Regresión Lineal Múltiple - Semana 06

Johnatan Cardona Jiménez
jcardonj@unal.edu.co
Profesor Asistente - Departamento de Estadística
Universidad Nacional de Colombia, Sede Medellín

Semestre 02-2024

Valores ajustados y residuales

Con los valores ajustados $\widehat{Y}_i$ se construye el vector de valores ajustados dado por

$$\underline{\widehat{\boldsymbol{y}}} = \boldsymbol{X}\underline{\widehat{\beta}} = \begin{bmatrix} \widehat{Y}_1 \\ \widehat{Y}_2 \\ \vdots \\ \widehat{Y}_n \end{bmatrix}$$

Note que el vector $\underline{\widehat{\boldsymbol{y}}}$ se puede reescribir como:

$$\underline{\widehat{\boldsymbol{y}}} = \boldsymbol{X}\underline{\widehat{\beta}} = \boldsymbol{X}\underbrace{\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\ \boldsymbol{X}'\underline{\boldsymbol{y}}}_{\underline{\widehat{\beta}}} = \overbrace{\boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\ \boldsymbol{X}'}^{H}\underline{\boldsymbol{y}} = \boldsymbol{H}\underline{\boldsymbol{y}}$$

{ Con $\boldsymbol{H}_{n\times n} = \boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\ \boldsymbol{X}'$, donde a la matriz $\boldsymbol{H}$ se le conoce como la matriz "**hat**" debido a que su multiplicación por el vector de observaciones $\underline{\boldsymbol{y}}$ lleva al vector de valores ajustados $\underline{\widehat{\boldsymbol{y}}}$ ($\underline{\boldsymbol{y}}$ "**hat**").}

Realmente, la matriz $\boldsymbol{H}$ es una matriz de proyección ortogonal (cuadrada, simétrica e idempotente) que proyecta a $\underline{\boldsymbol{y}}$ en el plano ajustado. **Esta matriz juega un papel muy importante en regresión tanto en la estimación como en la determinación de valores extremos, que será desarrollada más adelante**.

Los residuales del modelo corresponden como en el caso de RLS a las diferencias entre los valores observados y los valores ajustados, esto es, $e_i = Y_i - \widehat{Y}_i$ y el vector de residuales es:

$$\underline{\boldsymbol{e}} = \underline{\boldsymbol{y}} - \underline{\widehat{\boldsymbol{y}}} = \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}_{n\times 1}$$

El vector de residuales también puede expresarse en términos de la matriz $\boldsymbol{H}$, ya que $\underline{\boldsymbol{e}} = \underline{\boldsymbol{y}} - \underline{\widehat{\boldsymbol{y}}} = \underline{\boldsymbol{y}} - \boldsymbol{H}\boldsymbol{y} = (\boldsymbol{I}_n - \boldsymbol{H})\,\underline{\boldsymbol{y}}$.

Estimación de la varianza

Bajo los supuestos relativos a los errores del modelo

$$\varepsilon_i \overset{\text{iid}}{\sim} N\left(0, \sigma^2\right), \quad i = 1, 2, \ldots, n,$$

el estimador insesgado de la varianza corresponde a:

$$\widehat{\sigma}^2 = \text{MSE} = \frac{\text{SSE}}{n-p},$$

donde $p = k + 1$ es el número de parámetros del modelo y la suma de cuadrados del error SSE corresponde a:

$$\text{SSE} = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2 = \left(\underline{\boldsymbol{y}} - \underline{\widehat{\boldsymbol{y}}}\right)' \left(\underline{\boldsymbol{y}} - \underline{\widehat{\boldsymbol{y}}}\right) = \underline{\boldsymbol{y}}' \left(\boldsymbol{I} - \boldsymbol{H}\right) \underline{\boldsymbol{y}}.$$

Análisis de varianza

Al igual que en RLS en RLM se tiene un procedimiento de prueba basado en el análisis de varianza para probar la significancia de la regresión, que establece el siguiente juego de hipótesis:

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_K = 0, \quad \text{vs.}$$
$$H_1 : \text{algún } \beta_j \neq 0, j = 1, \ldots, K.$$

