Regresión múltiple y otras técnicas multivariadas

Tarea 10

Rivera Torres Francisco de Jesús
Rodríguez Maya Jorge Daniel
Samayoa Donado Víctor Augusto
Trujillo Barrios Georgina

Mayo 1°, 2019

Ejercicio 1

Enunciar los supuestos del modelo de regresión múltiple.

El modelo de regresión múltiple modela una variable aleatoria Y condicional a un conjunto de variables auxiliares $X_1, \ldots, X_p$ mismas que se asumen no aleatorias tales que

$$Y_i = \mathbf{x}_i^T \beta + \varepsilon_i, \quad i = 1, \ldots, n$$

donde $\mathbf{x}_i^T = (1, x_{1i}, \ldots, x_{pi})$ es la observación i-ésima, con $i = 1, \ldots, n$.

1. Linealidad

$$\mathrm{E}\left(Y_i | x_i\right) = \mathbf{x}_i^T \beta, \quad i = 1, \ldots, n$$

2. Homocedasticidad (varianza constante)

$$\mathrm{V}\left(Y_i | x_i\right) = \sigma^2, \quad i = 1, \ldots, n$$

3. No correlación

$$\mathrm{Cov}\left(Y_i, Y_j | x_i, x_j\right) = 0, \quad i, j = 1, \ldots, n \text{ e } i \neq j$$

Ejercicio 2

Enuncie correctamente el Teorema de Gauss-Markov para el estimador de β en el modelo de regresión múltiple.

En el modelo RLM $\mathbf{Y} = \mathbf{X}\beta + \varepsilon$, bajos las hipótesis:

- $\varepsilon \sim \left(\mathbf{0}, \sigma^2 \mathbf{I}\right)$
- $\mathbf{X}$ es una matriz de rango completo

el estimador de MCO de β es el MELI. Esto es, $\hat{\beta}$ es insesgado para β y si $\tilde{\beta}$ es otro estimador insesgado de β y $\mathbf{v}$ es un vector de dimensión $p+1$ distinto de $\mathbf{0}$, entonces $\mathbf{v}'V\left(\tilde{\beta}\right)\mathbf{v} \geqslant \mathbf{v}'V\left(\hat{\beta}\right)\mathbf{v}$. Lo anterior implica que no es posible encontrar otro estimador de β que siendo lineal e insesgado tenga una varianza menor que el estimador de MCO de β.

Ejercicio 3

Mostrar que el estadístico F utilizado para contrastar las hipótesis

$$H_0 : \beta_1 = \cdots = \beta_p = 0 \quad v.s. \quad H_1 : \beta_i \neq 0, \text{ para alguna } i$$

se puede escribir como

$$F = \frac{R^2(n-p-1)}{p(1-R^2)}$$

donde R^2 es el coeficiente de determinación del modelo.

Demostración. Sabemos que bajo la hipótesis nula (H_0) se tiene que

$$F = \frac{\dfrac{SC_{reg}}{p}}{\dfrac{SC_{error}}{(n-p-1)}}$$

además, el coeficiente de determinación del modelo RLM está dado por:

$$R^2 = \frac{SC_{reg}}{SC_{tot}} = 1 - \frac{SC_{error}}{SC_{tot}}$$

entonces se tiene que:

$$F = \frac{\dfrac{SC_{reg}}{p}}{\dfrac{SC_{error}}{(n-p-1)}} = \frac{\dfrac{SC_{reg}}{p}}{\dfrac{SC_{error}}{(n-p-1)}} \cdot \frac{\dfrac{1}{SC_{tot}}}{\dfrac{1}{SC_{tot}}} = \frac{\dfrac{SC_{reg}}{SC_{tot}} \cdot \dfrac{1}{p}}{\dfrac{SC_{error}}{SC_{tot}} \cdot \dfrac{1}{(n-p-1)}} = \frac{\dfrac{R^2}{p}}{\dfrac{1-R^2}{(n-p-1)}}$$
$$= \frac{R^2(n-p-1)}{p(1-R^2)}$$

□

Ejercicio 4

Suponer que se ha ajsutado un modelo de regresión lineal con $p = 2$ variables explicativas y $n = 25$ observaciones y que los resultados muestran que $R^2 = 0.90$.

Inciso 4.a

Contrastar la hipótesis de significancia de la regresión. Utilizar $\alpha = 0.05$

Del ejercicio 3, sabemos que

$$F = \frac{R^2(n-p-1)}{p(1-R^2)} = \frac{0.90(25-2-1)}{2(1-0.90)} = 99$$

Por otro lado,

```
alpha <- 0.05
f.a <- qf(1 - alpha, df1 = 2, df2 = 22)
```

Es decir, el cuantil de la distribución de referencia es $F_0 = 3.4433568$, el cual es un valor menor a F = 99. Esto implica que se debe rechazar H_0 con un nivel de significancia del $\alpha = 0.05$.

Inciso 4.b

¿Cuál es el mínimo valor de R^2 que nos lleva a concluir que la regresión es significativa?

