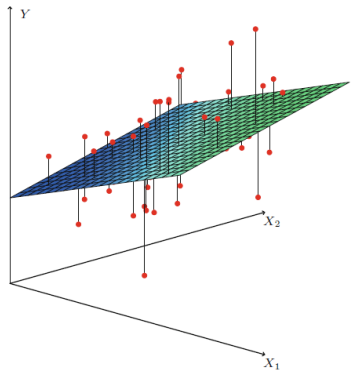
Pruebas de Hipótesis y Tablas ANVA

Teoría y Práctica

Christian Amao Suxo

Escuela Profesional de Ingeniería Estadística
Universidad Nacional de Ingeniería



Semestre I - 2020

Temario

1. **Prueba de hipótesis Lineal General**
   - Construcción de la prueba de hipótesis lineal general
   - Prueba de significancia global del modelo
   - Tabla ANVA para la prueba de significancia global del modelo
   - Prueba de la significancia de un subgrupo de variables
   - Tabla ANVA de la significancia de un subgrupo de variables

2. **Prueba de hipótesis por Ratio de Verosimilitud**
   - Construcción del ratio de verosimilitud
   - Relación entre la prueba de H.L.G. y la prueba por R.V.

¿Para qué sirven las pruebas de hipótesis?

Dado un modelo lineal clásico de la forma

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},\ con\ \boldsymbol{\varepsilon} \sim \mathcal{N}_n(\mathbf{0}, \sigma^2\mathbb{I}_n)$$

Surgen las siguientes interrogantes:

- ¿Cómo saber si el conjunto de variables explicativas que tengo en verdad influyen en Y?
- ¿Cómo saber si un subgrupo de estas variables influyen sobre Y?
- ¿Cómo saber si la variable explicativa X_j, para algún $j = 1, 2, \ldots, p$, influye significativamente sobre Y?

Entonces, ¿cómo construir pruebas de hipótesis estadísticas para responder estas interrogantes?

¿Para qué sirven las pruebas de hipótesis?

Dado un modelo lineal clásico de la forma

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},\ con\ \boldsymbol{\varepsilon} \sim \mathcal{N}_n(\mathbf{0}, \sigma^2\mathbb{I}_n)$$

Surgen las siguientes interrogantes:

- ¿Cómo saber si el conjunto de variables explicativas que tengo en verdad influyen en Y?
- ¿Cómo saber si un subgrupo de estas variables influyen sobre Y?
- ¿Cómo saber si la variable explicativa X_j, para algún $j = 1, 2, \ldots, p$, influye significativamente sobre Y?

Entonces, ¿cómo construir pruebas de hipótesis estadísticas para responder estas interrogantes?

¿Cómo construir pruebas de hipótesis?

Prueba de hipótesis lineal general

En general, las pruebas de hipótesis sobre los coeficientes β's se pueden expresar de la forma:

$$H_0 : R\boldsymbol{\beta} = r \;\; vs \;\; H_1 : R\boldsymbol{\beta} \neq r, \tag{1}$$

donde R es una matriz de orden $q \times (p+1)$, r es un vector de orden $q \times 1$ y $\boldsymbol{\beta}$ es el vector de coeficientes. Por ejemplo, si se desea probar:

1. *La significancia global del modelo:* $R = (\underset{p\times 1}{\mathbf{0}} \,|\, \mathbb{I}_p)$ y $r = \underset{p\times 1}{\mathbf{0}}$.
2. *La significancia de las últimas* $k(<p)$ *variables:* $R = (\mathbf{0}|\mathbb{I}_k)$ y $r = \underset{k\times 1}{\mathbf{0}}$.
3. *La significancia de la variable* $j-$*ésima:* $R = \zeta_j'$ y $r = 0$, donde ζ_j es un vector de ceros con el valor de 1 en la posición $j-$ésima.

¿Cómo plantear una prueba de hipótesis lineal general

Pasos para desarrollar la prueba de hipótesis lineal general

1.- Predefinir un nivel de significancia α y la prueba de hipótesis:

$$H_0 : R\boldsymbol{\beta} = r \;\; vs \;\; H_1 : R\boldsymbol{\beta} \neq r$$

2.- Hallar un estadístico pivote adecuado para probar la hipótesis.

3.- Calcular el estadístico de prueba suponiendo hipótesis nula verdadera:

$$F_c = \frac{(R\hat{\boldsymbol{\beta}} - r)'(R(\mathbf{X}'\mathbf{X})^{-1}R')^{-1}(R\hat{\boldsymbol{\beta}} - r)}{q\hat{\sigma}^2_{MCO}} \sim F(q, n-p-1) \qquad (2)$$

4.- Se compara el estadístico calculado con el valor en tablas:

- Si $F_c \leq F_\alpha(q, n-p-1)$ entonces no rechazo H_0.
- Si $F_c > F_\alpha(q, n-p-1)$ entonces rechazo H_0.