En este enfoque todavía es válida la identidad de suma de cuadrados que establece que:

$$\begin{array}{ccccc} \text{SST} & = & \text{SSR} & + & \text{SSE} \\ \sum_{i=1}^{n} (y_i - \bar{y})^2 & = & \sum_{i=1}^{n} (\widehat{y}_i - \bar{y})^2 & + & \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2 \end{array}$$

En RLM, las sumas de cuadrados se pueden expresar en forma matricial, así:

Sumas de cuadrados en forma matricial

En las siguientes fórmulas $\boldsymbol{J}$ es una matriz de dimensión $n \times n$ cuyas entradas son todas iguales a 1, e $\boldsymbol{I}$ es la matriz identidad de orden n, ie. $\boldsymbol{J}_{n\times n}$ e $\boldsymbol{I}_{n\times n}$:

Fuente	Suma de cuadrados
Regresión	$\text{SSR} = \underline{\boldsymbol{y}}' \left[\boldsymbol{H} - \left(\frac{1}{n}\right)\boldsymbol{J}\right] \underline{\boldsymbol{y}}$
Error	$\text{SSE} = \underline{\boldsymbol{y}}' \left(\boldsymbol{I} - \boldsymbol{H}\right) \underline{\boldsymbol{y}}$
Total	$\text{SST} = \underline{\boldsymbol{y}}' \left[\boldsymbol{I} - \left(\frac{1}{n}\right)\boldsymbol{J}\right] \underline{\boldsymbol{y}}$

El procedimiento de prueba se resume en la siguiente tabla.

Tabla de análisis de varianza para el modelo de RLM

Fuente de Variación	Suma de Cuadrados	Grados de Libertad	Cuadrado Media	F Calculado
Regresión o Modelo	SSR	$k = p - 1$	$MSR = \frac{SSR}{k}$	$F_0 = \frac{MSR}{MSE}$
Error o Residual	SSE	$n - p$	$MSE = \frac{SSE}{n-p}$	
Total	SST	$n - 1$		

Se rechaza H_0 a una significancia dada α si $F_0 > f_{1-\alpha;k,n-p}$. Equivalentemente, si se define el valor-P para la prueba como vp $= P(f_{k,n-p} > F_0)$, se rechaza H_0 si vp $< \alpha$. Al rechazar H_0, **se prueba que existe una relación de regresión**, sin embargo, esto no garantiza que el modelo resulte útil para hacer predicciones.

El coeficiente de determinación múltiple

Denotado por R^2 y definido como:

$$R^2 = \frac{\text{SSR}}{\text{SST}} = 1 - \frac{\text{SSE}}{\text{SST}},$$

mide la proporción de la variabilidad total observada en la respuesta que es explicada por el modelo propuesto (esto es, la asociación lineal con el conjunto de variables $X_1, X_2, \ldots, X_k$).

Por ser una proporción, esta cantidad varía entre 0 y 1:

- Siendo igual a 0, si todos los coeficientes de regresión ajustados son iguales a cero, y
- Siendo igual a 1, si todas las observaciones caen sobre la superficie de regresión ajustada.

Aunque es usado como una medida de bondad del ajuste de la función de regresión, es necesario tener presente que:

- Valores grandes de R^2 no implican necesariamente que la superficie ajustada sea útil. Puede suceder que se hayan observado pocos niveles de las variables predictoras y por tanto la superficie ajustada no sería útil para hacer extrapolaciones por fuera de tales rangos. Incluso, si esta cantidad es muy cercana a 1, todavía el MSE podría ser muy grande y por tanto las inferencias tendrían poca precisión.
- Cuando se agregan más variables predictoras al modelo, el R^2 tiende a no decrecer, aún cuando existan dentro del grupo de variables, un subconjunto de ellas que no aportan significativamente.

- Como **medida de bondad de ajuste** se prefiere usar otros estadísticos que penalicen al modelo por el número de variables incluidas, entre ellos se tienen el MSE, y el R^2 **ajustado**, estas dos medidas son equivalentes, dado que éste último se define como:

$$R^2_{\text{adj}} = 1 - \frac{(n-1)\,\text{MSE}}{\text{SST}}$$

  El R^2 ajustado disminuye cuando en el modelo se ingresan variables predictoras que no logran reducir al SSE, y que causan la pérdida de grados de libertad para este último.

