Análisis Econométrico

Práctica I

05 June 2025

1. Considere la base de datos Base ENCFT 20161 - 20164.xlsx que contiene información sobre el lado de la oferta del mercado de trabajo.

a) Importe la información a R, creando las bases de datos correspondientes.

b) Adjunte el módulo de MIEMBROS y al de OCUPACION creando una variable id que concatene la información de la vivienda, el hogar y miembro que le permita tener en una misma base de datos las características de los individuos y las variables de mercado laboral.

c) Cree la variable EDAD y MUJER donde esta última es igual a uno si el individuo es mujer.

d) Usando histogramas, grafique la distribución de la edad por género.

e) Cree la variable salario por hora (W) utilizando la información de ingreso laboral en la ocupación principal. Tome en cuenta que el ingreso reportado en la base de datos no está necesariamente en la misma escala para todos los individuos. Es decir, para algunos el salario es por hora, pero para otros es por mes, etc.

f) Muestre en una tabla la distribución de edad por percentil de ingreso. En particular, reporte el 5, 25, 50, 75, 95.

2. Para el caso del modelo de regresión lineal, muestre las propiedades de muestra finita del estimador de mínimos cuadrados.

3. Para el modelo de regresión lineal

$$y = \alpha + \beta x + \varepsilon$$

a) Muestre que las ecuaciones normales de mínimos cuadrados implica que $\sum_i e_i = 0$ y que $\sum_i x_i e_i = 0$.

b) Muestre que la solución para el término constante es: $a = \bar{y} - b\bar{x}$.

c) Muestre que la solución para b es $b = [\sum_i^n (x_i - \bar{x})(y_i - \bar{y})]/[\sum_i^n (x_i - \bar{x})^2]$

d) Muestre que estos dos valores son los que minimizan la suma de cuadrados mostrando que los elementos de la diagonal de la matriz de segundas derivadas de la suma de cuadrados respecto a los parámetros son ambas positivas y que su determinante es $4n[(\sum_i^n x_i^2) - n\bar{x}^2] = 4n[\sum_{i=1}^n (x_i - \bar{x})^2]$, positivo a menos que todos los valores de x sean los mismos.

4. Suponga que b es el vector de coeficientes de mínumos cuadrados en la regresión de y sobre X y que c es cualquier vector $K \times x1$. Demuestre que la diferencia entre las dos suma de residuos es:

$$(y - Xc)'(y - Xc) - (y - Xb)'(y - Xb) = (c - b)'X'X(c - b)$$

pruebe que esta diferencia es positiva.

5. Suponga que el modelo verdadero es:

$$Y = X_1\beta_1 + X_2\beta_2 + \varepsilon$$

a) Muestre qué sucede con el estimador de mínimos cuadrados si decide ignorar la variable X_2 del modelo.

b) Muestre qué sucede si decide estimar $Y = X_1\beta_1 + X_2\beta_2 + X_3\beta_3 + u$ en lugar del modelo verdadero.

6. Los datos en el archivo KoopandTobias2004.xls son una extracción de 15 observaciones de una muestra de 2,178 individuos de un conjunto de variables. Sea X_1 igual a la constante, educación, experiencia y habilidad. Mientras que sea X_2 igual a la eduación de la madre, la educación del padre, y el número de hermanos. Sea y el salario.

a) Compute los coeficientes de MCO en la regresión de y sobre X_1. Reporte los coeficientes.

b) Compute los coeficientes de MCO en la regresión de y sobre X_1 y X_2. Reporte los coeficientes.

c) Regrese cada una de las tres variables en X_2 sobre todas las variables en X_1. Estas nuevas variables denótenlas como X_2^*. ¿Cuáles son las medias muestrales de estas variables? Explique el resultado.

d) Compute $R^2 = 1 - \frac{e'e}{y'M_0y}$ para la regresión de y sobre X_1 y X_2. Repita el cómputo para el caso en la que el término constante es omitido de X_1. ¿Qué sucede con R^2?

e) Compute el R^2 para la regresión con todas las variables incluyendo el término constante. Interprete los resultados.

7. Suponga que usted tiene dos estimadores insegados e independientes del mismo parámetro θ, por ejemplo $\hat{\theta}_1$ y $\hat{\theta}_2$, con varianzas diferentes ν_1 y ν_2. ¿Qué combinación lineal $\hat{\theta} = c_1\hat{\theta}_1 + c_2\hat{\theta}_2$ es el estimador insesgado de varianza mínima.

8. Suponga que el modelo de regresión es $y_i = \alpha + \beta x_i + \varepsilon_i$, donde las perturbaciones ε_i tienen como distribución $f(\varepsilon_i) = (1/\lambda)exp(-\varepsilon_i/\lambda)$, $\varepsilon_i \geq 0$. En este modelo se asume que las perturbaciones son no negativas. Note que las perturbaciones tienenn $E[\varepsilon_i|x_i] = \lambda$ y $Var[\varepsilon_i|x_i] = \lambda^2$. Muestre que la pendiente obtenida por mínimos cuadrados es insesgada, pero el intercepto está sesgado.

