Maestría en Economía Aplicada
Teoría Econométrica
Práctica I

1. Considere la base de datos ENFT_2000 en Excel que contiene información del lado de la oferta del mercado de trabajo.

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

b. Cree la variable años de educación (EDUC) utilizando las variables de la encuesta que contiene información sobre la educación de las personas.

c. Cree las variables EDAD y MUJER donde esta última es igual a uno si el individuo es mujer.

d. Construya dos histogramas de la distribución de educación por edad, uno para cada 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 las mismas escalas para todos los individuos. Es decir, para algunos es salario por hora, pero para otros es salario por mes.

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

2.Considere la base de datos EEPS_hogares, la cual contiene información sobre las condiciones socioeconómicas de los hogares en condiciones de pobreza.

a. Importe esta información a STATA.

b. Resuma en una tabla la distribución de los hogares por tipo de vivienda de acuerdo a si la *"vivienda está conectada a una red de distribución de agua"*, *"vivienda está conectada a una red de distribución de energía eléctrica"*, como purifican el agua, si dispone los siguientes electrodomésticos: televisor, radio y microondas.

3. Considere la base datos EEPS_personas. Tabule la informacion relativa al uso de vacunas de distinto tipo. Presente la información por nivel educativo, asegurándose los promedios sean comparables.

4. 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: $a = \overline{y} - b\overline{x}$

c. Muestre que la solución para b es $b = \left[\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})\right] / \left[\sum_{i=1}^{n} (x_i - \overline{x})^2\right]$

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 seman de cuadrados respecto a los parametros son ambas positivas y que si determinante es $4n\left[\left(\sum_{i=1}^{n} x_i^2\right) - n\overline{x}^2\right] = 4n\left[\sum_{i=1}^{n} (x_i - \overline{x})^2\right]$, positivo a menos que todos los valores de x sean los mismos.

5. Suponga que $\mathbf{b}$ es el vector de coeficientes de mínimos cuadrados en la regresión de $\mathbf{y}$ sobre $\mathbf{X}$ y que $\mathbf{c}$ es cualquier vector $K \times 1$. Demuestre que la diferencia entre las dos sumas de residuos es

$$(\mathbf{y} - \mathbf{Xc})'(\mathbf{y} - \mathbf{Xc}) - (\mathbf{y} - \mathbf{Xb})'(\mathbf{y} - \mathbf{Xb}) = (\mathbf{c} - \mathbf{b})'\mathbf{X}'\mathbf{X}(\mathbf{c} - \mathbf{b})$$

pruebe que esta diferencia es positiva.

6. ¿Cuál sería el resultado del producto entre las matrices $\mathbf{M_1 M}$?

donde

$$\mathbf{M} = \left(\mathbf{I} - \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-\mathbf{1}}\mathbf{X}'\right)$$

$$\mathbf{M_1} = \left(\mathbf{I} - \mathbf{X_1}\left(\mathbf{X_1'}\mathbf{X_1}\right)^{-\mathbf{1}}\mathbf{X_1'}\right)$$

7. Considere una base de datos consistente con n observaciones en $\mathbf{X_n}$ y $\mathbf{y_n}$. El estimador de MCO basado en estas n observaciones es $\mathbf{b}_n = (X_n'X_n)^{-1}X_n'y_n$. Suponga que ahora está disponible otra observación, $\mathbf{x}_s$ y y_s. Muestre que el estimador computado utilizando esta observación adicional es:

$$\mathbf{b}_{n,s} = \mathbf{b}_n + \frac{1}{1 + \mathbf{x}_s'\left(\mathbf{X}_n'\mathbf{X}_n\right)^{-1}\mathbf{x}_s}\left(\mathbf{X}_n'\mathbf{X}_n\right)^{-1}\mathbf{x}_s\left(y_s - \mathbf{x}_s'\mathbf{b}_n\right)$$

note que el útimo término es e_s, el residuo de la predicción de $y_s$utilizando los coeficientes basados en $\mathbf{X}_n$ y $\mathbf{b}_n$. Concluya que el nuevo dato cambia los resultados de MCO solo si la nueva observación sobre y no puede ser perfectamente predicha utilizando la información ya disponible.

8. Una estrategia común para tratar el caso en el que falta una o más observaciones de una o más variables es llenar dichas variables con ceros y añadir una variable al modelo que toma el valor de 1 para esa observación y cero para el resto. Muestre que esta estrategia es equivalente a eliminar esta observación para el cómputo de $\mathbf{b}$ pero no tiene efectos sobre R^2. Considere el caso especial en el que $\mathbf{X}$ contiene una constante y una variable. Muestre que reemplazar valores ausentes de x con la media de todas las obsrvaciones tiene el mismo efecto que añadir una nueva variable.