Entre dos modelos ajustados se considera **mejor el de menor MSE o equivalentemente el de mayor R^2 ajustado**.

Inferencias sobre los parámetros del modelo de regresión

Se puede demostrar que bajo los supuestos del modelo de regresión, se cumple que:

$$T_j = \frac{\widehat{\beta}_j - \beta_j}{\text{ee}\left(\widehat{\beta}_j\right)} \sim t_{n-p},\ j = 0, 1, \ldots, k, \qquad (\star)$$

con $\text{ee}\left(\widehat{\beta}_j\right) = \sqrt{\widehat{Var}\left(\widehat{\beta}_j\right)}$ y t_{n-p} una variable aleatoria t-Student con $n - p$ grados de libertad.

Basados en este resultado se pueden construir pruebas de hipótesis e intervalos de confianza para los parámetros del modelo de RLM como se describe a continuación.
dar click para ir a icpi

Pruebas de hipótesis sobre los parámetros del modelo de RLM

Se tienen en total $p = k + 1$ pruebas de hipótesis sobre los coeficientes individuales del modelo de RLM. Veamos el procedimiento para el j-ésimo parámetro $(j = 0, 1, \ldots, k)$. Se quiere probar:

$$\begin{array}{l} H_0 : \beta_j = B_{j,0} \\ H_1 : \beta_j \neq B_{j,0} \end{array} \quad \text{con} \quad B_{j,0} \in \mathbb{R}$$

En resumen, para β_j se tiene que:

Estadístico de prueba	Criterio de rechazo
$T_{j,0} = \dfrac{\widehat{\beta}_j - B_{j,0}}{\text{se}\left(\widehat{\beta}_j\right)} \overset{\text{bajo } H_0}{\sim} t_{n-p}$	Rechazar H_0 si $\lvert T_{j,0} \rvert > t_{1-\alpha/2,n-p}$; con nivel de significancia α

NOTA: Un caso particular de las pruebas de hipótesis anteriores son las conocidas **pruebas de significancia de los parámetros individuales**, donde el procedimiento de prueba es idéntico al anteriormente mostrado haciendo $B_{j,0} = 0$. Acá, las hipótesis son:

$$H_0 : \beta_j = 0$$
$$H_1 : \beta_j \neq 0$$

cuyo procedimiento de prueba se resume como:

Estadístico de prueba	Criterio de rechazo
$T_{j,0} = \dfrac{\widehat{\beta}_j}{\text{se}\left(\widehat{\beta}_j\right)} \overset{\text{bajo } H_0}{\sim} t_{n-p}$	Rechazar H_0 si $\lvert T_{j,0}\rvert > t_{1-\alpha/2,n-p}$; con nivel de significancia α

Intervalos de confianza para los parámetros del modelo de RLM

De nuevo con base en el resultado dar cick aqui: $(\star)$ un intervalo de confianza (IC) del $(1-\alpha)\%$ para el j-ésimo parámetro β_j $(j = 0, 1, \ldots, k)$, es:

$$\widehat{\beta}_j \pm t_{1-\alpha/2, n-p} \, \text{se}\left(\widehat{\beta}_j\right)$$

donde $t_{1-\alpha/2,n-p}$ es el percentil $1-\alpha/2$ de la distribución t-Student con $n-p$ grados de libertad.

Prueba de la significancia de un subconjunto de coeficientes de la regresión

- Considere el caso en que se desea probar simultáneamente la significancia de uno o más coeficientes de la regresión, reunidos en un subconjunto A, dado que otro grupo de coeficientes reunidos en el subconjunto B ya se encuentran en el modelo.
- Se debe así separar la *importancia* de los coeficientes de regresión del subconjunto A dado que los coeficientes de regresión en el subconjunto B ya están presentes en el modelo.

Una forma de medir la importancia de un subconjunto de coeficientes en un modelo de RLM es a través de las denominadas **sumas extra de cuadrados**.

Una **sumas extra de cuadrados** (SSextra) mide la reducción marginal en la SSE (o el incremento marginal en la SSR) producida por uno o varios coeficientes de regresión, dado que los otros coeficientes de regresión están presentes en el modelo.

Una notación para las SSextra en un modelo de RLM debe definir:

- El subconjunto A de coeficientes de regresión del que se quiere obtener la SSextra.
- El subconjunto B de coeficientes de regresión que acompañan al subconjunto A en el modelo.

