat{y}_n)^\intercal$ y los residuos $\mathbf{e} = (e_1, \ldots, e_n)^\intercal$ son

$$\widehat{\mathbf{y}} = X\widehat{\boldsymbol{\beta}} = X(X^\intercal X)^{-1} X^\intercal \mathbf{y} := H\mathbf{y},$$
$$\mathbf{e} = \mathbf{y} - \widehat{\mathbf{y}} = (I_n - H)\mathbf{y}.$$

La matriz H es $n \times n$, simétrica, definida positiva e idempotente de rango $k+1$.

c) Sumas de cuadrados: $\text{TSS} = \text{MSS} + \text{RSS}$, con

$$\begin{aligned}
&\text{(total)} && \text{TSS} = \sum_{i=1}^{n}(y_i - \overline{y})^2 = nV_{\mathbf{y}} = \mathbf{y}^{\intercal}(I_n - \tfrac{1}{n}\,J_n)\,\mathbf{y},\\
&\text{(explicada por modelo)} && \text{MSS} = \sum_{i=1}^{n}(\widehat{y}_i - \overline{y})^2 = \mathbf{y}^{\intercal}(H - \tfrac{1}{n}\,J_n)\,\mathbf{y},\\
&\text{(residual)} && \text{RSS} = \sum_{i=1}^{n}(y_i - \widehat{y}_i)^2 = \sum_{i=1}^{n} e_i^2 = \mathbf{y}^{\intercal}(I_n - H)\,\mathbf{y},
\end{aligned}$$

donde J_n denota la matriz $n \times n$ con unos.

d) Estimación para σ^2:

$$\widehat{\sigma^2} = s_R^2 = \frac{1}{n-k-1}\,\sum_{i=1}^{n} e_i^2 = \frac{\text{RSS}}{n-k-1}\,.$$

e) Coeficiente R^2:

$$R^2 = \frac{\text{MSS}}{\text{TSS}} = 1 - \frac{\text{RSS}}{\text{TSS}}.$$

Obsérvese que $\text{MSS}/\text{RSS} = R^2/(1-R^2)$.

4. Distribución de estimadores

Consideramos los estimadores (estadísticos asociados a $\mathbb{Y} = (Y_1, \dots, Y_n)^{\intercal}$)

$$\widehat{\boldsymbol{\beta}} = (X^{\intercal}X)^{-1}X^{\intercal}\,\mathbb{Y} \qquad \text{y} \qquad s_R^2 = \frac{1}{n-k-1}\,\mathbb{Y}^{\intercal}(I_n - H)\,\mathbb{Y}.$$

En el caso $k = 1$,

$$\widehat{\beta}_1 = \frac{1}{V_{\mathbf{x}}}\,\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})(Y_i - \overline{Y}), \qquad \widehat{\beta}_0 = \overline{Y} - \frac{\overline{x}}{V_{\mathbf{x}}}\,\frac{1}{n}\,\sum_{i=1}^{n}(x_i - \overline{x})(Y_i - \overline{Y}),$$

donde $\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$.

Se tiene que

- $\widehat{\boldsymbol{\beta}} \sim \mathcal{N}(\boldsymbol{\beta}, \sigma^2(X^TX)^{-1})$,
- $(n-k-1)s_R^2/\sigma^2 \sim \chi^2_{n-k-1}$,
- y s_R^2 es independiente de $\widehat{\boldsymbol{\beta}}$.

En particular, para $j = 0, \dots, k$, y llamando $q_{j,j}$ al elemento j de la diagonal de $(X^TX)^{-1}$,

$$\frac{\widehat{\beta}_j - \beta_j}{s_R\sqrt{q_{j+1,j+1}}} \sim t_{n-k-1}.$$

En el caso $k = 1$,

$$\mathbf{V}(\widehat{\beta}_0) = \sigma^2\left[\frac{1}{n} + \frac{\overline{x}^2}{nV_{\mathbf{x}}}\right], \quad \mathbf{V}(\widehat{\beta}_1) = \sigma^2\,\frac{1}{nV_{\mathbf{x}}}, \quad \text{cov}(\widehat{\beta}_0, \widehat{\beta}_1) = -\sigma^2\,\frac{\overline{x}}{nV_{\mathbf{x}}}.$$

5. Intervalos de confianza para los parámetros

Dado α, y para $j = 0, \dots, k$,

$$\mathrm{IC}_{1-\alpha}(\beta_j) = \widehat{\beta}_j \pm t_{\{n-k-1;\alpha/2\}}\, s_R \sqrt{q_{j+1,j+1}}.$$