9. 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_1igual a la constante, educación, experiencia, y habilidad. Mientras que sea X_2igual a la educación de la madre, la educación del padre, y el número de hermano. 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_1y 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 denótelas como X_2^*. ¿Cuáles son las medias muestrales de estas variables? Explique el resultado.

d. Compute $R^2 = 1 - \frac{\mathbf{e'e}}{\mathbf{y'M_0y}}$ para la regresion de y sobre X_1y 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^2para la regresión con todas las variables incluyendo en término constnate. Interprete los resultados.

10. Suponga que usted tiene dos estimadores insesgados e independientes del mismo parámetro θ, por ejemplo $\hat{\theta}_1$ y $\hat{\theta}_2$, con varianzas diferentes ν_1y ν_2. ¿Qué combinación lineal $\hat{\theta} = c_1\hat{\theta}_1 + c_2\hat{\theta}_2$es el estimador insesgado de varianza mínima de θ?

11. Suponga que el modelo de regresión clásico aplica pero que el valor verdadero de la constante es cero. Compare la varianza del estimador de la pendiente del estimador computado sin el término constante con el estimador computado sin la constante (el verdadero).

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

13. El archivo GAS.xls contiene datos sobre el consumo de gasolina entre los años 1953 y 2004. Note que los datos de consumo está como gasto total. Para obtener el gasto percápita, divide GASEXP por GASP por Pop. Las otras variables no necesitan ser transformadas.

a. Compute la regresión múltiple del consumo pércapita de gasolina sobre el ingreso pércapita, el precio de la gasolina, el resto de las demás variables y una tendencia. Reporte los resultados. Son los esperados los signos de los coeficientes estimados?

b. Contraste la hipótesis que al menos en lo relativo a la demanda de gasolina, los consumidores no diferencian entre cambios en los precios de carros nuevos y carros viejos.

14. Una regresión múltiple de y sobre una constante, x_1 y x_2 produce los siguientes resultados.

$$\hat{y} = 4 + 0.4x_1 + 0.9x_2$$

$$R^2 = 8/60$$

$$\mathbf{e}'\mathbf{e} = 520$$

$$n = 29$$

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} 29 & 0 & 0 \\ 0 & 50 & 10 \\ 0 & 10 & 80 \end{bmatrix}$$

Contraste la hipótesis que las pendientes suman 1.

15. Utilizando los resultados en el ejercicio anterior, contraste la hipótesis que la pendiente de x_1 es 0 a través de una regresión restringida comparando las dos sumas de residuos al cuadrado.

16. Pruebe que $E[\mathbf{b}'\mathbf{b}] = \beta'\beta + \sigma^2 \sum_{k=1}^{K} \lambda_k$ donde $\mathbf{b}$ es el estimador de MCO y λ_k es la raíz característica de $\mathbf{X}'\mathbf{X}$.

17. Muestre que el estimador de MCO es un estimador insesgado y de varianza mínima entre todos los estimadores lineales.

18. Considere un conjunto de datos de n observaciones, n_c completas y n_m incompletas para los valores correspondientes de la variable dependiente y_i están ausentes. Los datos de las variables independientes, $\mathbf{x}_i$, están completas para todas las observaciones, $\mathbf{X_c}$ y $\mathbf{X_m}$. Deseamos utilizar los datos para estimar el MRL $\mathbf{y} = \mathbf{X}\beta + \varepsilon$. Considere la siguiente estrategia: Paso 1: Regrese linealmente $\mathbf{y_c}$sobre $\mathbf{X_c}$ y compute $\mathbf{b}$. Paso 2: Utilice $\mathbf{X_m}$ para predecir los valores ausentes de $\mathbf{y_m}$con $\mathbf{X_m b_c}$. Luego, regrese la muestra completa de las observaciones, $(\mathbf{y_c}, \mathbf{X_m b_c})$, sobre la muestra completa de regresores, $(\mathbf{X_c}, \mathbf{X_m})$.

a. Muestre los estimadores de mínimos cuadrados para el paso 1 y el paso 2 son idénticos.

b. Es el estimador del coeficiente en el segundo paso es insesgado?

c. Muestre que la suma de residuos al cuadrado es la misma en los dos pasos.

d. Muestre que estimador en el segundo paso de σ^2es sesgado a la baja.

