MAT-266: Test de hipótesis II

Felipe Osorio

fosorios.mat.utfsm.cl

Departamento de Matemática, UTFSM

Ejemplo (Modelo de análisis de varianza):

Considere el modelo:

$$Y_{ij} = \mu_i + \epsilon_{ij}, \qquad i = 1, \dots, p; j = 1, \dots, n. \tag{1}$$

Se tiene interés en probar la hipótesis

$$\begin{aligned} H_0 :& \mu_1 = \mu_2 = \cdots = \mu_p, \\ H_1 :& \mu_r \neq \mu_s, \quad \text{para algún } r \neq s. \end{aligned}$$

Evidentemente, podemos escribir la hipótesis anterior en la forma lineal $H_0 : \boldsymbol{G\beta} = \boldsymbol{g}$, con

$$\boldsymbol{G} = \begin{pmatrix} 1 & 0 & \dots & 0 & -1 \\ 0 & 1 & \dots & 0 & -1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & 1 & -1 \end{pmatrix} = (\boldsymbol{I}_{p-1}, -\mathbf{1}), \qquad \boldsymbol{g} = \mathbf{0}$$

El modelo en (1) puede ser escrito en la forma:

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

donde

$$\boldsymbol{Y} = \begin{pmatrix} \boldsymbol{Y}_1 \\ \boldsymbol{Y}_2 \\ \vdots \\ \boldsymbol{Y}_p \end{pmatrix}, \quad \boldsymbol{X} = \begin{pmatrix} \mathbf{1} & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \mathbf{1} & \dots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \dots & \mathbf{1} \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{pmatrix}, \quad \boldsymbol{\epsilon} = \begin{pmatrix} \boldsymbol{\epsilon}_1 \\ \boldsymbol{\epsilon}_2 \\ \vdots \\ \boldsymbol{\epsilon}_p \end{pmatrix}$$

con $\boldsymbol{Y}_i = (Y_{i1}, \dots, Y_{in})^\top$. Sabemos que

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{Y} = (\overline{Y}_1, \overline{Y}_2, \dots, \overline{Y}_p)^\top,$$

y

$$Q(\widehat{\boldsymbol{\beta}}) = \boldsymbol{Y}^\top \boldsymbol{Y} - \boldsymbol{Y}^\top \boldsymbol{X} \widehat{\boldsymbol{\beta}} = \sum_{j=1}^{n} \sum_{i=1}^{p} (Y_{ij} - \overline{Y}_i)^2.$$

Ahora, bajo[1] $H_0 : \mu_1 = \cdots = \mu_p \; (= \mu)$, tenemos que el modelo (reducido) puede ser escrito como:

$$\boldsymbol{Y} = \mu \mathbf{1}_{np} + \boldsymbol{\epsilon}. \tag{2}$$

De este modo,

$$\widetilde{\mu} = \frac{1}{np} \mathbf{1}^\top \boldsymbol{Y}, \qquad \widetilde{\boldsymbol{\beta}} = (\overline{Y}, \overline{Y}, \dots, \overline{Y})^\top = \overline{Y} \mathbf{1}_p,$$

mientras que

$$\begin{aligned} Q(\widetilde{\boldsymbol{\beta}}) = Q_R(\widetilde{\mu}) = \boldsymbol{Y}^\top \boldsymbol{Y} - \boldsymbol{Y}^\top \mathbf{1} \widetilde{\mu} &= \sum_{j=1}^{n} \sum_{i=1}^{p} Y_{ij}^2 - np\overline{Y}^2 \\ &= \sum_{j=1}^{n} \sum_{i=1}^{p} (Y_{ij} - \overline{Y})^2 \end{aligned}$$

[1] Es decir, suponiendo que H_0 es verdadera.