Sabemos que se rechaza H_0 (la regresión es significativa) si $F_0 \leqslant F$, por tal motivo

$$\begin{aligned} F_0 \leqslant \frac{R^2(n-p-1)}{p(1-R^2)} \quad &\Rightarrow \quad F_0 p(1-R^2) \leqslant R^2(n-p-1) \\ &\Rightarrow \quad F_0 p \leqslant R^2(n-p-1) + F_0 p R^2 \end{aligned}$$

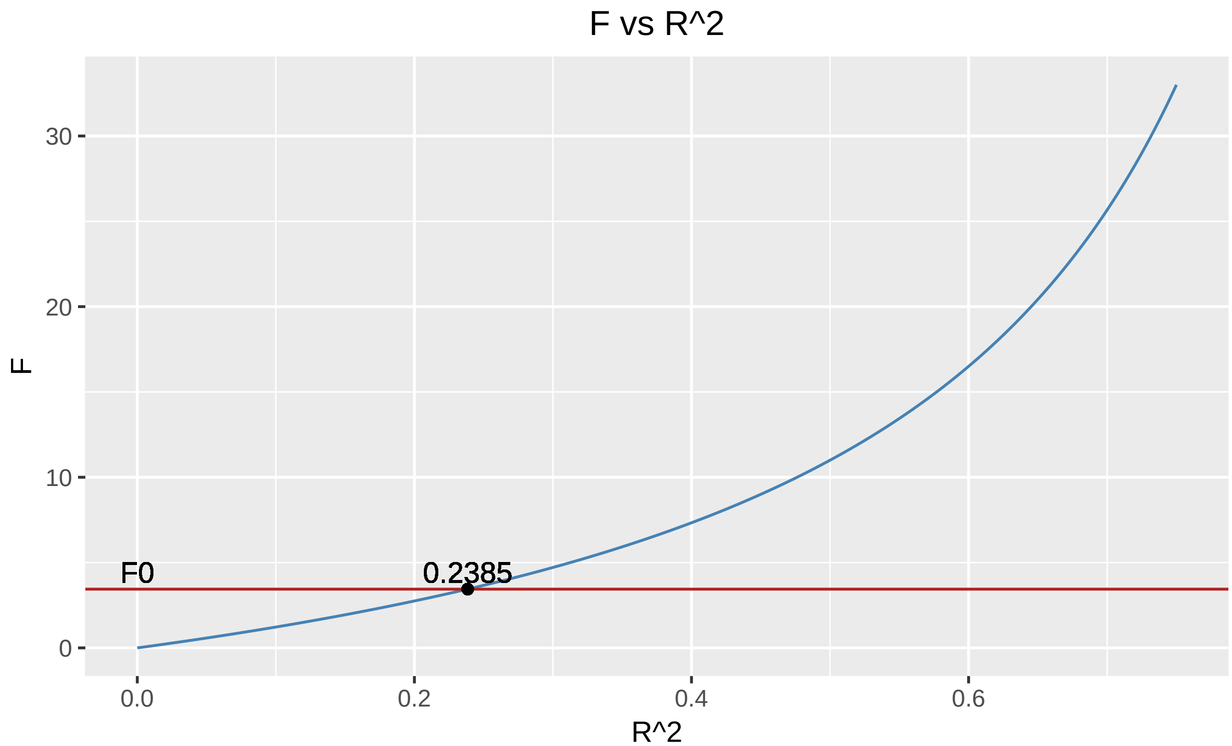
entonces

$$\frac{F_0 p}{n-p-1+F_0 p} \leqslant R^2$$

Por tal motivo, el mínimo valor de R^2 que lleva a concluir que la regresión es significativa es:

$$R^2_{min} = \frac{F_0 p}{n-p-1+F_0 p} = \frac{3.4433568 * 2}{25-2-1+3.4433568 * 2} = 0.2384042$$

Lo anterior lo podemos visualizar en la siguiente gráfica



Se observa que un valor de $R^2 = 0.2384042$ es el mínimo que cumple que $F_0 \leqslant F$.

Ejercicio 5

Suponer que se ajusta el modelo

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

En cada caso, indicar qué matriz **A** y qué vector **b** se deben utilizar para contrastar las siguientes hipótesis

Inciso 5.a

$$\beta_1 = \beta_2 = \beta_3 = \beta_4$$

Inciso 5.b

$$\beta_1 = \beta_2, \beta_3 = \beta_4$$

Inciso 5.c

$$\beta_1 - 2\beta_2 = 4\beta_3, \beta_1 + 2\beta_2 = 0$$

Ejercicio 6

Se ajustó con `R` un modelo lineal para explicar el ingreso por trabajo en los hogares a partir del gasto, un índice de características de la vivienda y un índice de equipamiento de las viviendas (bienes). Los resultados se muestran a continuación:

```
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) 273.09923 3843.94478 0.071 0.9434
Gasto 0.90400 0.02202 41.059 <2e-16 ***
1
Vivienda -25.67979 48.07923 -0.534 0.5938
Bienes 44.90692 17.99172 2.496 0.0132 *
---
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 2524 on 241 degrees of freedom
Multiple R-squared: 0.898,Adjusted R-squared: 0.8967
F-statistic: 707.3 on 3 and 241 DF, p-value: < 2.2e-16
```

Inciso 6.a

¿El modelo es significativo?

Inciso 6.b

Calcular intervalos de confianza simultáneos con el método de Hottelling-Scheffé para $\beta_0, \beta_1, \beta_2$ y β_3.

Inciso 6.c

¿Qué variables son significativas para explicar el ingreso?

Inciso 6.d

¿Qué porcentaje de la varianza del ingreso es explicada por el modelo?

Inciso 6.e

¿Cómo interpretaría la estimación del coeficiente del gasto?

Inciso 6.f

¿Sería mejor ajustar un modelo sin intercepto?

Inciso 6.g

¿Se podría afirmar que el índice de vivienda tiene un efecto negativo en el ingreso?

Inciso 6.h

¿Qué cambios propondría para mejorar el modelo?

Inciso 6.i

Construir la tabla ANOVA con los resultados del ajuste del modelo.