¿Cómo probar la significancia global de un modelo lineal?

Prueba de significancia global del modelo

Esta prueba sirve para probar si el modelo lineal propuesto es adecuado o no. Este tipo de hipótesis se puede plantear usando una hipótesis lineal general con $R = (\underset{p\times 1}{\mathbf{0}} \mid \mathbb{I}_p)$ y $r = \underset{p\times 1}{\mathbf{0}}$. Para esta prueba, se demuestra que el estadístico de prueba es:

$$F_c = \frac{\mathbf{Y}'(H_{\mathbf{X}} - \frac{1}{n}\mathbb{J}_n)\mathbf{Y}}{p\hat{\sigma}^2_{MCO}} \sim F(p, n-p-1) \tag{3}$$

Observación: Si usamos el enfoque del modelo centrado, entonces:

$$\mathbf{Y}'(H_{\mathbf{X}} - \frac{1}{n}\mathbb{J}_n)\mathbf{Y} = \mathbf{Y}'\mathbf{X}_C(\mathbf{X}_C'\mathbf{X}_C)^{-1}\mathbf{X}_C'\mathbf{Y} = \mathbf{Y}'H_{\mathbf{X}_C}\mathbf{Y} \tag{4}$$

¿Cómo construir la tabla ANVA para esta prueba?

Se parte de la identidad:

$$SCT = SCE + SCR$$

$$\mathbf{Y}'(\mathbb{I}_n - \frac{1}{n}\mathbb{J}_n)\mathbf{Y} = \mathbf{Y}'(H_{\mathbf{X}} - \frac{1}{n}\mathbb{J}_n)\mathbf{Y} + \mathbf{Y}'(\mathbb{I}_n - H_{\mathbf{X}})\mathbf{Y}$$

Tabla 1: Tabla ANVA para la prueba de hipótesis de significancia global del modelo

Fuente de Variación	Grados de Libertad	Suma de Cuadrados	Cuadrados Medios	F calculado
Regresión	p	$\mathbf{Y}'(H_{\mathbf{X}} - \frac{1}{n}\mathbb{J}_n)\mathbf{Y}$	$CME = \frac{\mathbf{Y}'(H_{\mathbf{X}} - \frac{1}{n}\mathbb{J}_n)\mathbf{Y}}{p}$	CME/CMR
Residual	$n-p-1$	$\mathbf{Y}'(\mathbb{I}_n - H_{\mathbf{X}})\mathbf{Y}$	$CMR = \frac{\mathbf{Y}'(\mathbb{I}_n - H_{\mathbf{X}})\mathbf{Y}}{n-p-1}$	
Total	$n-1$	$\mathbf{Y}'(\mathbb{I}_n - \frac{1}{n}\mathbb{J}_n)\mathbf{Y}$		

¿Cómo probar la significancia de un subgrupo de variables?

Prueba de significancia de un subgrupo de variables

Sin pérdida de generalidad, supóngase que se desea probar la significancia de las últimas k $(< p)$ variables regresoras. La hipótesis se puede plantear como:

$$H_0 : \beta_i = 0, \forall i \in N_k \;\; vs \;\; H_1 : \beta_j \neq 0, \; para\, algún\, j \in N_k, \tag{5}$$

donde $N_k = \{p-k+1, p-k+2, \ldots, p\}$.
Para plantear el estadístico de prueba, se particiona el modelo de la forma:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} = \mathbf{X}_1\boldsymbol{\beta_1} + \mathbf{X}_2\boldsymbol{\beta_2} + \boldsymbol{\varepsilon}, \tag{6}$$

donde $\mathbf{X} = (\underset{n\times(p-k+1)}{\mathbf{X}_1} \mid \underset{n\times k}{\mathbf{X}_2})$ y $\boldsymbol{\beta} = (\boldsymbol{\beta_1}'|\boldsymbol{\beta_2}')'$.

¿Cómo probar la significancia de un subgrupo de variables?

Prueba de significancia de un subgrupo de variables

Con la partición mencionada, la hipótesis se puede plantear como

$$H_0 : \boldsymbol{\beta_2} = \mathbf{0} \;\; vs \;\; H_1 : \boldsymbol{\beta_2} \neq \mathbf{0}. \tag{7}$$

La ecuación (7) se traduce en una prueba de hipótesis lineal general con $R = (\mathbf{0}|\mathbb{I}_k)$ y $r = \underset{k\times 1}{\mathbf{0}}$. Con esto, el estadístico de prueba es de la forma:

$$F_c = \frac{\mathbf{Y}'(H_{\mathbf{X}} - H_{\mathbf{X}_1})\mathbf{Y}}{k\hat{\sigma}^2_{MCO}} \sim F(k, n-p-1) \tag{8}$$

¿Cómo construir una tabla ANVA para un subgrupo de variables?