19. La función de costos Cobb-Douglas generalizada es un caso especial de la función de costo translog.

$$\begin{aligned} lnC &= \alpha + \beta lnQ + \delta_k lnP_k + \delta_l lnP_l + \delta_f lnP_f + \phi_{kk}\left[\frac{1}{2}(lnP_k)^2\right] + \phi_{ll}\left[\frac{1}{2}(lnP_l)^2\right] + \phi_{ff}\left[\frac{1}{2}(lnP_f)^2\right] \\ &\quad + \phi_{kl}[lnP_k][lnP_l] + \phi_{kf}[lnP_k][lnP_f] + \gamma\left[\frac{1}{2}(lnQ)^2\right] + \theta_{Qk}[lnQ][lnP_k] + \theta_{Ql}[lnQ][lnP_l] + \theta_{Qf}[lnQ][lnP_f] + \varepsilon \end{aligned}$$

el requerimiento teórico de homogeneidad lineal en los precios de los factores impone las siguientes restricciones

$$\delta_k + \delta_f + \delta_l = 1$$

$$\phi_{kk} + \phi_{kl} + \phi_{kf} = 0$$

$$\phi_{kf} + \phi_{lf} + \phi_{ff} = 0$$

$$\theta_{Qk} + \theta_{Ql} + \theta_{Qf} = 0$$

$$\phi_{kl} + \phi_{ll} + \phi_{lf} = 0$$

Note que aunque la teoría subyacente requiere estas restricciones, el modelo puede ser estimado (por mínimos cuadrados) sin imponerse estas restricciones.

Un número adicional de restricciones relacionadas con la hipótesis de homoteticidad de la estructura de producción puede añadirse a las restricciones anteriores.

$$\theta_{Qk} = 0 \;\; \theta_{Ql} = 0, \;\; \theta_{Qf} = 0$$

La restricción de homogeneidad lineal de la estructura de produccion le añade la restricción γ=0. La hipótesis de que la todas las elasticidades de la estructura de producción son iguales a -1 es impuesta por las seis restricciones $\phi_{ij} = 0$ para todo i y j.

Utilice la información de la base datos data_ejercicio16.xls para contrastar estas restricciones. Para propósitos de este ejercicio, denote por $\beta_1, ..., \beta_{15}$ los 15 parámetros que aparecen en la función de costos en el orden que aparecen en el modelo, empezando por la primera línea hasta el final de la ecuación.

a. Escriba la matriz **R** y el vector **q**necesarios para imponer la restricción de homogeneidad linal en los precios.

b. Contraste la teoría de producción utilizando las 158 observaciones disponibles. Utilice un test F para contrastar la restricción de homogeneidad lineal.

c. Contraste la hipótesis de homoteticidad de la estructura de producción bajo el supuesto de homogeneidad lineal en precios.

20. Considere un modelo para estudiar la salud de un individuo:

$$salud = \beta_0 + \beta_1 peso + \beta_2 altura + \beta_3 hombre + \beta_4 trabajo + \beta_5 ejercicio + \beta_6 edad + u$$

Donde *salud* es alguna medida cualitativa de la salud de la persona; *peso*, *altura*, *hombre* y *edad* se explican por sí mismas; *trabajo* es el número de horas trabajadas en la semana; y *ejercicio* son las horas de ejercicio por semana.

- ¿Por qué estarías preocupado de que la variable *ejercicio* esté correlacionada con el término error?
- Suponga que puedes colectar datos sobre dos variables adicionales, *distcasa* y *disttrabajo*, la distancia desde la casa y desde el trabajo al gimnasio mas cercano. Discuta porque estas variables probablemente no estén correlacionadas con el error.
- Ahora asuma que *distcasa* y *distrabajo* no están correlacionadas con u , asimismo tampoco esta correlacionada con las demás variables del modelo planteado, exceptuando *ejercicio*. Establezca las condiciones bajo las cuales los parámetros de la ecuación de interés están identificados.
- ¿Cómo puede el supuesto de identificación ser constrastado?

21. Considere el siguiente modelo para estimar el efecto de un conjunto de variables, incluyendo el consumo de cigarrillos, sobre el peso de recién nacidos:

$$log(pesonac) = \beta_0 + \beta_1 hombre + \beta_2 orden + \beta_3 log(ingfam) + \beta_4 cajetillas + u$$

Donde *hombre* es un indicador binario igual a uno si es un niño, *orden* es una variable que indica el orden de nacimiento del niño, *ingfam* es el ingreso familiar, y *cajetillas* es el número de cajetillas de cigarro fumadas por días durante el embarazo.

- ¿Por qué se espera que cajetilla esté correlacionada con u_i?

