Econometría I - 2020
Facultad de Ciencias Económicas
Universidad Nacional de La Plata

Trabajo Práctico 3

Modelo Lineal General (segunda parte)

Contenidos: Modelo lineal con *K* variables, notación matricial, propiedades estadísticas, inferencia, test "t", tests "*F*", *R*-cuadrado ajustado, estimación computacional, multicolinealidad.

Fecha de entrega: jueves 5 de noviembre. Ver reglas de formato y presentación en https://econometria1unlp.com/trabajos-practicos/.

PARTE I

1. Ejercicio teórico con matrices

Suponer que $Y = X\,b + \mu$, y que se cumplen todos los supuestos clásicos.

Considerar el estimador $\tilde{b} = HY$, en donde H es una matriz no estocástica K x N (K=número de variables, N=observaciones)

Se pide:

a) ¿Qué requisitos deberíamos imponerle a H para que el estimador sea insesgado?

b) Suponer que $H = (Z'X)^{-1}Z'$, donde Z es cualquier matriz no estocástica (N x K) pero distinta de X. Mostrar que este estimador es insesgado.

c) Justificar por qué el estimador del inciso *b*) es ineficiente comparado con el estimador de mínimos cuadrados ordinarios.

2. Completar

A continuación se presenta una salida de regresión de la variable *salario* en las variables *edad* y *aedu*. Lamentablemente se ha omitido cierta información (denotada con XXX en la tabla). Completar la información faltante y describir los pasos seguidos para obtenerla.

Source	SS	df	MS		Number of obs	45,961
Model	XXX	2	XXX		F(2, 45958)	2489.75
Residual	86919642.4	XXX	XXX		Prob > F	0
Total	XXX	XXX	XXX		R-squared	0.0978
					Adj R-squared	0.0977
					Root MSE	43.489

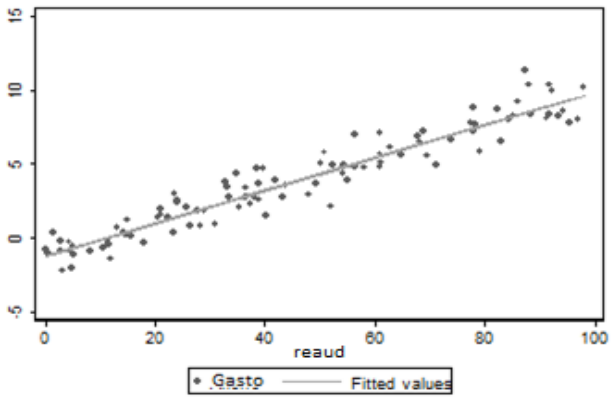
salario	Coef.	Std. Err.	t	P>t	[95% Conf.	Interval]
edad	XXX	0.016	34.94	0.000	0.5137804	0.57486
aedu	3.609	XXX	64.52	0.000	3.499467	3.71873
_cons	-15.498	0.962	XXX	0.000	-17.38455	-13.612

3. Ejercicio con variables dummies

Interesa estudiar la relación entre el gasto público municipal y la recaudación de impuestos mediante un análisis de regresión. Se dispone de datos de 100 municipios con información de gasto público (variable dependiente: G_i) y recaudación anual (variable explicativa: R_i) expresados en millones de pesos. Los municipios corresponden a dos provincias distintas, siendo 60 municipios de la provincia 1 y 40 municipios de la provincia 2. Suponer que el término aleatorio del modelo (μ_i) cumple con todos los supuestos clásicos.

3.1. Estimación con todos los datos

Se realizó una primera estimación con todos los datos disponibles (los 100 municipios). La siguiente figura muestra el gráfico de dispersión y la recta de regresión estimada.



Los resultados de la estimación en Stata son:

```
      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

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

3.2. Estimaciones separadas para cada provincia

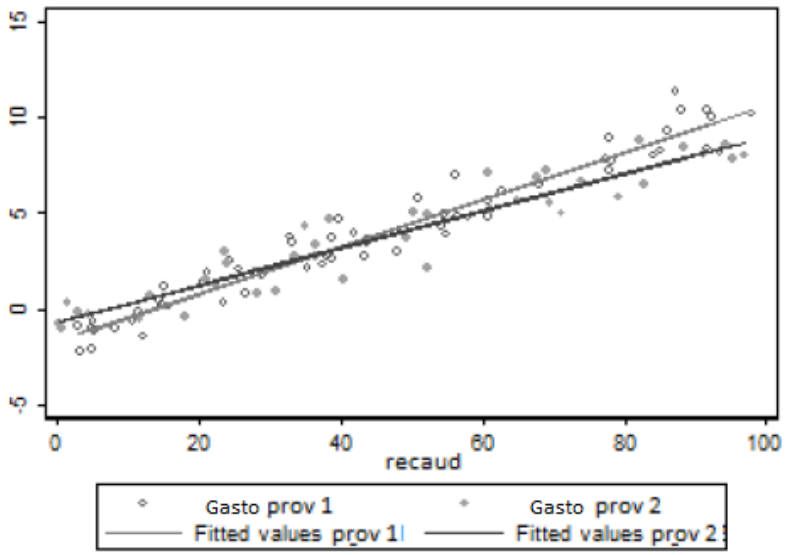
La estimación en Stata con los datos de la provincia 1 (60 municipios) es:

