MCO: inferencia y levantamiento de supuestos

Econometría I

Paula Pereda (ppereda@correo.um.edu.uy)
17 de setiembre de 2021

Bondad de ajuste

Estadístico t - ¡de nuevo!

El modelo poblacional lo escribimos de la siguiente manera:

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + \mu$$

Distribución t para estimadores estandarizados:

Bajo los supuestos del modelo lineal clásico (MLC),

$$\left(\hat{\beta}_j - \beta_j\right) / \operatorname{ee}\left(\hat{\beta}_j\right) \sim t_{n-k-1}$$

donde $k+1$ es la cantidad de parámetros desconocidos en el modelo poblacional $y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + \mu$, o sea k parámetros de pendiente y el intercepto β_0.

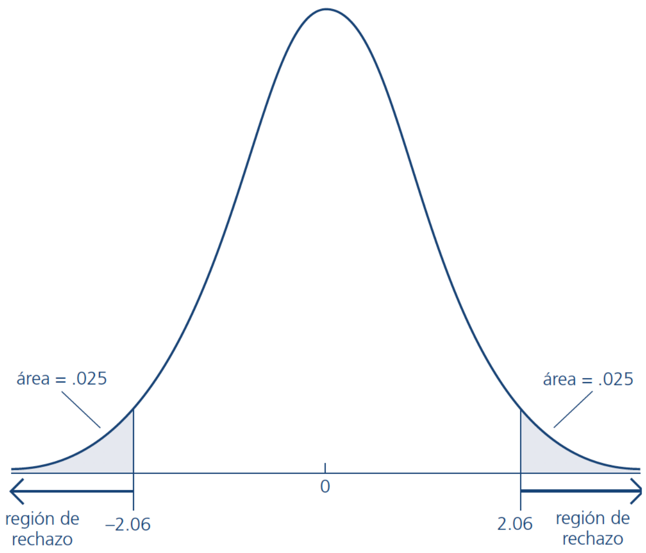
- Testeo: $H_0\ )\ \beta_j = 0$
- Ejemplo: $\textbf{salario}_i = \beta_0 + \beta_1\, \textbf{educación}_i + \beta_2\, \textbf{experiencia}_i + \beta_3\, \textbf{antigüedad}_i + \mu_i$
- ¿Qué significa testear: $H_0\ )\ \beta_j = 0$ en este caso?

Bondad de ajuste

Estadístico de prueba t

$$t_{\hat{\beta}_j} = \frac{\hat{\beta}_j}{\mathrm{ee}\left(\hat{\beta}_j\right)}$$

- A dos colas la regla de rechazo es $\left|t_{\hat{\beta}_j}\right| > c$
- Para encontrar c, definir α, en general 5%
- Suponiendo que $N - k - 1 = 25$



Valor-p para el estadístico t

Alternativamente, podemos calcular el **valor p** que acompaña a nuestro estadístico de prueba, que efectivamente nos da la probabilidad de ver nuestro estadístico de prueba *o nuestro estadístico de prueba más extrema* si la hipótesis nula fuera cierta.

Valores p muy pequeños, generalmente < 0.05, significan que sería poco probable que veamos nuestros resultados si la hipótesis nula fuera realmente cierta; tendemos a rechazar el valor nulo para valores p por debajo de 0.05.

Bondad de ajuste

Valor-p para el estadístico t

- Quitar arbitriariedad del nivel de significación elegido
- El valor p es la probabilidad de obtener valores de la prueba estadística que sean mayores o iguales (o más extremos) que el efectivamente observado si H_0) es cierto
- No rechazar H_0) si $p > \alpha$. En otro caso rechazar. Cuanto más chico es el p más fuerte es el rechazo.
- Recordar que, en principio, queremos rechazar H_0) pues esto implica decir que hay evidencia para decir que β_j es significativamente diferente de cero.

Bondad de ajuste

Estadístico F

- El estadístico F se utiliza para contrastar hipótesis conjuntas sobre los coeficientes de regresión.

- Las fórmulas para el estadístico F están integradas en los paquetes informáticos.

- **Caso de dos restricciones:**

Cuando la hipótesis nula conjunta tiene las dos restricciones de que $\beta_1 = 0$ y $\beta_2 = 0$, el estadístico F combina los dos estadísticos t, t_1 y t_2, mediante la fórmula:

$$F = \frac{1}{2}\left(\frac{t_1^2 + t_2^2 - 2\hat{\rho}_{t_1,t_2} t_1 t_2)}{1 - \hat{\rho}_{t_1,t_2}^2}\right)$$

donde $\hat{\rho}_{t_1,t_2}$ es un estimador de la correlación entre los dos estadísticos t.

Estadístico F

- **Caso general de q restricciones:**

Bajo la hipótesis nula, el estadístico F tiene una distribución muestral que, en muestras grandes, está dada por la distribución $F_{q,\infty}$. Es decir, en muestras grandes, bajo la hipótesis nula el estadístico F se distribuye $F_{q,\infty}$.