- Suponga que tienes datos sobre el precio promedio de las cajetillas de cigarrillos cada una de las localidades donde residen las mujeres que componen la muestra. Discuta si esta información satisface las condiciones para ser considerada un buen instrumento para la variable *cajetilla*.

- Use los datos en el archivo BWGHT.xls para estimar la ecuación anterior. Primero estime por MCO. Despues, use MC2E, donde *preciocig* es un instrumento para *cajetillas*. Discuta cualquier diferencia que observe entre las estimaciones de MCO y las de MC2E.

- Estime la forma reducida de *cajetillas*. ¿Qué puedes concluir acerca la identificación de la ecuación de interés utilizando *preciocig*? ¿Qué implicaciones tiene este resultado sobre tu respuesta de la sección c?

22. Utilice los datos del archivo CARD.xls para este problema.

- Estime una ecuación en $log(wage)$ por MCO con *educ*, *exper*, $exper^2$, *black*, *south*,*smsa*, *reg*661 hasta *reg*668, y *smsa*66 como variables explicativas. Compare sus resultados con la tabla 2 (columna 2) del paper Card(1995) que viene adjunto a la tarea.

- Ahora, estime la eucación de forma reducida para *educ* empleando todas las explicativas de la parte a. y la variable dummy *nearc*4. Es estadísticamente significativa la correlacion parcial de *educ* y *nearc*4? (revise la columna 1 d ela tabla 3 del paper de referencia). Estime por VI la ecuación para $log(wage)$, usando *nearc*4 como instrumento para *educ*. Compare el intervalo de 95% de confianza para el retorno de la educación con el obtenido de la parte a. (vea la columna 5 de la tabla 3 del paper).

- Ahora use *nearc*2 junto con *nearc*4 comom instrumentos para *educ*. Primero estime la forma reducida para *educ*, y comente sobre si *nearc*2 o *nearc*4 tiene una correlación mas estrecha con *educ*. ¿Cómo las estimaciones de MC2E se compara con las estimaciones anteriores?

23. Emplee la información contenida en el archivo SLEEP.xls para estimar el siguiente modelo reportando los resultados en una tabla resumen:

$$sleep=\beta_0 + \beta_1 totwrk + \beta_2 educ + \beta_3 age + \beta_4 age^2 + \beta_5 male + u$$

La variable *sleep* es minutos totales dormidos en la noche por semana, *totwrk*es el total de minutos trabajando a la semana, *educ* y *edad* son medidas en años, y *male* es una variable ficticia o dummy de género.

(i) Manteniendo todos los demás factores iguales, existe evidencia de que los hombres duermen más que las mujeres? ¿Cuán consistente es esa evidencia?

(ii) ¿Existe un tradeoff significativo entre trabajar y dormir? ¿Cuál es el estimado de esta disyuntiva?

(iii) ¿Qué otra regresión necesitas para contrastar la hipótesis nula que, manteniendo los demás factores constantes, la edad no tiene efecto sobre la cantidad de minutos dormidos por semena?

24. Estime la siguiente ecuación utilizando los datos contenidos en BWGHT.xls y presente sus resultados:

$$log(bwght) = \beta_0 + \beta_1 cigs + \beta_2 log(faminc) + \beta_3 parity + \beta_4 male + \beta_5 white + u$$

Donde *bwght* es el peso al nacer en libras, *cigs* es el promedio diario de cigarrillos fumados por la madre durante el embarazo, $faminc$ es el ingreso anual de la familia, *parity* es el orden de nacimiento el niño, *male* es una variable binaria igual a 1 si niño y *white* es una dummy si el infante es clasificado como blanco.

(i) Interprete el coeficiente de la variable *cigs*. En particular, ¿cuál es la diferencia en el peso al nacer de fumar 10 o más cigarrillos por día?

(ii) Manteniendo los demás factores constantes, ¿cuál es la diferencia de peso predecia entre infantes clasificados como blancos y los no blancos? ¿Es esta diferencia estadísticamente significativa?

(iii) Ahora añada a la ecuación anterior las variables *motheduc* y $fatheduc$, educación de la madre y del padre respectivamente. Estime el nuevo modelo, reporte los resultados en una tabla resumen y comente la significancia estadística de *motheduc*.

(iv) De la información proporcionada, por qué no es posible computar el estadístico F para contrastar la significancia conjunta de *motheduc* y $fatheduc$. ¿Qué se puede hacer para computar el estadístico F?