```
      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

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

La estimación en Stata con los datos de la provincia 2 (40 municipios) es:

```
      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

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

En la figura siguiente se presenta nuevamente el gráfico de dispersión, pero ahora es posible diferenciar qué datos pertenecen a la provincia 1 (marcador sin relleno) y qué datos a la provincia 2 (marcador con relleno). En el gráfico también se muestran las rectas de regresión estimadas para cada una de las provincias.



3.3. Estimación con variable dummy aditiva para distinguir provincias

Se incorpora a la estimación una variable dummy que identifica la provincia 1 (*prov1*).

$$G_i = \beta_1 + \beta_2 R_i + \beta_3 prov1_i + \mu_i$$

Donde $prov1_i = 1$ cuando el municipio *i* corresponde a la provincia 1 y $prov1_i = 0$ en caso contrario.

```
      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

------------------------------------------------------------------------------
      gasto  |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      recaud |   .1119337   .0034008    32.91   0.000     .1051842    .1186833
       prov1 |   .2107189   .2010644     1.05   0.297    -.1883383    .6097761
       _cons |  -1.389611   .2227599    -6.24   0.000    -1.831728   -.9474942
```

3.4. Estimación con variables dummies aditiva y multiplicativa para distinguir provincias

Se incorpora a la estimación anterior una variable dummy multiplicativa definida por $M_i = R_i \, prov1_i$.

$$G_i = \beta_1 + \beta_2 R_i + \beta_3 prov1_i + \beta_4 M_i + \mu_i$$

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

------------------------------------------------------------------------------
      gasto  |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      recaud |   .0965929   .0048845    19.78   0.000     .0868973    .1062885
       prov1 |   -.998437     .34817    -2.87   0.005    -1.689549   -.3073249
           M |   .0262895   .0063942     4.11   0.000     .0135972    .0389818
       _cons |  -.6707318    .270561    -2.48   0.015    -1.207791   -.1336725
------------------------------------------------------------------------------
```

Se pide:

a) Interpretar económica y estadísticamente los resultados de la regresión 3.1, es decir, la que utiliza todos los datos (de ambas provincias) sin incorporar variables dummies al modelo.

b) Interpretar económica y estadísticamente los resultados de la regresión 3.3, es decir, la que incluye la variable dummy aditiva *prov1*.

c) Sobre la base de los resultados de la estimación 3.4 (con variables dummies aditiva y multiplicativa), computar la ordenada al origen estimada y la pendiente estimada de la recta de regresión para las provincias 1 y 2.

d) Comparar los resultados del inciso anterior con lo que se obtiene en la estimación 3.2, es decir, estimando un modelo para cada provincia por separado. ¿A qué conclusión se llega?

e) Evaluar la hipótesis de que los coeficientes de las variables *prov1* y *M* son simultáneamente iguales a cero considerando una hipótesis alternativa bilateral y un nivel de confianza del 95%. Interpretar el significado de esta prueba.

4. Ecuaciones de ingresos y retornos a la educación

Retomar el modelo de ecuaciones de Mincer visto en el Trabajo Práctico 2:

(1) $\ln W = X\,\theta + \mu$

donde *lnW* es el logaritmo de los salarios horarios (vector $N \times K$), X es una matriz $N \times K$ que incluye las variables explicativas, en este caso las características de los individuos que se consideran relevantes para explicar el salario, θ es el vector $K \times 1$ de parámetros desconocidos y μ el vector $N \times 1$ de variables aleatorias inobservables, que cumple los supuestos clásicos. El objetivo de este ejercicio es estimar e interpretar ecuaciones de ingresos utilizando información proveniente de la Encuesta Permanente de Hogares (EPH).

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_2019_s2.dta* cuenta con información extraída de la EPH del segundo semestre del año 2019 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:[1]

- Código de identificación para distinguir viviendas (codusu)
- Código de identificación para distinguir hogares (nro_hogar)
- Trimestre de la encuesta (trimestre)
- Edad del individuo (ch06)
- Género del individuo (ch04)
- Total de horas trabajada semanales (pp3e_tot)

[1] Adicionalmente se incluye como material el archivo EPH_registro_4t19.pdf que contiene una descripción de la EPH.

