Econometría I - 2009
Universidad de San Andrés

Trabajo Práctico 3

Modelo Lineal General y algunas extensiones

Contenidos: Variables explicativas binarias, modelos no lineales en las variables, R-cuadrado ajustado, test "F".

Parte I [1]

1. Ejercicio con variables dummies

Disponemos de datos experimentales de 100 familias con información de ahorro e ingreso anual expresados en miles de pesos. Las familias corresponden a dos regiones distintas, siendo 60 familias de la región 1 y 40 familias de la región 2.

Los datos de ahorro e ingresos para las 60 familias de la región 1 son construidos a partir de la siguiente ecuación:

$A_i = -1.5 + 0.12\ Y_i + u_i$

Los datos de ahorro e ingresos para las 40 familias de la región 2 son construidos a partir de la siguiente ecuación:

$A_i = -0.5 + 0.09\ Y_i + u_i$

Donde:

A_i es el ahorro de la familia *i*.

Y_i es el ingreso de la familia *i*.

u_i es un valor obtenido de una distribución normal con media 0 y varianza σ.

Estimación con todos los datos

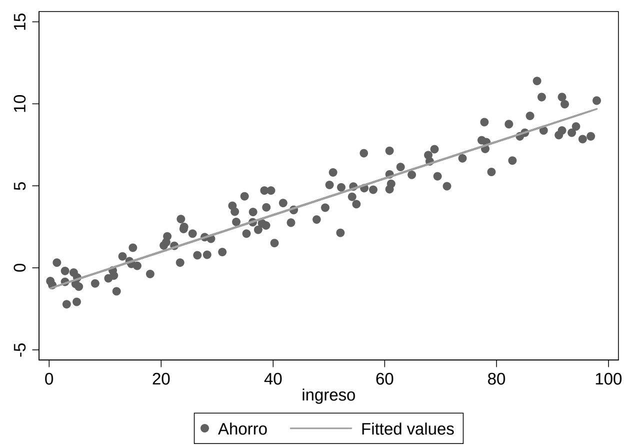
Se realizó una primera estimación con todos los datos disponibles (las 100 familias) obteniéndose los siguientes resultados.

---

[1] Puede utilizar calculadora o planilla de cálculo para hacer sumas o productos de datos, pero las operaciones con matrices deben hacerse *a mano*. Para los ejercicios de la Parte I no puede utilizarse STATA.

```
      Source |       SS       df       MS              Number of obs =     100
-------------+------------------------------           F(  1,    98) = 1081.18
       Model |  1048.76522     1  1048.76522           Prob > F      =  0.0000
    Residual |  95.0616049    98  .970016377           R-squared     =  0.9169
-------------+------------------------------           Adj R-squared =  0.9160
       Total |  1143.82682    99  11.5538063           Root MSE      =  .98489

------------------------------------------------------------------------------
      ahorro |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     ingreso |   .1118083   .0034004    32.88   0.000     .1050604    .1185562
       _cons |  -1.257459    .183726    -6.84   0.000    -1.622057   -.8928604
------------------------------------------------------------------------------
```

En la figura siguiente se presenta el gráfico de dispersión de todos los datos (ambas regiones) y la recta de regresión obtenida en la estimación anterior.



<u>Estimaciones separadas por regiones</u>

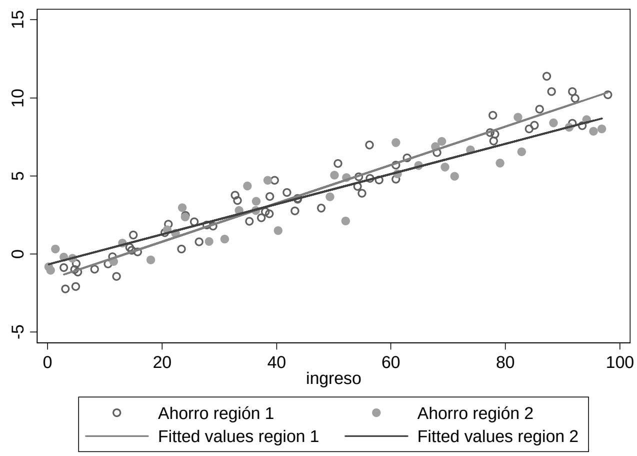
La estimación con los datos de la región 1 (60 familias) se presenta en la siguiente tabla.

```
      Source |       SS       df       MS              Number of obs =      60
-------------+------------------------------           F(  1,    58) = 1029.70
       Model |   738.30989     1   738.30989           Prob > F      =  0.0000
    Residual |  41.5868721    58  .717015036           R-squared     =  0.9467
-------------+------------------------------           Adj R-squared =  0.9458
       Total |  779.896763    59  13.2185892           Root MSE      =  .84677

------------------------------------------------------------------------------
      ahorro |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     ingreso |   .1228824   .0038294    32.09   0.000     .1152169    .1305478
       _cons |  -1.669169   .2033616    -8.21   0.000    -2.076241   -1.262096
------------------------------------------------------------------------------
```