Test de hipótesis: Ejemplos

En efecto, tenemos

$$\boldsymbol{G} = (\boldsymbol{G}_q, \boldsymbol{G}_r) = (\boldsymbol{I}_{p-1}, -\boldsymbol{1}), \qquad \boldsymbol{g} = \boldsymbol{0}, \qquad \boldsymbol{X} = (\boldsymbol{X}_q, \boldsymbol{X}_r),$$

donde

$$\boldsymbol{X}_q = \begin{pmatrix} \boldsymbol{1} & \boldsymbol{0} & \dots & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{1} & \dots & \boldsymbol{0} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{0} & \boldsymbol{0} & \dots & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \dots & \boldsymbol{1} \end{pmatrix}, \qquad \boldsymbol{X}_r = \begin{pmatrix} \boldsymbol{0} \\ \boldsymbol{0} \\ \vdots \\ \boldsymbol{0} \\ \boldsymbol{1} \end{pmatrix}$$

luego

$$\boldsymbol{Y}_R = \boldsymbol{Y} - \boldsymbol{X}_q \boldsymbol{G}_q^{-1} \boldsymbol{g} = \boldsymbol{Y}.$$

Test de hipótesis: Ejemplos

Mientras que

$$\boldsymbol{X}_q \boldsymbol{G}_q^{-1} \boldsymbol{G}_r = -\begin{pmatrix} \mathbf{1} & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \mathbf{1} & \dots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \dots & \mathbf{1} \end{pmatrix} \mathbf{1} = -\begin{pmatrix} \mathbf{1} \\ \mathbf{1} \\ \vdots \\ \mathbf{1} \\ \mathbf{0} \end{pmatrix}$$

De este modo

$$\boldsymbol{X}_R = \boldsymbol{X}_r - \boldsymbol{X}_q \boldsymbol{G}_q^{-1} \boldsymbol{G}_r = \begin{pmatrix} \mathbf{0} \\ \vdots \\ \mathbf{0} \\ \mathbf{1} \end{pmatrix} + \begin{pmatrix} \mathbf{1} \\ \vdots \\ \mathbf{1} \\ \mathbf{0} \end{pmatrix} = \mathbf{1}_{np},$$

lo que lleva al modelo (2) (como esperado).

De este modo,

$$Q(\widetilde{\boldsymbol{\beta}}) - Q(\widehat{\boldsymbol{\beta}}) = \sum_{j=1}^{n} \sum_{i=1}^{p} (Y_{ij} - \overline{Y})^2 - \sum_{j=1}^{n} \sum_{i=1}^{p} (Y_{ij} - \overline{Y}_i)^2,$$

y

$$s^2 = \frac{1}{np - p} \sum_{j=1}^{n} \sum_{i=1}^{p} (Y_{ij} - \overline{Y}_i)^2.$$

Finalmente, se rechaza $H_0 : \mu_1 = \mu_2 = \cdots = \mu_p$, si

$$\begin{aligned} F &= \frac{\sum_{j=1}^{n} \sum_{i=1}^{p} \{(Y_{ij} - \overline{Y})^2 - (Y_{ij} - \overline{Y}_i)^2\}/(p-1)}{s^2} \\ &\geq F_{1-\alpha}(p-1, (n-1)p). \end{aligned}$$

Ejemplo (Modelo de regresión lineal múltiple):

Considere el modelo:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + \epsilon_i, \qquad i = 1, \dots, n, \tag{3}$$

con los supuestos habituales, y suponga que se desea probar Se tiene interés en probar la hipótesis

$$\begin{aligned} H_0 :& \beta_1 = \beta_2 = \cdots = \beta_k = 0, \\ H_1 :& \beta_j \neq 0, \quad \text{para algún } j = 1, \dots, k. \end{aligned}$$

Sea $\boldsymbol{\beta} = (\beta_0, \boldsymbol{\beta}_*^\top)^\top$ con $\boldsymbol{\beta}_* = (\beta_1, \dots, \beta_k)^\top$. Entonces, podemos escribir:

$$H_0 : \boldsymbol{\beta}_* = \mathbf{0}, \qquad H_1 : \boldsymbol{\beta}_* \neq \mathbf{0},$$

o alternativamente, como $H_0 : \boldsymbol{G}\boldsymbol{\beta} = \boldsymbol{g}$, con

$$\boldsymbol{G} = (\mathbf{0}, \boldsymbol{I}_k), \qquad \boldsymbol{g} = \mathbf{0}.$$