- Nivel educativo (nivel_ed)
- Ingreso monetario en la ocupación principal en el mes anterior a la entrevista (p21)

Se pide:

a) Estimar el modelo (1). Como variables explicativas incluidas en el vector *X* utilice: los años de educación, la edad, el género y el tipo de contrato (full time o part time) del individuo. Interpretar económica y estadísticamente el efecto estimado de cada una de las variables sobre los salarios horarios.

Ayuda: para construir las variables que se usarán en la estimación se debe tener en cuenta lo siguiente:

- El salario horario se obtiene a partir de las variables p21 y pp3e_tot.
- Los años de educación se pueden computar usando la variable nivel_ed
- La variable full-time/part-time debe construirse utilizando la variable pp3e_tot. Definir trabajo full-time si las horas trabajadas por semana son 30 o más.

b) Considerar la siguiente hipótesis de investigación: "Los retornos a la educación varían entre géneros". A partir del modelo estimado en (a), proponer una nueva especificación que permita evaluar la validez de la hipótesis anterior. Estimar el modelo y evaluar la hipótesis. Graficar.

c) Estimar nuevamente el modelo en (a) pero en lugar de los años de educación utilizar 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? ¿Es esta diferencia estadísticamente significativa? Explicar detalladamente el test de hipótesis empleado.

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 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 estimar un modelo similar al estimado en (a) pero agregando un término con la edad al cuadrado. Evaluar las siguientes hipótesis nulas: (d.1) la edad tiene un efecto constante 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 33 años? ¿Y para una persona de 60 años? Graficar la relación entre el logaritmo del salario horario y la edad para el rango de edad relevante en la muestra.

5. Modelos no lineales en variables e inferencia

Retomar el ejemplo de la curva de Kuznets ambiental del TP 2 y, utilizando los datos de la base "*polucion2006.dta*", realizar los ejercicios de inferencia estadística para evaluar las hipótesis nulas planteadas en los siguientes incisos. Explicar con detalle los pasos en cada prueba de hipótesis, incluyendo la decisión del nivel de confianza empleado.

a) La contaminación ambiental es un fenómeno independiente del nivel educativo del país.

b) No hay relación entre el desarrollo económico y el nivel de polución.

c) La relación entre el PBI per cápita y la contaminación es lineal.

d) El efecto marginal del desarrollo económico sobre la contaminación en un país con un PBI per cápita de 20 mil dólares es nulo.

6. Modelo del ingreso familiar

La base de datos *itf_2019_s2.dta* cuenta con información extraída de la EPH del segundo semestre del año 2019 para la submuestra de jefes de hogar. Se incluyen las siguientes variables:

- *itf*: ingreso total familiar medido en pesos.
- *aedu*: años de educación del jefe de hogar.
- *n_ocup*: número de personas ocupadas en el hogar.
- *propieta*: variable binaria que vale 1 si el hogar es propietario de la vivienda y 0 en caso contrario.
- *region*: variable categórica donde cada valor indica una región de Argentina, 1 = Gran Buenos Aires (GBA), 2 = Pampeana, 3 = Cuyo, 4 = Noroeste (NOA), 5 = Patagonia y 6 = Noreste (NEA).

Se pide:

a) Estimar un modelo para explicar el logaritmo del ingreso total familiar incluyendo como variables explicativas *aedu*, *n_ocup*, *propieta* y una variable indicadora de regiones del norte del país (=1 si la región es NOA o NEA, y 0 en caso contrario). Interpretar económica y estadísticamente el efecto estimado de cada una de las variables sobre el logaritmo del ingreso total familiar.

b) Proponer un nuevo modelo y evaluar la siguiente hipótesis: "El efecto de un año adicional de educación del jefe de hogar sobre el logaritmo del ingreso total familiar difiere entre el norte y el resto del país".

c) Volviendo al modelo estimado en a), agregar como variable explicativa adicional un término cuadrático para los años de educación del jefe de hogar. Evaluar las siguientes hipótesis: (i) Los años de educación del jefe no son relevantes para explicar el logaritmo del ingreso total familiar; (ii) El efecto marginal de los años de educación del jefe de hogar sobre el logaritmo del ingreso total familiar es nulo para un jefe con 10 años de educación. En ambos casos, plantear las hipótesis nula y alternativa y especificar el estadístico de prueba que se utiliza.