Se debe cumplir que $A \cup B$ debe estar incluido en el conjunto de todos los coeficientes de regresión del modelo, y $A \cap B = \phi$.

Así, **una suma de cuadrados extra para el subconjunto A dado un subconjunto B se denota y calcula como**:

$$\text{SSR}\,(\boldsymbol{A}|\boldsymbol{B}) = \text{SSR}\,(\boldsymbol{A} \cup \boldsymbol{B}) - \text{SSR}\,(\boldsymbol{B}) = \text{SSE}\,(\boldsymbol{B}) - \text{SSE}\,(\boldsymbol{A} \cup \boldsymbol{B})$$

Ejemplos de sumas de cuadrados extra

Suponga un modelo de regresión múltiple de una respuesta Y en función de tres variables predictoras X_1, X_2, X_3, esto es,

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$$

Veamos algunas de las posibles sumas de cuadrados extras:

- $$\begin{aligned} \mathrm{SSR}(\beta_1 \mid \beta_0, \beta_2, \beta_3) &= \mathrm{SSR}\left(\beta_0, \beta_1, \beta_2, \beta_3\right) - \mathrm{SSR}\left(\beta_0, \beta_2, \beta_3\right) \\ &= \mathrm{SSE}\left(\beta_0, \beta_2, \beta_3\right) - \mathrm{SSE}\left(\beta_0, \beta_1, \beta_2, \beta_3\right), \end{aligned}$$

la cual representa la suma de cuadrados extra de β_1 dado que β_0, β_2 y β_3 ya están presentes en el modelo de regresión.

- $$\begin{aligned}\mathrm{SSR}(\beta_1, \beta_2 \mid \beta_0, \beta_3) &= \mathrm{SSR}\left(\beta_0, \beta_1, \beta_2, \beta_3\right) - \mathrm{SSR}\left(\beta_0, \beta_3\right) \\ &= \mathrm{SSE}\left(\beta_0, \beta_3\right) - \mathrm{SSE}\left(\beta_0, \beta_1, \beta_2, \beta_3\right),\end{aligned}$$

  la cual representa la suma de cuadrados extra de β_1 y β_2 dado que β_0 y β_3 ya están presentes en el modelo de regresión.

- $$\begin{aligned}\mathrm{SSR}(\beta_1 \mid \beta_0, \beta_3) &= \mathrm{SSR}\left(\beta_0, \beta_1, \beta_3\right) - \mathrm{SSR}\left(\beta_0, \beta_3\right) \\ &= \mathrm{SSE}\left(\beta_0, \beta_3\right) - \mathrm{SSE}\left(\beta_0, \beta_1, \beta_3\right)\end{aligned}$$

  la cual la suma de cuadrados extras de β_1 dado que β_0 y β_3 ya están presentes en el modelo de regresión.

  (**Tarea:** defina la suma de cuadrados extra $\mathrm{SSR}(\beta_2 \mid \beta_0, \beta_1)$)

Ejemplo: suponga que para el modelo

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \varepsilon,$$

se desea probar si el subconjunto de coeficientes de regresión β_1, β_2 y β_5 es significativo en el modelo, esto es, se desea probar que:

$$\begin{aligned} H_0 &: \ \beta_1 = \beta_2 = \beta_5 = 0 \\ H_1 &: \ \text{Algún } \beta_j \neq 0, \quad j = 1, 2, 5. \end{aligned}$$

Para este tipo de pruebas se requiere calcular las sumas de cuadrados extra asociada al subconjunto de los coeficientes de regresión de $A = \{\beta_1, \beta_2, \beta_5\}$ **dado el subconjunto de coeficientes restante** $B = \{\beta_0, \beta_3, \beta_4\}$.

