Universidad de Chile
Facultad de Ciencias Físicas y Matemáticas
Departamento de ingeniería industrial
IN3242-1 – Estadística

Auxiliar 9

Repaso OLS

Profesor: Raimundo Undurraga
Auxiliares: Brandon Galarza, Camila Jáuregui, Leonardo Meneses, Francisca Monetta, Matías Reyes, Bastián Urzúa, Antonia Villegas.

Parte 1: Comentes

1. Si los supuestos del teorema de Gauss-Markov se cumplen, entonces no es posible hallar un estimador con error cuadrático medio menor que MCO.

   Solución: El teorema de Gauss-Markov prueba que el estimador MCO es el que tiene la **mínima varianza** dentro de la categoría de los estimadores lineales e insesgados. Por otro lado, el error cuadrático medio se descompone como: $\text{ECM}(\hat{\beta}) = \text{Sesgo}(\hat{\beta})^2 + \text{Var}(\hat{\beta})$.

   Con estos antecentes, se concluye que la afirmación es incorrecta. Eventualmente se podrían encontrar estimadores sesgados y/o no lineales que reduzcan ECM.Si un estimador tiene sesgo mayor a cero, su varianza podría ser considerablemente menor a la del estimador MCO.

2. Suponga que quiere construir un modelo OLS para determinar los factores que influyen en el desempeño académico de los estudiantes que cursan segundo año medio. Para lo cual, piensa utilizar, como variables independientes, el nivel de ingresos mensuales recibidos en el hogar del estudiante y su calificación en el Registro Social de Hogares. Con lo cual, se tiene el siguiente modelo:

   $$DesempeñoAcadémico_i = \beta_0 + \beta_1 IngresosMensuales_i + \beta_2 CalificaciónRSH_i + \epsilon_i$$

   Antes de ejecutar sus modelo en R, anticípese comentando qué problemas podría presentar el estimador que construyó. (Hint: investigue cómo se determina el RSH de un hogar)

   Solución: El Registro Social de Hogares es un sistema de información cuyo fin es apoyar los procesos de selección de personas beneficiarias de un conjunto amplio de subsidios y programas sociales. Cuya selección de hogares para estos beneficios, se realiza mediante una caracterización socioeconómica, la cual considera los ingresos efectivos mensuales

que recibe cada hogar junto a las necesidades del grupo familiar. Por ejemplo, integrantes con capacidades diferentes, adultos mayores o menores de 18 años.

Como se menciona, la calificación del RSH considera los ingresos de las familias. Pero la variable $IngresosMensuales_i$ ya es parte del modelo construido. Por lo tanto, habrá un problema de multicolinealidad, debido a que habrá una correlación positiva entre $IngresosMensuales_i$ y $CalificaciónRSH_i$. Con lo cual, los errores estándar de los coeficientes, probablemente, sean grandes, es decir, es más imprecisa la estimación de los $\vec{\beta}$.

3. Para su práctica, lo contratan desde un centro de salud, para que construya un modelo causal que explique el efecto del consumo frecuente de alcohol en los jóvenes, sobre a el estado de salud que tienen cuando superan los 40 años. Así que construye el siguiente modelo:

$$Salud40_i = \beta_0 + \beta_1 ConsumoCig_i + \epsilon_i$$

Donde Salud40 es un indicador de la salud de la persona i cuando esta cumple 40 años. El cual toma valores en el rango [1,10], donde 1 representa un estado de salud precario, mientras que un 10 significa una salud perfecta.

Al correr su modelo en R, tiene que $\beta 1$ = -9.5. Su tutora, una enfermera experimentada, dice estar de acuerdo con tener un efecto negativo asociado al consumo, pero le parece exagerada su magnitud. Entonces, usted piensa que su coeficiente puede estar sesgado. En ese sentido, comente qué está pasando exactamente con su modelo y cómo mejorar el coeficiente asociado.

Solución: Tiene sentido pensar que, efectivamente, la estimación esté sesgada. Y, en particular, se trate de un sesgo de variable omitida, ya que este tipo de sesgo puede sobreestimar o subestimar los coeficientes. Para identificarlo, hay que pensar qué variable o factor se correlaciona con el consumo de cigarrillos. Por ejemplo, pensemos en si la persona sufre de ansiedad o si es propensa a sufrir vicios o adicciones. En este caso, trabajemos con la variable de Ansiedad.