PARTE II: Multicolinealidad

7. Ecuación de Mincer con diferencias regionales

El objetivo de este ejercicio es mostrar dos situaciones típicas donde hay multicolinealidad perfecta. El archivo *mincer_2019_s2.dta* contiene una muestra de hombres de 18 a 65 años de edad que proviene de la Encuesta Permanente de Hogares de Argentina del segundo semestre de 2019. Las variables que contiene son:

- *salario*: ingreso laboral por hora en pesos corrientes.
- *edad*: edad medida en años.
- *educ*: educación medida en años de educación formal.
- *region*: variable categórica donde cada valor indica una región de Argentina, 1 = Gran Buenos Aires (GBA), 2 = Pampeana, 3 = Cuyo, 4 = Noroeste (NOA), 5 = Patagonia y 6 = Noreste (NEA).
- *sexo*: indicador del género, toma el valor 1 si el encuestado es hombre y 2 si es mujer.

El modelo a estimar es:

(3) $ln(salario)_i = \beta_1 + \beta_2\, educ_i + \beta_3\, edad_i + \beta_4\, edad^2_i + Z_i'\gamma + \mu_i$

donde Z es un vector columna que incluye las variables binarias por región y gamma el vector columna de los parámetros asociados a las regiones. Para simplificar el análisis, considere solo tres áreas geográficas: *norte* (regiones NOA y NEA), *centro* (GBA, Cuyo y Pampa) y *sur* (Patagonia).

a) Calcular estadísticas descriptivas de las variables (medias, desvíos, rangos) y comentar.

b) Estimar por MCO el modelo (3) incluyendo las tres dummies regionales como variables explicativas. ¿Qué se observa? ¿A qué se debe ese resultado? Hacer explícito el argumento en términos analíticos.

c) Estimar por MCO el modelo (3) pero sin incluir a la constante del modelo. ¿A qué se debe el cambio de resultado? Explicar e interpretar los coeficientes estimados.

d) Construir una variable dummy *hombre* (=1 si es hombre y 0 en caso contrario) e incluirla como variable explicativa. Estimar el modelo por MCO. Comentar sobre la causa detrás del resultado observado.

8. Movilidad intergeneracional de la educación

La literatura sobre movilidad intergeneracional estudia el proceso por el cual los individuos van cambiando de estrato social a lo largo de las generaciones. La relación entre la educación de los hijos con sus padres es considerada por algunos autores como un indicador simple de movilidad social entre generaciones. A los efectos de estudiar este aspecto, considerar el siguiente modelo:

(4) $educ_i = \beta_1 + \beta_2\, educpadre_i + \beta_3\, educmadre_i + \beta_4\, hombre_i + \mu_i$

Donde:

- *educ*: años de educación del individuo.
- *educpadre*: años de educación de su padre.
- *educmadre*: años de educación de su madre.
- *hombre*: variable dicotómica que toma valor 1 si el encuestado es hombre y 0 si es mujer.

Los datos en el archivo *movilidad1998.dta*. Se trata de una muestra de 490 personas mayores de 30 años de edad encuestadas en 1998 en distintas regiones de Uruguay. Adicionalmente, la base incluye una variable categórica *area* que indica las 5 regiones a las que pertenecen las observaciones.

a) Considerando solamente las observaciones de la Región 1, estimar por MCO el modelo (4). Comentar los resultados en términos económicos y estadísticos (inferencia y bondad de ajuste).

b) Con los mismos datos, estimar otro modelo que solo incluya al género y la educación del padre como variables explicativas. ¿Qué cambios observa? ¿Puede concluir que la educación del padre es relevante para explicar la educación de los hijos? Comentar.

c) Estimar la correlación entre la educación del padre y la madre para la Región 1. ¿A qué se puede atribuir el cambio de comportamiento del estimador de MCO observado entre los incisos (*a*) y (*b*)?

d) Algunos economistas opinan que lo relevante para explicar el nivel educativo de los individuos es el capital humano de la familia en su conjunto. Construir una nueva variable que contenga el máximo nivel educativo entre el padre y la madre. Estimar por MCO el modelo reemplazando las variables educativas de los padres por esta medida resumen. Comentar los resultados en términos económicos y estadísticos (inferencia y bondad de ajuste).

e) Un investigador sugiere que tal vez se puedan apaciguar las consecuencias de la alta multicolinealidad aumentando el número de observaciones en la estimación. Estimar

nuevamente los parámetros de la ecuación (4) pero ahora utilizando las observaciones de todas las regiones. Interpretar cada uno de los coeficientes y comentar sobre la utilidad de la sugerencia del investigador a la luz de sus resultados.

f) Considerando las distintas alternativas seguidas en los incisos anteriores, ¿es la multicolinealidad alta (no perfecta) un problema grave para la estimación por el método de MCO? Comentar brevemente.