1. **Suma de cuadrados explicada de un modelo saturado:** Dado un modelo de la forma $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, entonces se define:

$$SCE(\boldsymbol{\beta}) = SCE(\beta_0, \beta_1, \ldots, \beta_p) := \mathbf{Y}' H_{\mathbf{X}} \mathbf{Y} \tag{9}$$

2. **Suma de cuadrados extra:** Dado un modelo de la forma $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$. Si $\boldsymbol{\beta_1} = (\beta_{i_1}, \ldots, \beta_{i_k})'$ y $\boldsymbol{\beta_2} = (\beta_{i_{k+1}}, \ldots, \beta_{i_{p+1}})$ donde $\{i_1, i_2, \ldots, i_{p+1}\} = \{0, 1, \ldots, p\}$. Se define la suma de cuadrados extra de las variables $X_{i_{k+1}}, X_{i_{k+2}}, \ldots, X_{i_{p+1}}$ dado que las variables $X_{i_1}, X_{i_2}, \ldots, X_{i_k}$ están en el modelo como:

$$SCE(\boldsymbol{\beta_2}|\boldsymbol{\beta_1}) := SCE(\boldsymbol{\beta}) - SCE(\boldsymbol{\beta_1}), \tag{10}$$

donde $SCE(\boldsymbol{\beta_1}) = \mathbf{Y}' H_{\mathbf{Z}} \mathbf{Y}$ con $\mathbf{Z} = (\mathbf{x}_{(i_1)}| \ldots |\mathbf{x}_{(i_k)})$.

¿Cómo construir una tabla ANVA para un subgrupo de variables?

Dado un modelo lineal particionado como en la ecuación (6) y la definición de la suma de cuadrados extra, el estadístico de prueba para probar la significancia de las últimas k variables regresoras se puede expresar como:

$$F_c = \frac{SCE(\boldsymbol{\beta_2}|\boldsymbol{\beta_1})}{k\hat{\sigma}^2_{MCO}} \sim F(k, n-p-1), \tag{11}$$

donde:

$$SCE(\boldsymbol{\beta_2}|\boldsymbol{\beta_1}) = \mathbf{Y}'(H_{\mathbf{X}} - H_{\mathbf{X}_1})\mathbf{Y}.$$

¿Cómo construir una tabla ANVA para un subgrupo de variables?

Para construir la tabla ANVA se hace uso de la siguiente identidad

$$\mathbf{Y}'\mathbf{Y} = \mathbf{Y}'H_{\mathbf{X}_1}\mathbf{Y} + \mathbf{Y}'(H_{\mathbf{X}} - H_{\mathbf{X}_1})\mathbf{Y} + \mathbf{Y}'(\mathbb{I}_n - H_{\mathbf{X}})\mathbf{Y} \tag{12}$$

$$SCT^* = SCE(\boldsymbol{\beta_1}) + SCE(\boldsymbol{\beta_2}|\boldsymbol{\beta_1}) + SCR \tag{13}$$

Tabla 2: Tabla ANVA para probar la significancia de un subgrupo de variables

Fuente de Variación	G. L.	Suma de Cuadrados	Cuadrados Medios	F calculado
Debido a $\boldsymbol{\beta_1}$	$p+1-k$	$\mathbf{Y}'H_{\mathbf{X}_1}\mathbf{Y}$	$CME(\boldsymbol{\beta_1}) = \frac{\mathbf{Y}'H_{\mathbf{X}_1}\mathbf{Y}}{p+1-k}$	
Debido a $\boldsymbol{\beta_2}\|\boldsymbol{\beta_1}$	k	$\mathbf{Y}'(H_{\mathbf{X}}-H_{\mathbf{X}_1})\mathbf{Y}$	$CME(\boldsymbol{\beta_2}\|\boldsymbol{\beta_1}) = \frac{\mathbf{Y}'(H_{\mathbf{X}}-H_{\mathbf{X}_1})\mathbf{Y}}{k}$	$\frac{CME(\boldsymbol{\beta_2}\|\boldsymbol{\beta_1})}{CMR}$
Residual	$n-p-1$	$\mathbf{Y}'(\mathbb{I}_n - H_{\mathbf{X}})\mathbf{Y}$	$CMR = \frac{\mathbf{Y}'(\mathbb{I}_n-H_{\mathbf{X}})\mathbf{Y}}{n-p-1}$	
Total*	n	$\mathbf{Y}'\mathbf{Y}$		

¿En qué consiste la prueba de hipótesis por ratio de verosimilitud?