Podría ser razonable pensar que los fumadores más ansiosos tengan un mayor consumo de cigarrillos. Por lo tanto, Cov(ConsumoCig, Ansiedad) > 0. Mientras que las personas que sufren ansiedad pueden verse más perjudicadas en su estado de salud al futuro. Por lo tanto, si consideramos un nuevo modelo:

$$Salud40_i = \beta_0 + \beta_1 ConsumoCig_i + \beta_2 Ansiedad_i + \epsilon_i$$

Tendríamos que $\beta_2 < 0$ y:

$$\hat{\beta_1} = \beta_1 + \beta_2 \frac{Cov(ConsumoCig, Ansiedad)}{Var(ConsumoCig)}$$

Donde el sesgo de variable omitida es:

$$\beta_2 \frac{Cov(ConsumoCig, Ansiedad)}{Var(ConsumoCig)} < 0$$

Por lo tanto este es el sesgo que est´a exagerando el valor de $\hat{\beta}_1$, haciendo que no se aproxime a su valor real β_1.

Parte 2: Diagnóstico de OLS

La ecuación de Mincer, que busca explicar el retorno a la educación (es decir, cuánto más rentable es cada año adicional de estudio) se usa y define tradicionalmente como:

$$ln(ingresoporhora_i) = \beta_0 + \beta_1 escolaridad_i + \beta 2 experienciap_i + \beta 3 experienciap_i^2 + \mu_i$$

Donde experienciap representa la experiencia potencial del individuo i. A partir de la Ecuación de Mincer, realice lo siguiente:

- Estudie si agregar un término cúbico de experiencia potencial mejora el modelo propuesto. Para las siguientes partes, trabaje con el modelo elegido.

  Solución: ejecutamos los modelos de regresión lineal con la ecuación de Mincer y aquel que se le agrega un término cúbico de la Experiencia Potencial.

  Vemos que el R^2 aumenta, pero el error de la experiencia potencial cuadrada también aumenta. Aunque estas variaciones son mínimas. Entonces el modelo más adecuado no necesariamente es uno u otro. El hecho de que el R^2 aumente no necesariamente implica que un modelo sea mejor que el otro, especialmente si la diferencia en el R^2 es mínima entre los modelos.

  Además, como se ha observado, el cambio en el error asociado a la variable de ExperienciaPotencial cuadrada también puede ser un factor a considerar. En la comparación entre modelos, es fundamental tener en cuenta varios aspectos y no depender solamente de una métrica específica, como el R^2. Aquí hay algunas consideraciones adicionales: complejidad del modelo, interpretación, capacidad predictiva, sesgo-varianza.

- Estudie si el modelo elegido presenta heterocedasticidad.

  Solución: Se observa que la varianza es pequeña en cierta región y mucho mayor en otros sitios. O sea, no es constante.

  Para tener más seguridad, comprobémoslo con un test de hipótesis. Un test de hipótesis que sirve para saber si el modelo presenta heterocedasticidad es el Test de Breusch Pagan, cuya formulación es:

$$H_0 : \text{La varianza es constante}$$
$$H_A : \text{La varianza no es constante}$$

Este test compara la suma de los residuos al cuadrado del modelo con una versión transformada de las variables independientes para verificar si la varianza de los residuos es constante o no.

Interpretación de resultados:

- La hipótesis nula del test de Breusch Pagan es que no hay heterocedasticidad en los residuos (es decir, la varianza de los errores es constante).
- Si el valor p (p-value) asociado al test es menor que un nivel de significancia elegido (por ejemplo, 0.05), se rechaza la hipótesis nula, lo que indica la presencia de heterocedasticidad en los residuos del modelo.

El último término Prob > Chi2 es el p-valor correspondiente al test, el cual,en este caso es aproximable a cero ($1.17212e-48 \approx 0$). Por lo tanto, a un 99 % de significancia, se rechaza que la varianza es constante. O sea, hay heterocedasticidad, y no se cumple el supuesto de homocedasticidad de OLS.

- Estudie si el modelo presenta multicolinealidad.

  Solución: Existe una alta correlación entre las variables de experiencia potencial, lo cual es esperable debido a que una se puede construir a partir de una transformación no lineal de la otra. Además, la escolaridad también tiene una alta correlación con las variables de experiencia potencial, debido a que la experiencia potencial se construye usando los años de escolaridad.

- Estudie si el modelo presenta endogeneidad.

  Solución: Para analizar la heterogeneidad se debe pensar si la escolaridad está correlacionada con alguna otra variable que no se incluye en el modelo de Mincer, ¿con qué se relaciona la escolaridad? Del mismo modo, ¿con qué se relaciona la experiencia de una persona?