Para el caso $k = 1$,

$$\mathrm{IC}_{1-\alpha}(\beta_0) = \widehat{\beta}_0 \pm t_{\{n-2;\alpha/2\}}\, s_R \sqrt{\frac{1}{n} + \frac{\overline{x}^2}{n\,V_{\mathbf{x}}}}, \quad \mathrm{IC}_{1-\alpha}(\beta_1) = \widehat{\beta}_1 \pm t_{\{n-2;\alpha/2\}}\, s_R \sqrt{\frac{1}{n\,V_{\mathbf{x}}}}.$$

Para σ^2,

$$\mathrm{IC}_{1-\alpha}(\sigma^2) = \Big(\frac{(n-k-1)\, s_R^2}{\chi^2_{\{n-k-1;\alpha/2\}}}, \frac{(n-k-1)\, s_R^2}{\chi^2_{\{n-k-1;1-\alpha/2\}}}\Big).$$

6. Contrastes de hipótesis

a) Hipótesis individuales $H_0 : \beta_j = 0$, con $j \in \{1, \dots, k\}$. Región de rechazo con nivel de significación α:

$$\mathcal{R}_j = \Big\{\Big|\frac{\widehat{\beta}_j}{s_R\sqrt{q_{j+1,j+1}}}\Big| > t_{\{n-k-1;\alpha/2\}}\Big\}.$$

b) Hipótesis global $H_0 : \beta_1 = \cdots = \beta_k = 0$. Bajo H_0, se tiene que

$$\frac{\textsc{mss}/k}{\textsc{rss}/(n-k-1)} \sim F_{k,n-k-1}.$$

Región de rechazo con nivel de significación α:

$$\mathcal{R} = \Big\{\frac{\textsc{mss}/k}{\textsc{rss}/(n-k-1)} > F_{\{k,n-k-1;\alpha\}}\Big\}.$$

Tabla ANOVA:

Fuente	suma cuadrados	g.l.	varianza	estadístico F
explicada por regresión	MSS	k	$\textsc{mss}/k$	$(\textsc{mss}/k)/s_R^2$
residual	RSS	$n-k-1$	$\textsc{rss}/(n-k-1) = s_R^2$	
total	TSS	$n-1$		

7. Predicciones

Condicionando sobre una observación $\mathbf{x}_0 = (x_{0,1}, \dots, x_{0,k})$, y si llamamos $\widetilde{\mathbf{x}}_0 = (1, x_{0,1}, \dots, x_{0,k})$, la predicción, tanto sobre la media de Y como sobre el valor de Y, es

$$\widehat{y}_0 = \widetilde{\mathbf{x}}_0^{\intercal} \cdot \widehat{\boldsymbol{\beta}}.$$

Intervalos de confianza:

$$\mathrm{IC}_{1-\alpha}(\text{media de } Y \,|\, \mathbf{x}_0) = \widehat{y}_0 \pm t_{\{n-k-1;\alpha/2\}} \cdot s_R \cdot \sqrt{\widetilde{\mathbf{x}}_0^{\intercal}\, (X^{\intercal}X)^{-1}\, \widetilde{\mathbf{x}}_0}$$

$$\mathrm{IC}_{1-\alpha}(\text{valor de } Y \,|\, \mathbf{x}_0) = \widehat{y}_0 \pm t_{\{n-k-1;\alpha/2\}} \cdot s_R \cdot \sqrt{1 + \widetilde{\mathbf{x}}_0^{\intercal}\, (X^{\intercal}X)^{-1}\, \widetilde{\mathbf{x}}_0}$$

En el caso $k = 1$, dada la observación x_0,

$$\mathrm{IC}_{1-\alpha}(\text{media de } Y \,|\, x_0) = \widehat{y}_0 \pm t_{\{n-2;\alpha/2\}} \cdot s_R \cdot \sqrt{\frac{1}{n} + \frac{(x_0 - \overline{x})^2}{n\,V_{\mathbf{x}}}},$$

$$\mathrm{IC}_{1-\alpha}(\text{valor de } Y \,|\, x_0) = \widehat{y}_0 \pm t_{\{n-2;\alpha/2\}} \cdot s_R \cdot \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{n\,V_{\mathbf{x}}}}.$$

Espacio para tus anotaciones adicionales.