Esto es,

$$\begin{aligned}\mathrm{SSR}(\beta_1, \beta_2, \beta_5 \mid \beta_0, \beta_3, \beta_4) \\ &= \mathrm{SSR}\,(\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5) - \mathrm{SSR}\,(\beta_0, \beta_3, \beta_4) \\ &= \mathrm{SSE}\,(\beta_0, \beta_3, \beta_4) - \mathrm{SSE}\,(\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5)\end{aligned}$$

Note que en este cálculo se deben definir dos modelos:

- **Un modelo completo:** que incluye todos los coeficientes de regresión que se consideran inicialmente en el modelo (el conjunto $A \cup B$). Para el caso de nuestro ejemplo:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \varepsilon.$$

- **Un modelo nulo o reducido:** que se obtiene al aplicar lo establecido en H_0 al modelo completo, es decir, eliminando los coeficientes de regresión en A (quedando los coeficientes de regresión en B). Para el caso de nuestro ejemplo:

$$Y = \beta_0 + \beta_3 X_3 + \beta_4 X_4 + \varepsilon.$$

Al igual que en las sumas de cuadrados vistas en la tabla ANOVA, las sumas de cuadrados extra tienen asociados unos grados de libertad, que en este caso se obtienen como **el tamaño del subconjunto** A **que se está probando**, o equivalentemente como **la diferencia en grados de libertad de la SSR (o SSE) de los dos modelos definidos anteriormente**.

Para el ejemplo:

$$
\begin{aligned}
\text{g.l } \; \text{SSR}(\beta_1, \beta_2, \beta_5 \; | \beta_0, \beta_3, \beta_4) \\
&= \text{g.l } \; \text{SSR}\,(\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5) - \text{g.l } \; \text{SSR}\,(\beta_0, \beta_3, \beta_4) \\
&= 5 - 2 = 3 \quad \rightarrow (k = p - 1) \\
&= \text{g.l } \; \text{SSE}\,(\beta_0, \beta_3, \beta_4) - \text{g.l } \; \text{SSE}\,(\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5) \\
&= (n - 3) - (n - 6) \quad \rightarrow (n - p) \\
&= 3
\end{aligned}
$$

También se define el **cuadrado medio extra** (MSextra) como la razón entre la **suma de cuadrados extra** y sus respectivos grados de libertad. **Para el ejemplo**:

$$\mathrm{MSR}(\beta_1, \beta_2, \beta_5 \mid \beta_0, \beta_3, \beta_4) = \frac{\mathrm{SSR}(\beta_1, \beta_2, \beta_5 \mid \beta_0, \beta_3, \beta_4)}{3}$$

Finalmente, el **estadístico de prueba** es igual a la razón **del cuadrado medio extra sobre la media cuadrática de error del modelo completo**. Para el ejemplo, sería:

$$F_0 = \frac{\mathrm{MSR}(\beta_1, \beta_2, \beta_5 \mid \beta_0, \beta_3, \beta_4)}{\mathrm{MSE}(\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5)}$$

$$= \frac{\mathrm{SSR}(\beta_1, \beta_2, \beta_5 \mid \beta_0, \beta_3, \beta_4)/3}{\mathrm{MSE}^1}$$

A un nivel de significancia α, el criterio de rechazo es $F_0 > f_{1-\alpha;3,n-6}$.

[1] siempre en el denominador estará el MSE del modelo completo.

Otro ejemplo:

En el modelo $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \varepsilon$, para probar la hipótesis:

$$H_0 : \beta_2 = \beta_4 = 0$$
$$H_1 : \text{Algún } \beta_j \neq 0,\ j = 2, 4.$$

se usa como estadístico de prueba a

$$F_0 = \frac{\text{SSR}(\beta_2, \beta_4 \mid \beta_0, \beta_1, \beta_3, \beta_5)/2}{\text{MSE}} \overset{\text{bajo } H_0}{\sim} F_{2,n-6}$$

y con un nivel de significancia α el criterio de rechazo de la hipótesis nula es $F_0 > f_{1-\alpha;\ 2,n-6}$.

Uso de SSextra para la prueba de significancia de un coeficiente individual

En un modelo de RLM con k predictoras, esta prueba establece que:

$$\begin{array}{l} H_0 : \beta_j = 0 \\ H_1 : \beta_j \neq 0 \end{array}, \quad j = 1, 2, \ldots, k,$$

donde $A = \{\beta_j\}$ y $B = \{\beta_0, \beta_1, \beta_2, \ldots, \beta_{j-1}, \beta_{j+1}, \beta_{j+2}, \ldots, \beta_k\}$. Luego, usando SSextra el estadístico de prueba es:

$$F_{j,0} = \frac{\mathrm{SSR}(\beta_j \mid \beta_0, \beta_1, \beta_2, \ldots, \beta_{j-1}, \beta_{j+1}, \beta_{j+2}, \ldots, \beta_k)}{\mathrm{MSE}}.$$

Observe que la SSextra solo tiene un grado de libertad, de forma que es igual al MSextra, y bajo la hipótesis nula $F_{j,0} \sim f_{1,n-p}$, por lo cual, a un nivel de significancia α, el criterio de rechazo de la hipótesis nula es: $F_{j,0} > f_{1-\alpha;\ 1,n-p}$.