Ratio de verosimilitud

Dada la prueba de hipótesis $H_0 : R\boldsymbol{\beta} = r \;\; vs \;\; H_1 : R\boldsymbol{\beta} \neq r$, el ratio de verosimilitud se define como:

$$\lambda = \frac{\max\limits_{\boldsymbol{\beta}\in\Phi} L(\theta)}{\max\limits_{\boldsymbol{\beta}\in\Omega} L(\theta)}, \tag{14}$$

donde $\theta = (\beta', \sigma^2)'$ son los parámetros del modelo lineal, $L(.)$ la función de verosimilitud, Ω es el **espacio parametral** y Φ el **espacio parametral restringido a** H_0, esto es

$$\Phi = \{\beta^* \in \Omega : R\beta^* = r\} \tag{15}$$

¿En qué consiste la prueba de hipótesis por ratio de verosimilitud?

Observaciones:

- Note que $0 < \lambda \leq 1$.
- Si $\lambda \approx 1$, entonces nos aproximariamos a creer que H_0 es "cierta".
- Si $\lambda \approx 0$, entonces nos aproximariamos a creer que H_0 es "falsa".

Interrogante: ¿Para qué valor umbral del ratio de verosimilitud se podría tomar una decisión estadística conclusiva?

¿En qué consiste la prueba de hipótesis por ratio de verosimilitud?

Observaciones:

- Note que $0 < \lambda \leq 1$.
- Si $\lambda \approx 1$, entonces nos aproximariamos a creer que H_0 es "cierta".
- Si $\lambda \approx 0$, entonces nos aproximariamos a creer que H_0 es "falsa".

Interrogante: ¿Para qué valor umbral del ratio de verosimilitud se podría tomar una decisión estadística conclusiva?

¿En qué consiste la prueba de hipótesis por ratio de verosimilitud?

Para responder esta pregunta primero obtenemos la forma explícita del ratio de verosimilitud. Se demuestra que

$$\lambda = \left(\frac{\hat{\boldsymbol{\varepsilon}}' \hat{\boldsymbol{\varepsilon}}}{\hat{\boldsymbol{\varepsilon}}'_{\omega} \hat{\boldsymbol{\varepsilon}}_{\omega}} \right)^{n/2} \tag{16}$$

donde $\hat{\boldsymbol{\varepsilon}} = \mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}_{MV}$ y $\hat{\boldsymbol{\varepsilon}}_{\omega} = \mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}_{MVR}$ con

$$\hat{\boldsymbol{\beta}}_{MVR} = \hat{\boldsymbol{\beta}}_{MV} + (\mathbf{X}'\mathbf{X})^{-1} R' (R(\mathbf{X}'\mathbf{X})^{-1} R')^{-1} (r - R\hat{\boldsymbol{\beta}}_{MV}) \tag{17}$$

¿Cómo construir la prueba por ratio de verosimilitud?

Prueba de hipótesis por ratio de verosimilitud

Sea un modelo lineal clásico de la forma $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ con p variables explicativas y término independiente. Si se desea probar la hipótesis estadística general

$$H_0 : R\boldsymbol{\beta} = r \;\; vs \;\; H_1 : R\boldsymbol{\beta} \neq r$$

entonces el estadístico de prueba por ratio de verosimilitud es

$$F_{RV} = \frac{n-p-1}{q}\left(\lambda^{-2/n} - 1\right) \sim F(q, n-p-1) \tag{18}$$

donde con un nivel de significancia α el criterio de decisión es

- Si $F_{RV} > F_\alpha(q, n-p-1)$ entonces se rechaza H_0.
- Si $F_{RV} \leq F_\alpha(q, n-p-1)$ entonces no se rechaza H_0.

¿Qué relación existe entre la prueba de H.L.G. y la prueba por R.V.?

Teorema

Sea un modelo lineal clásico de la forma $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ con p variables explicativas y término independiente. Si se desea probar la hipótesis estadística general

$$H_0 : R\boldsymbol{\beta} = r \;\; vs \;\; H_1 : R\boldsymbol{\beta} \neq r$$

entonces el estadístico de prueba clásico de la hipótesis lineal general (ecuación (2)) y por ratio de verosimilitud (ecuación (18)) son equivalentes. Esto es

$$\frac{n-p-1}{q}\left(\lambda^{-2/n} - 1\right) = \frac{(R\hat{\boldsymbol{\beta}} - r)'(R(\mathbf{X}'\mathbf{X})^{-1}R')^{-1}(R\hat{\boldsymbol{\beta}} - r)}{q\hat{\sigma}^2_{MCO}} \tag{19}$$

¿Preguntas?