Análisis Econométrico

Práctica 1: Econometría

José Burgos

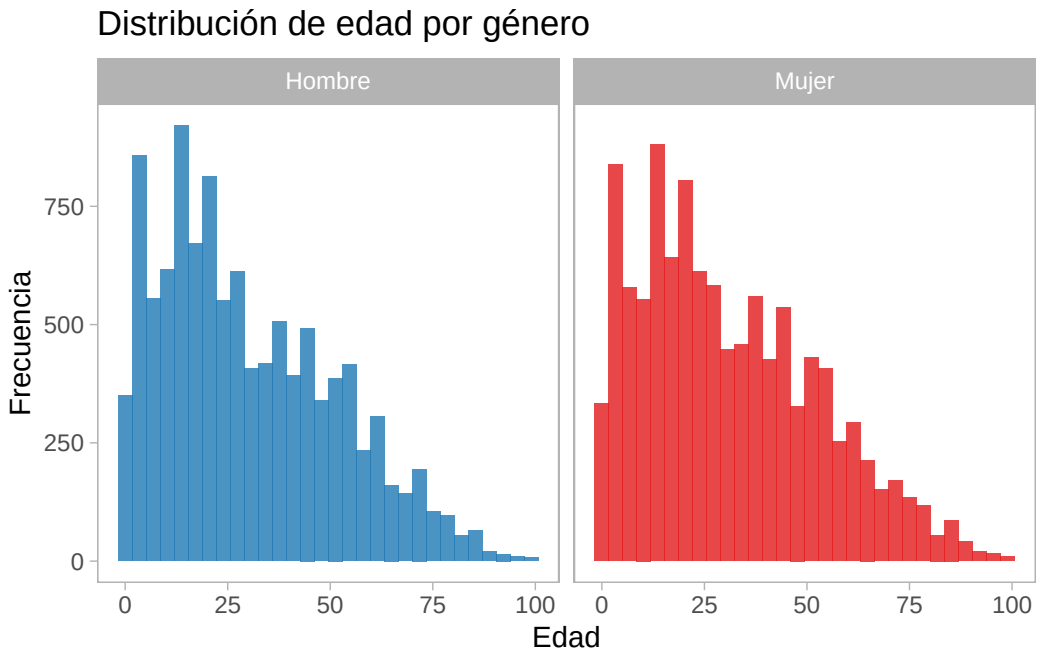
Ejercicio 1: Considere la base de datosBase *ENCFT 20161 – 20164.xlsx* que contiene información sobre el lado de la oferta del mercado de trabajo.

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

b) Adjunte el módulo de MIEMBROS y el de OCUPACIÓN, creando una variable `id` que concatene la información de la vivienda, el hogar y el miembro, de modo que en una misma base de datos queden las características de los individuos y las variables de mercado laboral.

c) Cree la variable `EDAD` y `MUJER`, donde esta última sea igual a 1 si el individuo es mujer y 0 en caso contrario.

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. Tenga en cuenta que el ingreso reportado en la base de datos no está necesariamente en la misma escala para todos los individuos (p. ej., algunos reportan salario por hora, otros por mes, etc.).

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

Table 1: Distribución de edad por percentil de ingreso

Rango de Edad	Percentil 5	Percentil 25	Percentil 50	Percentil 75	Percentil 95
35-44	25.00	48.38	66.25	102.81	218.75
Menos de 25	16.12	37.50	50.00	75.00	125.00
25-34	25.00	45.00	62.50	87.50	187.50
45-54	18.75	43.75	62.50	100.00	250.00
55-64	13.44	37.50	61.88	97.81	251.87
65 o más	12.50	32.25	43.75	70.00	148.75

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

   Propiedades de muestra finita:

1. **Insesgadez**:
$$\hat{\beta} = (X'X)^{-1}X'y \ \rightarrow \ E[\hat{\beta}|X] = \beta$$
por expectativas iteradas $E[\hat{\beta}] = \beta$. Esto demuestra que, para cualquier muestra (de tamaño $n >\_$), el estimador de MCO está centrado en el valor verdadero del parámetro.

2. **Varianza condicional**

$$Var(\hat{\beta}|X) = Var[(X'X)^{-1}X'\epsilon|X] = \sigma^2(X'X)$$
De aquí se obtienen las varianzas individuales y las covarainzas entre componentes de $\hat{\beta}$.