La prueba anterior es equivalente a la prueba t definida en una clase anterior. De hecho se puede demostrar que.

- $F_{j,0} = T_{j,0}^2$.
- Si se calculan los valores-P de los dos procedimientos de prueba, se llega a que:

$$\text{vp}_F = P\left(f_{1,n-p} > F_{j,0}\right) \equiv P\left(|t_{n-p}| > |T_{j,0}|\right) = \text{vp}_T$$

Por otro lado, también se puede ver la prueba de significancia de la regresión como un caso particular de una prueba basada en SSextra donde $A = \{\beta_1, \beta_2, \dots, \beta_k\}$ y $B = \{\beta_0\}$.

Prueba de la hipótesis lineal general

Suponga un modelo de RLM con k variables predictoras,
$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k + \varepsilon$, al que llamaremos **modelo completo (FM)**.

En este modelo se tiene una suma de cuadrados de la regresión
$\text{SSR(FM)} = \text{SSR}\left(\beta_0, \beta_1, \ldots, \beta_k\right)$ con $k = p - 1$ g.l y una suma de cuadrados del error
$\text{SSE(FM)} = \text{SSE}\left(\beta_0, \beta_1, \ldots, \beta_k\right)$ con $(n - p)$ g.l.

Considere además una matriz $m \times p$ de constantes $\boldsymbol{L}$, con $r \leq m$ **filas linealmente independientes**. Se puede formular una prueba de hipótesis lineal general como:

$$H_0 : \boldsymbol{L}\underline{\boldsymbol{\beta}} = \underline{\mathbf{0}} \quad \text{vs.} \quad H_1 : \boldsymbol{L}\underline{\boldsymbol{\beta}} \neq \underline{\mathbf{0}},$$

donde, $\underline{\mathbf{0}}$ es un vector de ceros de dimensión $m \times 1$.

$\boldsymbol{L}\underline{\beta} = \underline{\mathbf{0}}$ **es simplemente un sistema de ecuaciones donde se formulan** m **hipótesis que se prueban simultáneamente**.

- Si al modelo completo se le aplica lo establecido en H_0 algunos coeficientes serán iguales a cero y se llega a **un modelo reducido (RM)**, que tiene asociado tanto una suma de cuadrados de la regresión SSR(RM) como una suma de cuadrados del error SSE(RM).
- Para probar la hipótesis se debe definir una **suma de cuadrados debida a la hipótesis** (SSH) que se calcula como **la diferencia entre las sumas de cuadrados de la regresión (o del error) de los modelos completo y reducido**. Esto es,

$$\text{SSH} = \text{SSE(RM)} - \text{SSE(FM)} = \text{SSR(FM)} - \text{SSR(RM)},$$

que tiene tantos grados de libertad como el número r de filas linealmente independientes en $\boldsymbol{L}$. O equivalentemente:

$$r = \text{g.l SSE(RM)} - \text{g.l SSE(FM)} = \text{g.l SSR(FM)} - \text{g.l SSR(RM)}$$

Luego, se define el **cuadrado medio debido a la hipótesis** (MSH) como:

$$\text{MSH} = \frac{\text{SSH}}{r}.$$

Finalmente, se define como estadístico de prueba a la razón **entre el cuadrado medio de la hipótesis y la media cuadrática de error del modelo completo**:

$$F_0 = \frac{\text{MSH}}{\text{MSE}\,(\beta_0, \beta_1, \ldots, \beta_4)} = \frac{\text{SSH}/r}{\text{MSE}} \sim F_{r,n-p}$$

Se puede demostrar que bajo H_0 el estadístico $F_0 \sim F_{r,n-p}$. Lo cual permite a un nivel de significancia α, rechazar H_0 si $F_0 > f_{1-\alpha;\ r,n-p}$.