9. Considere el siguiente modelo de función de producción para un conjunto de N industrias.

$$y_i = A_i K_i^{\alpha} L_i^{\beta}$$

$$A_i = B\exp(\varepsilon_i)$$

Donde y_i es el nivel de producción en la industria i, A es la productividad en la industria que depende de la productividad agregada(B) común a todos los sectores y de un componente idiosincrático asociado a la industria. Asimismo, K y L son el factor capital y trabajo, respectivamente.

a) Tal como está formulado ¿por qué no es posible estimar este modelo por mínimos cuadrados ordinarios?

b) Aplique la transformación apropiada para que este modelo sea estimable por MCO y formule el nuevo modelo econométrico.

Importante: *Las bases de datos de esta sección de la práctica están en el paquete "wooldridge". Si no lo tiene instalado use: install.package("wooldridge") y después lo carga: library(wooldridge). Si ya lo tiene instalado, solo lo carga. Para llamar una base de datos: data("nombre de la base de datos"). Por ejemplo, para el ejercicio 3: data(k401k).*

10) La base de datos **k401k** es un subconjunto de los datos analizados por Papke (1995) para estudiar la relación entre la participación en un plan de pensión y la generosidad del plan. La variable *prate* es el porcentaje de trabajadores que están inscritos en el plan y que tienen cuenta activa; esta es la variable que se quiere explicar. La medida de la generosidad es la tasa de contribución (de la empresa) al plan, *mrate*. Esta variable es la cantidad promedio con la que la empresa contribuye al plan de cada trabajador por cada peso que aporte el trabajador. Por ejemplo, si $mrate = 0.50$, entonces a una contribución de 1 del trabajador corresponde una contribución de 50 centavos de la empresa.

i) Encuentre el promedio de la tasa de participación y el promedio de la tasa de contribución para la muestra.

ii) Ahora, estime el modelo

$$prate = \beta_0 + \beta_1 mrate + u$$

y de los resultados, el tamaño de la muestra y R-cuadrada.

iii) Interprete el intercepto de la ecuación. Interprete también el coeficiente de *mrate*.

iv) Determine la *prate* que se predice para $mrate = 3.5$. ¿Es razonable esta predicción? Explique qué ocurre aquí.

v) ¿Qué tanto de la variación en prate es explicada por *mrate*? En su opinión, ¿es mucho?

11) Un problema de interés para los funcionarios de salud (y para otros) es determinar los efectos que el fumar durante el embarazo tiene sobre la salud infantil. Una medida de la salud infantil es el peso al nacer; un peso demasiado bajo puede ubicar al niño en riesgo de contraer varias enfermedades. Ya que es probable que otros factores que afectan el peso al nacer están correlacionados con fumar, deben considerarse. Por ejemplo, un nivel de ingresos más alto en general da como resultado el acceso a mejores cuidados prenatales y a una mejor nutrición de la madre.

Una ecuación que reconoce estos factores es:

$$bwght = \beta_0 + \beta_1 cigs + \beta_2 faminc + u$$

i) ¿Cuál es el signo más probable para β_2?

ii) ¿Cree que *cigs* y $faminc$ están correlacionados? Explique por qué la correlación puede ser positiva o negativa.

iii) Ahora, estime la ecuación con y sin $faminc$ utilizando los datos del archivo **bwght**. Dé los resultados en forma de ecuación incluyendo el tamaño de la muestra y la R cuadrada. Explique sus resultados enfocándose en si el añadir $faminc$ modifica de manera sustancial el efecto esperado de *cigs* sobre *bwght*.

12) El modelo siguiente puede usarse para estudiar si los gastos de campaña afectan los resultados de las elecciones:

$$voteA = \beta_0 + \beta_1 expendA + \beta_2 expendB + \beta_3 prtysstrA + u$$

donde $voteA$ es el porcentaje de votos recibidos por el candidato A, $expendA$ y $expendB$ son los gastos de campaña del candidato A y del candidato B y $prtystrA$ es una medida de la fortaleza del partido del candidato A (el porcentaje de votos que obtuvo el partido de A en la elección presidencial más reciente).

i) ¿Cuál es la interpretación de β_1?

ii) Use los datos en **vote1** y Estime el modelo e interprete los coeficientes y la bonda de ajuste.

13) Para este ejercicio emplee los datos del archivo **wage2**.