El modelo en (3) puede ser escrito como

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} = (\mathbf{1}, \boldsymbol{X}_*)\begin{pmatrix}\beta_0\\ \boldsymbol{\beta}_*\end{pmatrix} + \boldsymbol{\epsilon} = \mathbf{1}\beta_0 + \boldsymbol{X}_*\boldsymbol{\beta}_* + \boldsymbol{\epsilon}.$$

Sabemos que

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{Y}, \qquad Q(\widehat{\boldsymbol{\beta}}) = \boldsymbol{Y}^\top(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y} = RSS,$$

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

Bajo $H_0 : \boldsymbol{\beta}_* = \mathbf{0}$, tenemos

$$\boldsymbol{Y} = \beta_0\mathbf{1} + \boldsymbol{\epsilon}.$$

Así, $\widetilde{\beta}_0 = \overline{Y}$ y

$$Q(\widetilde{\beta}_0) = (\boldsymbol{Y} - \overline{Y}\mathbf{1})^\top(\boldsymbol{Y} - \overline{Y}\mathbf{1}) = \sum_{i=1}^n (Y_i - \overline{Y})^2 = SS_{\mathsf{Total}}.$$

Sea

$$SS_{\text{Regr}} = SS_{\text{Total}} - RSS.$$

Así, el test F para probar $H_0 : \beta_1 = \cdots = \beta_k = 0$ adopta la forma:

Rechazar $H_0 : \boldsymbol{\beta}_* = \mathbf{0}$, si

$$F = \frac{SS_{\text{Regr}}/k}{s^2} \geq F_{1-\alpha}(k, n-k-1),$$

con

$$s^2 = \frac{1}{n-k-1} \boldsymbol{Y}^\top (\boldsymbol{I} - \boldsymbol{H}) \boldsymbol{Y}.$$

Es habitual construir la tabla de análisis de varianza (ANOVA)

Fuente de variación	Suma de cuadrados	Grados de libertad	Media de cuadrados
Regresión	SS_{Regr}	k	SS_{Regr}/k
Residual	RSS	$n-k-1$	$RSS/(n-k-1)$
Total	SS_{Total}	$n-1$	

Test de hipótesis: Ejemplos

Evidentemente

$$(\boldsymbol{Y} - \overline{Y}\boldsymbol{1})^\top(\boldsymbol{Y} - \overline{Y}\boldsymbol{1}) = \boldsymbol{Y}^\top\boldsymbol{C}\boldsymbol{Y} = \boldsymbol{Y}^\top\Big(\boldsymbol{I} - \frac{1}{n}\boldsymbol{1}\boldsymbol{1}^\top\Big)\boldsymbol{Y},$$

mientras que

$$\begin{aligned} SS_{\mathsf{Regr}} &= \boldsymbol{Y}^\top\Big(\boldsymbol{I} - \frac{1}{n}\boldsymbol{1}\boldsymbol{1}^\top\Big)\boldsymbol{Y} - \boldsymbol{Y}^\top(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y} \\ &= \boldsymbol{Y}^\top\Big(\boldsymbol{I} - \frac{1}{n}\boldsymbol{1}\boldsymbol{1}^\top - \boldsymbol{I} + \boldsymbol{H}\Big)\boldsymbol{Y} = \boldsymbol{Y}^\top\Big(\boldsymbol{H} - \frac{1}{n}\boldsymbol{1}\boldsymbol{1}^\top\Big)\boldsymbol{Y}. \end{aligned}$$