Veamos, ejemplos de cómo trabaja este procedimiento de prueba.

Ejemplo 1

Suponga un modelo de RLM con $k = 4$ variables predictoras, entonces se puede formular la siguiente prueba de hipótesis:

$$H_0 : \beta_1 = \beta_2, \ \beta_3 = \beta_4 \ \text{ vs. } \ H_1 : \beta_1 \neq \beta_2 \ \text{ ó } \ \beta_3 \neq \beta_4$$

Podemos reescribir la hipótesis nula de la siguiente manera:

$$H_0 : \beta_1 - \beta_2 = 0, \ \beta_3 - \beta_4 = 0,$$

de manera que la hipótesis nula contiene $m = 2$ ecuaciones y se puede escribir como:

$$H_0 : \begin{cases} \beta_1 - \beta_2 = 0 \\ \beta_3 - \beta_4 = 0 \end{cases}$$

que en forma matricial se puede expresar como:

$$H_0 : \begin{bmatrix} 0 & 1 & -1 & 0 & 0 \\ 0 & 0 & 0 & 1 & -1 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

por tanto, se tiene una prueba de hipótesis lineal general, con:

$$\boldsymbol{L} = \begin{bmatrix} 0 & 1 & -1 & 0 & 0 \\ 0 & 0 & 0 & 1 & -1 \end{bmatrix},$$

que tiene $r = 2$ filas linealmente independientes (**observe que una fila no puede escribirse como un múltiplo escalar de la otra**).

El modelo reducido en este caso es:

$$\begin{aligned}\text{RM:}\quad Y &= \beta_0 + \beta_1 (X_1 + X_2) + \beta_3 (X_3 + X_4) + \varepsilon \\ &= \beta_0 + \beta_1 X_{1,2} + \beta_3 X_{3,4} + \varepsilon\end{aligned},$$

donde $X_{1,2} = X_1 + X_2$, y $X_{3,4} = X_3 + X_4$.

En este modelo se tiene una suma de cuadrados del error $\text{SSE(RM)} = \text{SSE}(\beta_0, \beta_1, \beta_3)$ con $(n-3)$ grados de libertad.

Luego, **la SSH se calcula como**:

$$\text{SSH} = \text{SSE(RM)} - \text{SSE(FM)},$$

que tiene 2 grados de libertad, de manera que el cuadrado medio debido a la hipótesis es:

$$\text{MSH} = \frac{\text{SSH}}{2}.$$

Finalmente, **se define como estadístico de prueba a**:

$$F_0 = \frac{\text{MSH}}{\text{MSE}} = \frac{\text{SSH}/2}{\text{MSE}} \sim F_{2,n-5}$$

NOTA: Observe que en el denominador se encuentra la media cuadrática de error (o cuadrado medio de error) del modelo completo que tiene $(n-5)$ grados de libertad.

Bajo H_0 y los supuestos sobre los errores, $F_0 \sim F_{2,n-5}$. Se rechaza para valores grandes de este estadístico, esto es, si $VP = P\left(f_{2,n-5} > F_0\right)$ es pequeño. O bien, si $F_0 > f_{1-\alpha;\ 2,n-5}$, el valor crítico a un nivel de significancia α.

Ejemplo 2

Bajo el mismo modelo de RLM con $k = 4$ considere la siguiente prueba:

$$H_0:\ \beta_1 = \beta_2 = 0,\ \beta_3 = \beta_4 \quad \text{vs.} \quad H_1:\ \beta_1 \neq 0 \ \text{ ó } \ \beta_2 \neq 0 \ \text{ ó } \ \beta_3 \neq \beta_4$$

Como en el ejemplo anterior, también se puede reescribir la hipótesis nula en términos de ecuaciones igualadas a cero:

$$H_0:\ \beta_1 = 0,\ \beta_2 = 0,\ \beta_3 - \beta_4 = 0$$

Luego, en H_0 se tiene un sistema de $m = 3$ ecuaciones que se puede expresar como:

$$H_0: \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & -1 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

por tanto, se tiene una prueba de hipótesis lineal general, con:

$$\boldsymbol{L} = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & -1 \end{bmatrix},$$

que tiene $r = 3$ filas linealmente independientes (compruebe que ninguna de las filas se puede escribir como combinación lineal de las otras dos filas).