25. Utilice la base de datos CEOSAL1.xls para estimar el siguiente modelo que explica el comportamiento de los salarios de ejecutivos a cargo de empresas:

$$log(salary)=\beta_0 + \beta_1 log(sales) + \beta_2 roe + \beta_3 finance + \beta_4 consprod + \beta_5 utility + u$$

Donde *salary*,*sales*, *roe*,*finance*,*consprod* y *utility* son el salario del CEO, las ventas, el retorno sobre equity, dummy igual a 1 si el CEO trabaja en la industria financiera, dummy igual a 1 si trabaja en la industria de productos de consumo y dummy igual a 1 si trabaja en la industra de servicios públicos. La industria omitida es la de transporte.

(i) Calcule la diferencia porcentual aproximada en términos de salario estimado entre las industrias de servicios públicos y la de transporte, manteniendo *sales* y *roe* constantes. ¿Es esta diferencia significativa al 1% ?

(ii) Calcule la diferencia porcentual exacta de salario estimado entre las industrias de servicios públicos y la de transporte y compare estos resultados con la respuesta obtenida en la parte (i).

(iii) ¿Cuál es la diferencia porcentual aproximada en términos de salario estimado entre las industrias de productos de consumo y la industria financiera? Escriba una ecuación que le permita contrastar si la diferencia es estadísticamente significativa.

26. Suponga que usted obtiene datos de una encuesta sobre salarios, educación, experiencia, y género. Adicionalmente, usted obtiene información acerca del uso de marihuna. La pregunta original es: "¿En el último mes cuantas veces, en diferentes ocasiones, usted fumó marihuana?".

(i) Escriba una ecuación que le permita estimar el efecto del uso de marihuna sobre el salario, mientras controla por otros factores. Debe ser tal que le permita hacer una afirmació como: "consumir marihuana cinco o más veces por mes en promedio afecta el salario estimado en x%".

(ii) Escriba un modelo que le permita contrastar si el uso de drogas tiene efectos sobre el salario diferenciados entre hombres y mujeres. ¿Cómo puede contrastar que no existen diferencias en el efectos del uso de drogas entre hombres y mujeres?

(iii) Suponga que usted piensa que es mejor medir el uso de marihuana clasificando a los individuos en cuatro categorias: no usuario, usuario menor (de 1 a 5 veces por mes), usuario moderado (de 6 a 10 veces por mes), y usuario mayor (más de 10 veces por mes). Ahora escriba un modelo que le permita estimar los efectos del uso de marihuana sobre el salario.

(iv) Utilice el modelo de la parte (iii) para explicar en detalle como contrastar la hipótesis nula de que el uso de marihuana no tiene efectos sobre el salario. Sea muy específico e incluya una lista de los grados de libertad.

(v) ¿Cuáles algunos de los problemas potenciales de extraer inferencias causales empleando los datos de la encuesta que usted realizó?

27. Use los datos en WAGE2 para realizar el siguiente ejercicio:

(i) Estime el modelo

$$log(wage)=\beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 tenure + \beta_4 married + \beta_5 black + \beta_6 south + \beta_7 urban + u$$

y reporte los resultados. Manteniendo los demás factores constante, cual es la diferencia aproximada en términos de salario mensual entre negros y no negros? ¿ Es esta diferencia estadísticamente significativa?

(ii) Añada las variables $exper^2$ y $tenure^2$a la ecuación y muestre que conjuntamente son no significativas inclusive al 20%.

(iii) Extienda el modelo origina para permitir que el retorno de la educación dependa de la raza y contraste si el retorno de la educación depende o no de la raza.

(iv) Empiece con el modelo original, epro ahora permita que los salarios difieran a través de cuatro grupos: casados y negros, casados y no negros, solteros y negros, y solteros y no negros. ¿Cuál es la diferencia salarial estimada entre los casados negros y los casados no negros?

28. Considere la siguiente función de ahorro familiar de una familia:

$$sav = \beta_0 + \beta_1 inc + \beta_2 hhsize + \beta_3 educ + \beta_4 age + u$$

donde *hhsize* es el tamaño del hogar, *educ* son los años de educación del cabeza de hogar, y *age* es la edad del cabeza de hogar. Asuma que $E(u|inc, hhsize, educ, age) = 0$.

(i) Suponga que la muestra incluye solo familias cuya cabeza de hogar es mayor a 25 años de edad. Si se emplea MCO sobre esta muestra, son los esimadores de $\beta_j$sesgados? Explique.

(ii) Ahora suponga que la muestra incluye solo parejas casadas sin niños. ¿Se pueden estimar todos los parámetros de la ecuación? ¿Cuáles pueden ser estimados?

(iii) Suponga que se excluyen de la muestra de familias aquellas que ahorren mas de \$25 mil a año. ¿Puede MCO producir estimadores consistentes de β_j?