- De acuerdo a sus resultados, responda: ¿Qué tan adecuado es el modelo de regresión para explicar el salario obtenido por una persona?

  Solución: El modelo presenta un fuerte problema de multicolinealidad, por lo tanto, las estimaciones pueden no ser del todo certeras. Y también presenta heterocedasticidad y se sospechan problemas de endogeneidad. Por lo tanto, no cumple el teorema de Gauss Markov, entonces este modelo no representa al mejor estimador lineal insesgado (MELI).

  Entonces, puede que este modelo no sea tan adecuado para explicar perfectamente la relación que existe entre los ingresos con la experiencia y la escolaridad.

Parte 3: Inferencia Causal

El gobierno está evaluando el impacto que podría causar un nuevo programa social que llaman "Escoge ser saludable", el cual buscaría inculcar los hábitos de vida sana en los niños, para así reducir los niveles de obesidad en los menores. Dicha campaña consta en impartir talleres deportivos a estos niños, y otorgarles una canasta mensual de alimentos saludables.

Para lo cual, disponen del siguiente modelo de regresión lineal:

$$IMC_i = \beta_0 + \beta_1 Inscripción_i + \epsilon_i$$

Donde el IMC_i es es Índice de Masa Corporal del niño i. E $Inscripción_i$ es una variable binaria que indica si el niño i se inscribe en el programa.

Responda:

1. Si se decide que la inscripción será voluntaria, o sea, que cada apoderado del niño i se autoselecciona para participar en el programa, ¿a qué problema nos enfrentar´ıamos? Justifique.

   Solución: El efecto de la inscripción sobre el IMC está dado por:

   $$\hat{\beta}_1 = \mathbb{E}[IMC_i|Inscripción_i = 1] - \mathbb{E}[IMC_i|Inscripción_i = 0]$$
   $$\hat{\beta}_1 = \beta_0 + \beta_1 + \mathbb{E}[\epsilon_i|Inscripción_i = 1] - (\beta_0 + \mathbb{E}[\epsilon_i|Inscripción_i = 0])$$
   $$\hat{\beta}_1 = \beta_1 + \mathbb{E}[\epsilon_i|Inscripción_i = 1] - \mathbb{E}[\epsilon_i|Inscripción_i = 0]$$

   A priori, no sabemos si las inscripciones están relacionadas con factores o variables que no estamos considerando en el modelo, es por eso que no podemos cancelar las esperanzas de los residuos. Para saber cómo tratar con ellos, pensemos en qué pasa cuando el programa es voluntario. En este caso, tiene sentido pensar que algunas personas se sientan más incentivadas o animadas a participar que otras, ¿por qué? ¿Qué las puede animar a inscribirse?

   Pensemos que el programa consta en brindar canastas de alimentos gratuitos a las personas, entonces quienes tengan más necesidades económicas se verán con mayores incentivos a inscribirse. O sea, la inscripción estaría dependiendo del nivel socioeconómico de la persona. Por lo tanto:

   $$\mathbb{E}[\epsilon_i|Inscripción_i = 1] - \mathbb{E}[\epsilon_i|Inscripción_i = 0] \neq 0$$

   Ya que los grupos se están diferenciando por factores no observados. A este término se le denomina sesgo de selección. Con lo cual $\hat{\beta}_1 = \beta_1$, ya que el estimador está siendo sesgado por el término anterior.

2. Explique por qué una asignación aleatoria de los niños al programa solucionaría el problema anterior. O sea, el gobierno decide, al azar, qué niños formarán un grupo de tratamiento y otro de control.

   Solución: En este caso, la inscripción es totalmente exógena, ya que la realiza el mismo gobierno de manera completamente aleatoria. Por lo tanto, la variable de inscripción ya no depende de los incentivos de los apoderados, ni de otra variable que no se esté considerando. Es decir, existe independencia entre los residuos y los distintos valores que tome la variable de inscripción. O sea, $\mathbb{E}[\epsilon_i|Inscripción_i = 1] = \mathbb{E}[\epsilon_i|Inscripción_i = 0] = E[\epsilon i]$

   Reemplazando, tenemos que:

$$\hat{\beta}_1 = \beta_1 + \mathbb{E}[\epsilon_i|Inscripción_i = 1] - \mathbb{E}[\epsilon_i|Inscripción_i = 0]$$
$$\hat{\beta}_1 = \beta_1 + \mathbb{E}[\epsilon_i] - E[\epsilon_i]$$
$$\hat{\beta}_1 = \beta_1$$

   Por lo tanto, la estimación del efecto coincide con el parámetro real, lo que asegura un efecto causal.