Considere la ecuación estándar para salario

$$wage = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 tenure + u$$

i) Interprete los coeficientes del modelo.

ii) Estime el modelo e interprete los coeficientes estimados.

iii) Establezca la hipótesis nula de que un año más de experiencia en la fuerza de trabajo general tiene el mismo efecto sobre $log(wage)$ que un año más de antigüedad en el empleo actual.

iv) Al nivel de significancia de 5% pruebe la hipótesis nula del inciso iii) contra la alternativa de dos colas. ¿Qué puede concluir?

v) Estime nuevamente el modelo, pero ahora la variable dependiente está en logaritmo (natural) e interprete los coeficientes.

$$\log(wage) = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 tenure + u$$

14) Utilice los datos **hprice1** para estimar el siguiente modelo:

$$price = \beta_0 + \beta_1 sqrft + \beta_2 bdrms + u$$

donde $price$ es el precio de las casas en miles de dólares.

i) Presente los resultados de la estimación.

ii) Cuál es el incremento en el precio estimado para una casa con una habitación más, manteniendo constante la superficie en pies cuadrados ($sqrft$)?

iii) Cuál es el incremento en el precio estimado para una casa con una habitación adicional de 140 pies cuadrados? Compare con su respuesta del inciso ii).

iv) Qué porcentaje de la variación en el precio se explica por la extensión en pies cuadrados y el número de habitaciones?

v) La primera casa de la muestra tienen $sqrft = 2,348$ y $bdrms = 4$. Determine el precio de venta estimado para esta con la línea de regresión de MCO.

vi) El precio de venta de la primera casa fue de $\$300,000$ (así que $price = 300$). Determine el residual para esta casa. Sugiere esto que el comprador pagó de más o de menos por la casa?

vii) Estime el modelo con el logaritmo de *price*. Cómo cambia la interpretación de los parámetros del modelo. En especial, interprete correctamente la constante de esta especificación.

15) La base de datos **k401Ksubs** contiene información acerca de la riqueza financiera neta ($nettfa$), edad de la persona entrevistada (age), ingreso familiar (inc), tamaño de la familia ($fsize$), y participación en los planes de pensiones. Las variables riqueza e ingreso están dadas en miles de dólares. Para esta ecuación, emplee solo los datos de hogares de solo una persona ($fsize = 1$).

i) ¿Cuántos hogares de una sola persona hay en la base de datos?

ii) Use MCO para estimar el modelo:

$$nettfa = \beta_0 + \beta_1 inc + \beta_2 age + u$$

Y de los resultados empleando el formato habitual. Compruebe que solo utiliza los hogares de una sola persona que hay en la muestra. Interprete los coeficientes de pendiente. ¿Hay algo que sorprenda en las estimaciones de pendiente?

iii) ¿Tiene algún significado interesante el intercepto en la regresión del inciso ii)? Explique.

iv) Encuentre el valor-p de la prueba $H_0 : \beta_2 = 1$ contra $H_1 : \beta_2 < 1$. ¿Rechaza usted H_0 al nivel de significancia de 1%?

v) Si realiza una regresión simple de $nettfa$ sobre inc, ¿es el coeficiente estimado de inc muy diferente al estimado en el inciso ii)? Justifique su respuesta.

vi) Considere el modelo:

$$nettfa = \beta_0 + \beta_1 inc + \beta_2 age + \beta_3 age^2 + u$$

¿Cuál es la interpretación literal de β_2? ¿Tiene mucho interés en si misma?

vii) Estime el modelo anterior y de los resultados de manera habitual. ¿Le preocupa que el coeficiente de *age* (edad) sea negativo? Explique.

viii) Dado que las personas más jóvenes de la muestra tienen 25 años, es razonable pensar que, dado un determinado nivel de ingreso, la menor cantidad promedio de activo financiero neto es a la edad de 25 años. Recuerde que el efecto parcial de age sobre $nettfa$ es $\beta_2 + 2\beta_3 age$, de manera que este efecto parcial a la edad de 25 años es $\beta_2 + 2\beta_3(25) = \beta_2 + 50\beta_3$;llámese a esto θ_2. Determine θ_2 y obtenga el valor–p de dos colas para probar $H_0 : \theta_2 = 0$. Debe concluir que θ_2 es pequeño y estadísticamente muy poco significativo. [Sugerencia: una manera de hacer esto es estimar el modelo $nettfa = \alpha_0 + \beta_1 inc + \theta_2 age + \beta_3(age - 25)^2 + u$,donde el intercepto α_0 es diferente a β_0].

16) Para el siguiente ejercicio utilice los datos por hogar disponibles en la Encuesta Nacional de Gastos e Ingresos de los Hogares de 2018 (ENGIH2018).Los datos los encuentra la siguiente dirección: https: //www.bancentral.gov.do/a/d/4796-engih-2018, sección bases de datos.Note que tendra que combinar módulos, filtrar observaciones y crear las variables observadas relevantes.

a) Para cada grupo de gasto estime un modelo que relacione el gasto percápita y el ingreso percápita de los hogares dominicanos. Utilice como controles las variables que considere necesarias (usualmente, características de los hogares). Construya distintas definiciones de ingreso del hogar, incluya en el modelo variables relacionadas a los subsidios y ayuda que reciba el hogar por concepto de transferencias del gobierno.

b) Elabore un cuadro para cada modelo comparar las elasticidades de cada uno de los tipos de gasto respecto al ingreso