La estimación con los datos de la región 2 (40 familias) se presenta en la siguiente tabla.

```
      Source |       SS       df       MS              Number of obs =      40
-------------+------------------------------           F(  1,    38) =  322.72
       Model |  325.581367     1  325.581367           Prob > F      =  0.0000
    Residual |  38.3368901    38  1.00886553           R-squared     =  0.8947
-------------+------------------------------           Adj R-squared =  0.8919
       Total |  363.918257    39  9.33123736           Root MSE      =  1.0044

------------------------------------------------------------------------------
      ahorro |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     ingreso |   .0965929   .0053769    17.96   0.000     .0857079    .1074778
       _cons |  -.6707318   .2978376    -2.25   0.030    -1.273673   -.0677911
------------------------------------------------------------------------------
```

En la figura siguiente se presenta el gráfico de dispersión de todos los datos (distinguiendo las regiones) y las rectas de regresión para ambas regresiones.



Estimación con variable dummy aditiva para distinguir regiones

Se incorpora a la estimación una variable dummy que identifica la región 1.

$A_i = \beta_0 + \beta_1\ Y_i + \beta_2\ R_i + u_i$

Donde R_i = 1 cuando la familia *i* corresponde a la región 1 y R_i = 0 en caso contrario. En la estimación con Stata la variable *R* se denomina `reg1`.

```
      Source |       SS       df       MS              Number of obs =     100
-------------+------------------------------           F(  2,    97) =  541.68
       Model |  1049.82956     2  524.914779           Prob > F      =  0.0000
    Residual |  93.9972657    97  .969043976           R-squared     =  0.9178
-------------+------------------------------           Adj R-squared =  0.9161
       Total |  1143.82682    99  11.5538063           Root MSE      =   .9844

------------------------------------------------------------------------------
      ahorro |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     ingreso |   .1119337   .0034008    32.91   0.000     .1051842    .1186833
        reg1 |   .2107189   .2010644     1.05   0.297    -.1883383    .6097761
       _cons |  -1.389611   .2227599    -6.24   0.000    -1.831728   -.9474942
------------------------------------------------------------------------------
```

Estimación con variables dummies aditiva y multiplicativa para distinguir regiones

Se incorpora a la estimación anterior una variable dummy multiplicativa definida por $Mi = Y_i\ R_i$. En la estimación con Stata la variable *M* se denomina `ingreg.`

$A_i = \beta_0 + \beta_1\ Y_i + \beta_2\ R_i + \beta_3\ M_i + u_i$

Los resultados obtenidos se presentan a continuación:

```
      Source |       SS       df       MS              Number of obs =     100
-------------+------------------------------           F(  3,    96) =  425.97
       Model |  1063.90306     3  354.634354           Prob > F      =  0.0000
    Residual |  79.9237622    96  .832539189           R-squared     =  0.9301
-------------+------------------------------           Adj R-squared =  0.9279
       Total |  1143.82682    99  11.5538063           Root MSE      =  .91244

------------------------------------------------------------------------------
      ahorro |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     ingreso |   .0965929   .0048845    19.78   0.000     .0868973    .1062885
        reg1 |   -.998437     .34817    -2.87   0.005    -1.689549   -.3073249
      ingreg |   .0262895   .0063942     4.11   0.000     .0135972    .0389818
       _cons |  -.6707318    .270561    -2.48   0.015    -1.207791   -.1336725
------------------------------------------------------------------------------
```

Se pide|

a) Interprete económicamente y estadísticamente los resultados de la regresión con todos los datos sin variables dummies.

b) Interprete económicamente y estadísticamente los resultados de la regresión sólo con variable dummy aditiva.

c) Calcule la estimación de la ordenada al origen y la pendiente de la recta de regresión para la región 1 y la región 2, según los resultados de la estimación con variables dummies aditiva y multiplicativa.

d) Compare los resultados del inciso anterior con la estimación separada por regiones y con los datos reales de β_0 y β_1.

e) Evalúe si la recta de regresión para la región 2 se distingue significativamente de la recta de regresión de la región 1.