Aplicaciones en `R`

Levantamiento de supuestos

Recordemos los supuestos...

Supuesto 1. Linealidad en parámetros.

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + u$$

Supuesto 2. Muestra aleatoria.

$$\{(y_i, x_i) : i = 1, \ldots, n\}$$

son variables aleatorias i.i.d.

Supuesto 3. Exogeneidad estricta.

$$E\left(u \mid x_1, \ldots, x_k\right) = 0$$

Supuesto 4. No multicolinealidad. En la muestra, ninguna de las variables independientes es constante y no hay relaciones lineales exactas entre las variables independientes.

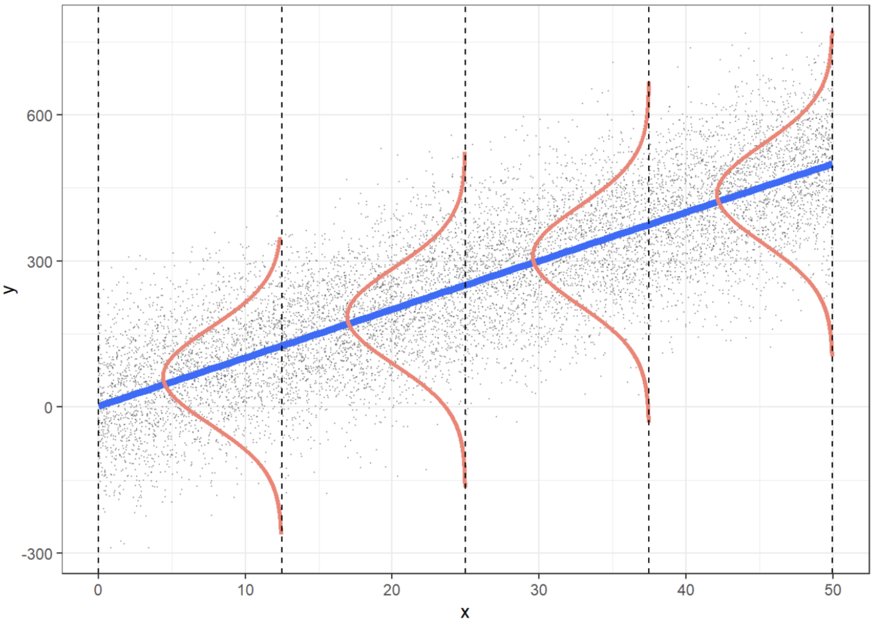
Recordemos los supuestos...

Supuesto 5. Homocedasticidad y ausencia de autocorrelación.

$\mathbf{Var}(u \mid x_1, \ldots, x_k) = \sigma^2$ (homocedasticidad) y $\mathbf{COV}\left(u_i, u_j \mid x_1, \ldots, x_k\right) = 0$

Supuesto 6. Normalidad.

$$u \left| x \sim \mathcal{N}\left(0, \sigma^2\right) \right|$$



Omisión de variable relevante

Omisión de variable relevante

Sesgo de variable omitida Si el regresor está correlacionado con una variable que ha sido omitida en el análisis y ésta determina, en parte, la variable dependiente, el estimador MCO presentará sesgo de variable omitida.

El sesgo de variable omitida se produce cuando se cumplen dos condiciones:

(1) cuando la variable omitida está correlacionada con los regresores incluidos en la regresión y

(2) cuando la variable omitida es un factor determinante de la variable dependiente.

Ejemplos:

- Ejemplo #1: Porcentaje de estudiantes de inglés.
- Ejemplo #2: La hora del día de la prueba.
- Ejemplo #3: Espacio de aparcamiento por alumno.

Omisión de variable relevante

El sesgo de variable omitida significa que el tercer supuesto de mínimos cuadrados, que $E(u \mid x_1, \ldots, x_k) = 0$, no se cumple.

¿Por qué se incumple el supuesto 3?

Recordemos el término de error u_i en el modelo de regresión lineal con un único regresor representa todos los factores, distintos de X_i, que son determinantes de Y_i.

- Si uno de esos otros factores está correlacionado con X_i, esto significa que el término de error (que contiene a este factor) está correlacionado con X_i. En otras palabras, si una variable omitida es un determinante de Y_i, entonces está en el término de error, y si está correlacionada con X_i, entonces el término de error está correlacionado con X_i.

Omisión de variable relevante

¿Por qué se incumple el supuesto 3?

- Debido a que u_i y X_i están correlacionados, la media condicional de u_i dado X_i es distinta de cero. Esta correlación por lo tanto, viola el tercer supuesto de mínimos cuadrados, y la consecuencia es grave: el estimador MCO es sesgado. Este sesgo no desaparece incluso en muestras muy grandes, y el estimador MCO es inconsistente.