Tenemos

$$\Big(\boldsymbol{H} - \frac{1}{n}\boldsymbol{1}\boldsymbol{1}^\top\Big)(\boldsymbol{I} - \boldsymbol{H}) = \boldsymbol{H} - \frac{1}{n}\boldsymbol{1}\boldsymbol{1}^\top - \boldsymbol{H}^2 + \frac{1}{n}\boldsymbol{1}\boldsymbol{1}^\top\boldsymbol{H}.$$

La matriz asociada al modelo (3) es $\boldsymbol{X} = (\boldsymbol{1}, \boldsymbol{X}_*)$. Sabemos que

$$\boldsymbol{H}\boldsymbol{X} = \boldsymbol{X} \quad \Rightarrow \quad \boldsymbol{H}(\boldsymbol{1}, \boldsymbol{X}_*) = (\boldsymbol{1}, \boldsymbol{X}_*) \quad \Rightarrow \quad \boldsymbol{H}\boldsymbol{1} = \boldsymbol{1}.$$

De ahí que

$$\Big(\boldsymbol{H} - \frac{1}{n}\boldsymbol{1}\boldsymbol{1}^\top\Big)(\boldsymbol{I} - \boldsymbol{H}) = \boldsymbol{0},$$

es decir SS_{Regr} y RSS son independientes (como es esperado).[2]

[2] También podemos argumentar la independencia usando el Teorema de Cochran.

Cemento Portland (Woods, Steinour y Starke, 1932)[3]

Ejemplo (Datos de cemento Portland):

Estudio experimental relacionando la emisión de calor durante la producción y endurecimiento de 13 muestras de cementos Portland. Woods et al. (1932) consideraron cuatro compuestos para los clinkers desde los que se produce el cemento.

La respuesta (Y) es la emisión de calor después de 180 días de curado, medido en calorías por gramo de cemento. Los regresores son los porcentajes de los cuatro compuestos: aluminato tricálcico (X_1), silicato tricálcico (X_2), ferrito aluminato tetracálcico (X_3) y silicato dicálcico (X_4).

[3] Industrial and Engineering Chemistry **24**, 1207-1214.

Cemento Portland (Woods, Steinour y Starke, 1932)

```
# base de datos
> load("portland.rda")
> portland
       y x1 x2 x3 x4
1   78.5  7 26  6 60
2   74.3  1 29 15 52
3  104.3 11 56  8 20
4   87.6 11 31  8 47
5   95.9  7 52  6 33
6  109.2 11 55  9 22
7  102.7  3 71 17  6
8   72.5  1 31 22 44
9   93.1  2 54 18 22
10 115.9 21 47  4 26
11  83.8  1 40 23 34
12 113.3 11 66  9 12
13 109.4 10 68  8 12

# en efecto,
> apply(portland[,-1], 1, sum)
 1  2  3  4  5  6  7  8  9 10 11 12 13
99 97 95 97 98 97 97 98 96 98 98 98 98
```

Cemento Portland (Woods, Steinour y Starke, 1932)

```
# Ajuste usando función 'lm'
> fm <- lm(y ~ x1 + x2 + x3 + x4, data = portland)
> fm

Call:
lm(formula = y ~ x1 + x2 + x3 + x4, data = portland)

Coefficients:
(Intercept)           x1           x2           x3           x4
    62.4054       1.5511       0.5102       0.1019      -0.1441
```

```
# Salida de función 'summary'
> summary(fm)

Call:
lm(formula = y ~ x1 + x2 + x3 + x4, data = portland)

Residuals:
    Min      1Q  Median      3Q     Max
-3.1750 -1.6709  0.2508  1.3783  3.9254

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  62.4054    70.0710   0.891   0.3991
x1            1.5511     0.7448   2.083   0.0708 .
x2            0.5102     0.7238   0.705   0.5009
x3            0.1019     0.7547   0.135   0.8959
x4           -0.1441     0.7091  -0.203   0.8441

Residual standard error: 2.446 on 8 degrees of freedom
Multiple R-squared:  0.9824,    Adjusted R-squared:  0.9736
F-statistic: 111.5 on 4 and 8 DF,  p-value: 4.756e-07
```