Parte II[2]

2. Curva de Engel

En este ejercicio se retoma el análisis de la curva de Engel del gasto en alimentos del Trabajo Práctico Nº 1. La base de datos disponible contiene información sobre el ingreso per cápita familiar (ipcf) y el gasto en alimentos per cápita (gasto_alimentos) de 1748 hogares con ingresos y gastos en alimentos positivos en el año 2002. Las dos variables se encuentran medidas en pesos corrientes.

Se empleará el siguiente modelo:

$$g_i = \alpha + \beta . y_i + u_i \qquad i = 1,...,1748$$

donde g es el logaritmo natural del gasto en alimentos per cápita, y es el logaritmo del ingreso per cápita familar y u es un término de error.

a) Explique el significado económico de los parámetros de este modelo.
b) Transforme las variables. Realice el diagrama de dispersión de las variables transformadas y compárelo con el diagrama obtenido en el inciso b) del ejercicio de curva de Engel del Trabajo Práctico Nº 1.
c) Estime el modelo por MCO.
d) Evalúe la hipótesis de significatividad individual de cada coeficiente e interprételos económicamente.
e) Utilice el modelo para validar o refutar la siguiente afirmación: "el gasto en alimentos posee elasticidad unitaria"
f) Compare el valor de la elasticidad-ingreso del gasto en alimentos con las obtenidas en el inciso h) del Trabajo Práctico 1. Comente. ¿Se mantienen las conclusiones referidas al tipo de bien?

En la literatura de curvas de Engel se encuentra muy difundida la especificación conocida como Working- Leser:

$$s_i = \beta_0 + \beta_1 y_i + u_i \qquad i = 1,...,1748$$

donde s es la participación del gasto en alimentos en el ingreso, y es el logaritmo natural del ingreso per cápita familar y u es un término de error.

g) Construya la nueva variable y estime el modelo por MCO.
h) Interprete cuidadosamente los resultados obtenidos.
i) Analice si se cumple la ley de Engel.
j) Compare con los modelos anteriores.

[2] Los ejercicios de la Parte II están pensados para resolver usando STATA.

3. Ecuaciones de ingresos y retornos a la educación

Una rama de la economía laboral se concentra en el estudio de los determinantes de los ingresos laborales. ¿Qué hace que un individuo gane más que otro?. Citando a Adam Smith en *La Riqueza de las Naciones* (libro I, Cap. 10): "Los salarios crecen con la dificultad y el costo de aprender una tarea". Esto plantea un nexo teórico entre la educación y los salarios: la educación aumenta la productividad del trabajador y consecuentemente su salario.

Una herramienta para estudiar empíricamente la relación entre la educación y los salarios son las llamadas *ecuaciones de ingresos* ó *ecuaciones de Mincer*.[3] La ecuación (1) es la especificación más básica de una ecuación de ingresos:

$$(1) \qquad \ln Y_i = \alpha + \beta\, edu_i + Z_i'\,\delta + \mu_i$$

donde Y son los ingresos laborales por hora o salarios horarios, *edu* son los años de educación formal, Z son otras variables que se considera que pueden afectar los salarios y μ es un término aleatorio que supondremos que cumple con los supuestos clásicos.

La ecuación de ingresos (1) es lo que se conoce como un modelo de precios hedónicos. El bien mano de obra tiene múltiples características. Su precio en equilibrio (el salario de mercado) puede interpretarse como que surge de la interacción de diferentes mercados implícitos, cada uno correspondiente a una de esas características. El análisis empírico de las ecuaciones de Mincer permite estimar la contribución de cada una de las características del trabajador a su salario.

