

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:

$$Y_{ij} = \mu_i + \epsilon_{ij}, \quad i = 1, \dots, p; j = 1, \dots, n. \quad (1)$$

Se tiene interés en probar la hipótesis

$$H_0 : \mu_1 = \mu_2 = \dots = \mu_p,$$

$$H_1 : \mu_r \neq \mu_s, \quad \text{para algún } r \neq s.$$

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

$$G = \begin{pmatrix} 1 & 0 & \dots & 0 & -1 \\ 0 & 1 & \dots & 0 & -1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & 1 & -1 \end{pmatrix} = (I_{p-1}, -1), \quad g = 0$$



Test de hipótesis: Ejemplos

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

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

donde

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_p \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{pmatrix}, \quad \boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_p \end{pmatrix}$$

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

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

y

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



Test de hipótesis: Ejemplos

Ahora, bajo¹ $H_0 : \mu_1 = \dots = \mu_p$ ($= \mu$), tenemos que el modelo (reducido) puede ser escrito como:

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

De este modo,

$$\tilde{\mu} = \frac{1}{np} \mathbf{1}^\top \mathbf{Y}, \quad \tilde{\boldsymbol{\beta}} = (\bar{Y}, \bar{Y}, \dots, \bar{Y})^\top = \bar{Y} \mathbf{1}_p,$$

mientras que

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

¹Es decir, suponiendo que H_0 es verdadera.



Test de hipótesis: Ejemplos

En efecto, tenemos

$$G = (G_q, G_r) = (I_{p-1}, -1), \quad g = 0, \quad X = (X_q, X_r),$$

donde

$$X_q = \begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 \\ 0 & 0 & \dots & 0 \end{pmatrix}, \quad X_r = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{pmatrix}$$

luego

$$Y_R = Y - X_q G_q^{-1} g = Y.$$



Mientras que

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

De este modo

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

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

Test de hipótesis: Ejemplos

De este modo,

$$Q(\tilde{\beta}) - Q(\hat{\beta}) = \sum_{j=1}^n \sum_{i=1}^p (Y_{ij} - \bar{Y})^2 - \sum_{j=1}^n \sum_{i=1}^p (Y_{ij} - \bar{Y}_i)^2,$$

y

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

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

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



Ejemplo (Modelo de regresión lineal múltiple):

Considere el modelo:

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

con los supuestos habituales, y suponga que se desea 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 $\beta = (\beta_0, \beta_*^\top)^\top$ con $\beta_* = (\beta_1, \dots, \beta_k)^\top$. Entonces, podemos escribir:

$$H_0 : \beta_* = \mathbf{0}, \quad H_1 : \beta_* \neq \mathbf{0},$$

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

$$G = (\mathbf{0}, I_k), \quad g = \mathbf{0}.$$



El modelo en (3) puede ser escrito como

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

Sabemos que

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

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

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

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

Así, $\tilde{\beta}_0 = \bar{Y}$ y

$$Q(\tilde{\beta}_0) = (\mathbf{Y} - \bar{Y}\mathbf{1})^\top (\mathbf{Y} - \bar{Y}\mathbf{1}) = \sum_{i=1}^n (Y_i - \bar{Y})^2 = SS_{\text{Total}}.$$



Test de hipótesis: Ejemplos

Sea

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

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

Rechazar $H_0 : \beta_* = \mathbf{0}$, si

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

con

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



Test de hipótesis: Ejemplos

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

$$(\mathbf{Y} - \bar{Y}\mathbf{1})^\top (\mathbf{Y} - \bar{Y}\mathbf{1}) = \mathbf{Y}^\top \mathbf{C} \mathbf{Y} = \mathbf{Y}^\top \left(\mathbf{I} - \frac{1}{n} \mathbf{1}\mathbf{1}^\top \right) \mathbf{Y},$$

mientras que

$$\begin{aligned} SS_{\text{Regr}} &= \mathbf{Y}^\top \left(\mathbf{I} - \frac{1}{n} \mathbf{1}\mathbf{1}^\top \right) \mathbf{Y} - \mathbf{Y}^\top (\mathbf{I} - \mathbf{H}) \mathbf{Y} \\ &= \mathbf{Y}^\top \left(\mathbf{I} - \frac{1}{n} \mathbf{1}\mathbf{1}^\top - \mathbf{I} + \mathbf{H} \right) \mathbf{Y} = \mathbf{Y}^\top \left(\mathbf{H} - \frac{1}{n} \mathbf{1}\mathbf{1}^\top \right) \mathbf{Y}. \end{aligned}$$

Tenemos

$$\left(\mathbf{H} - \frac{1}{n} \mathbf{1}\mathbf{1}^\top \right) (\mathbf{I} - \mathbf{H}) = \mathbf{H} - \frac{1}{n} \mathbf{1}\mathbf{1}^\top - \mathbf{H}^2 + \frac{1}{n} \mathbf{1}\mathbf{1}^\top \mathbf{H}.$$

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

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

De ahí que

$$\left(\mathbf{H} - \frac{1}{n} \mathbf{1}\mathbf{1}^\top \right) (\mathbf{I} - \mathbf{H}) = \mathbf{0},$$

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

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



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 tetra-cálcico** (X_3) y **silicato dicálcico** (X_4).

³Industrial 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



Cemento Portland (Woods, Steinour y Starke, 1932)

```
# 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