Entonces, el modelo nulo es:

$$\begin{aligned} \text{RM:} \quad Y &= \beta_0 + \beta_3 (X_3 + X_4) + \varepsilon \\ &= \beta_0 + \beta_3 X_{3,4} + \varepsilon \end{aligned},$$

donde $X_{3,4} = X_3 + X_4$.

El estadístico de prueba es,

$$F_0 = \frac{\mathrm{SSH}/3}{\mathrm{MSE}} \sim F_{3,n-5}$$

Bajo H_0 y los supuestos sobre los errores, $F_0 \sim F_{3,n-5}$. Se rechaza para valores grandes de este estadístico, esto es, si $\mathrm{VP} = P\left(f_{3,n-5} > F_0\right) < \alpha$, donde α es el nivel de significancia de la prueba. O bien, si $F_0 > f_{1-\alpha;\ 3,n-5}$.

Ejemplo 3

Considere ahora la prueba de significancia del modelo de RLM con $k = 4$ **variables predictoras**:

$$H_0: \ \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0 \quad \text{vs.} \quad H_1: \ \text{Algún } \beta_j \neq 0, \ j = 1, 2, 3, 4.$$

Note que H_0 se puede reescribir como:

$$H_0: \ \beta_1 = 0, \ \beta_2 = 0, \ \beta_3 = 0, \ \beta_4 = 0.$$

En este caso también se puede reformular la hipótesis nula en la forma de una hipótesis lineal general, **considerando las** $m = r = 4$ **ecuaciones linealmente independientes** como sigue:

$$H_0: \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

El modelo reducido es simplemente RM: $Y = \beta_0 + \varepsilon$, donde el intercepto representa la media de la variable respuesta. **Así el estimador de mínimos cuadrados del intercepto es simplemente la media muestral de** Y, es decir, $\widehat{\beta}_0 = \bar{Y}$, por tanto, $\widehat{Y} = \bar{Y}$, y en consecuencia tiene una suma de cuadrados del error igual a la suma de cuadrados totales ($\mathrm{SSE}\,(\beta_0) = \mathrm{SST}$) con $(n-1)$ grados de libertad, mientras que la suma de cuadrados de la regresión es igual a cero ($\mathrm{SSR}\,(\beta_0) = 0$).

Al calcular la SSH en función de la diferencia entre las SSE de los modelos RM y FM, se obtiene:

$$\begin{aligned} \mathrm{SSH} &= \mathrm{SSE}\,(\beta_0) - \mathrm{SSE}\,(\beta_0, \beta_1, \beta_2, \beta_3, \beta_4) \\ &= \mathrm{SST} - \mathrm{SSE}\,(\beta_0, \beta_1, \beta_2, \beta_3, \beta_4) \qquad , \\ &= \mathrm{SSR}\,(\beta_0, \beta_1, \beta_2, \beta_3, \beta_4) = \mathrm{SSR} \end{aligned}$$

con $r = m = k = 4 = p - 1$ grados de libertad, **cuyo MSextra coincide con el MSR del modelo completo**.

Así, **el estadístico de prueba coincide con el visto en la prueba de significancia de la regresión**

$$F_0 = \frac{\text{MSH}}{\text{MSE}} = \frac{\text{SSH}/4}{\text{MSE}} = \frac{\text{SSR}/4}{\text{MSE}} = \frac{\text{MSR}}{\text{MSE}} \sim F_{4,n-5}$$

Por lo tanto, bajo H_0 y los supuestos sobre los errores se cumple que, $F_0 \sim F_{4,n-5}$. Se rechaza para valores grandes de este estadístico, esto es, si $\text{VP} = P\left(f_{4,n-5} > F_0\right) < \alpha$, donde α es el nivel de significancia de la prueba. O bien, si $F_0 > f_{1-\alpha;\ 4,n-5}$.

NOTA: También es posible probar hipótesis lineales generales del tipo $H_0 : \boldsymbol{L}\underline{\boldsymbol{\beta}} = \boldsymbol{c}$ vs. $H_1 : \boldsymbol{L}\underline{\boldsymbol{\beta}} \neq \boldsymbol{c}$, donde $\boldsymbol{c}$ es un vector de constantes arbitrario.