¿Cuáles son las características que determinan el salario? Capacidades adquiridas como consecuencia de la inversión en capital humano (educación formal, experiencia laboral), capacidades debidas a habilidades innatas del individuo ("inteligencia"), tipo de empleo (rama de actividad, por ejemplo), base de contrato part-time o full-time, género y otros factores como la pertenencia a una unión sindical, la raza del individuo o su región de residencia.

Concentrémonos en la contribución de la educación a los salarios. Derivando (1) con respecto a *edu* se obtiene:

$$(2) \qquad \beta = \frac{\partial \ln Y}{\partial edu} = \frac{\partial Y}{\partial edu}\frac{1}{Y} \cong \frac{dY/Y}{d(edu)} = \textit{retorno a la educación}$$

Es decir, el coeficiente β es la semielasticidad del salario horario con respecto a la educación: esperaríamos que un año más de educación formal aumente el salario por hora en $\beta\times100\%$. Por este motivo β se conoce como el *retorno a la educación.*

El objetivo de este ejercicio es estimar e interpretar ecuaciones de ingresos utilizando información proveniente de la Encuesta Permanente de Hogares (EPH).

[3] Mincer, J. (1974). Schooling, experience, and earnings. National Bureau of Economic Research. Columbia University Press. New York.

La EPH es relevada por el Instituto Nacional de Estadísticas y Censos (INDEC) y es la principal encuesta de hogares que se realiza en Argentina. Actualmente cubre 31 aglomerados urbanos de más de 100.000 habitantes, que representan al 71% del total de la población urbana y 62% de la población total del país. Mediante la EPH se recaba información sobre características demográficas, educativas y laborales de los hogares y los individuos que los componen.

La base de datos *eph.xls* cuenta con información extraída de la EPH del primer trimestre del año 2003 para la submuestra de individuos entre 25 y 65 años de edad del Area Metropolitana (Conurbano y Ciudad de Buenos Aires). Se incluyen las siguientes variables:[4]

- Código de identificación (codusu)
- Edad del individuo (h12)
- Género del individuo (h13)
- total de horas trabajada semanales (p15t)
- categorías educativas (p56)
- finalización o no del nivel educativo (p58)
- salario horario (inghora)

Se pide

a) Estime el modelo (1). Como variables explicativas incluidas en el vector *Z* utilice los años de educación[5], la edad, el género y el tipo de contrato (full time o part time)[6] del individuo. Interprete económica y estadísticamente el efecto estimado de cada una de las variables sobre los salarios horarios.

b) Considere la siguiente hipótesis de investigación: “Los retornos a la educación varían entre géneros”. A partir del modelo estimado en (a), proponga una nueva especificación que permita evaluar la validez de la hipótesis anterior. Estime el modelo y evalúe la hipótesis. Grafique.

c) Estime nuevamente el modelo en (a) pero en lugar de emplear los años de educación utilice dummies por grupos educativos (usar como categoría base u omitida un nivel de primaria incompleta). ¿Cuál es el retorno a la educación estimado de completar la escuela primaria? ¿Y de pasar de un nivel educativo de primaria completa a secundaria completa? ¿Y de secundaria completa a superior completa?. Grafique.

d) El “retorno a la edad” (semielasticidad de los salarios respecto de la edad) refleja el hecho de que a mayor edad potencialmente mayor sería la experiencia laboral del individuo y, consecuentemente, mayor su productividad. Sin embargo, podría esperarse que el retorno a la edad no sea constante: la capacidad de

[4] Adicionalmente se incluye como material el archivo eph.pdf que contiene una descripción de la eph.
[5] La variable años de educación debe construirse utilizando las variables p56 y p58.
[6] La variable full-time/part-time debe construirse utilizando la variable p15t.

aprender a hacer una tarea puede variar a lo largo de la vida, las técnicas aprendidas pueden depreciarse cuando aparecen nuevas tecnologías, etc. Para explorar esta posibilidad se pide que estime el mismo modelo estimado en (a) agregando un término con la edad al cuadrado. Evalúe las siguientes hipótesis nulas: (d.1) la edad tiene un efecto lineal sobre el logaritmo de los salarios horarios y (d.2) la edad no es relevante para explicar al logaritmo de los salarios. ¿Cuál es el retorno estimado a la edad para una persona de 30 años?¿Y para una persona de 55 años?. Grafique la relación edad – logaritmo del salario horario para el rango de edad relevante en la muestra.