3. **Insesgadez del estimados de σ^2**

El estiamdor habitual de la varianza del error es:

$$s^2 = \frac{e'e}{n-(k+1)}, \qquad e = y - X\hat{\beta}$$
Bajo los supuestos clásicos se tiene $E[s^2|X] = \sigma^2$, es decir, s^2 es insesgado para σ^2 en muestras finitas.

4. **Optimalidad (Teorema de Gauss-Markov)**

3. **Modelo de regresión lineal**
$$y = \alpha + \beta x + \varepsilon$$

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

Demostración: Para demostrar que las ecuaciones normales de mínimos cuadrados implican que $\sum_i e_i = 0$ y $\sum_i x_i e_i = 0$, partimos de la definición del error $e_i = y_i - \hat{y}_i$, donde $\hat{y}_i$ es la predicción del modelo.

La ecuación de regresión lineal es:

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

c) Muestre que la solución para b es $b = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}(x_i-\bar{x})^2}$

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 $X1$ igual a la constante, educación, experiencia y habilidad. Mientras que sea $X2$ igual a la educación 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 X1. Reporte los coeficientes.

Modelo:

$$wage = \beta_0 + \beta_1 education + \beta_2 experience + \beta_3 ability + \epsilon$$

Table 2: Coeficientes de MCO en la regresión de y sobre X_1

Variable	Estimación	Error estándar	Valor t	Valor p
(Intercept)	1.664	0.619	2.690	0.021
education	0.015	0.049	0.297	0.772
experience	0.071	0.048	1.479	0.167
ability	0.027	0.099	0.269	0.793

b) Compute los coeficientes de MCO en la regresión de y sobre X1 y X2. Reporte los coeficientes.

$$wage = \beta_0 + \beta_1 education + \beta_2 experience + \beta_3 ability + \beta_4 education01 + \beta_5 education02 + \beta_5 siblings + \epsilon$$

Table 3: Coeficientes de MCO en la regresión de y sobre X_1 y X_2

Variable	Estimación	Error estándar	Valor t	Valor p
(Intercept)	0.049	0.949	0.052	0.960
education	0.026	0.045	0.578	0.579
experience	0.103	0.047	2.184	0.061
ability	0.031	0.121	0.254	0.806
education01	0.102	0.070	1.448	0.186
education02	0.002	0.045	0.037	0.972
siblings	0.059	0.069	0.857	0.416

c) Regrese cada una de las tres variables en X2 sobre todas las variables en X1. Estas nuevas variables denótenlas como X_2^* ¿Cuáles son las medias muestrales de estas variables? Explique el resultado.

Table 4: Medias muestrales de las variables X_2^*

Variable	Media
education01	12.067
education02	12.667
siblings	2.200

Las medias muéstrales de las variables en X_2^* representan el valor promedio de cada variable después de haber sido ajustadas por las variables en X_1. Esto significa que estas medias reflejan el efecto de la educación, experiencia y habilidad sobre la educación de los padres y el número de hermanos, eliminando la variabilidad que podría estar asociada con estas variables. En otras palabras, las medias muéstrales indican el nivel promedio de educación de los padres y el número de hermanos, controlando por las características del individuo como la educación, experiencia y habilidad.

d) Compute $R2 = 1 - \frac{e'e}{y'M_0y}$ para la regresión de y sobre X1 y X2. Repita el cómputo para el caso en la que el término constante es omitido de X1. ¿Qué sucede con R2?

Table 5: Valores de R^2 para los modelos

Modelo	Condición	R^2
y ~ X1	Con constante	0.183
y ~ X1	Sin constante	0.980
y ~ X2	Con constante	0.516
y ~ X2	Sin constante	0.993

En esta regresión cuando se omite el término constante, el R^2 aumenta, lo que indica que el modelo sin constante explica una mayor proporción de la variabilidad de la variable dependiente. El problema de esto, es que al eliminar el término constante estamos omitiendo la media de la variable dependiente, en este caso la media del salario, lo que puede llevar a una interpretación errónea de los resultados. Con esto el modelo ya no explica la variación respecto a la media de y.

e) Compute el R2 para la regresión con todas las variables incluyendo el término constante. Interprete los resultados.

El R^2 para la regresión con todas las variables incluyendo el término constante es 0.5161341 . Esto nos indica que aproximadamente el 51.6% de la variabilidad del salario puede ser explicada por las variables independientes incluidas en el modelo (educación, experiencia, habilidad, educación de los padres y número de hermanos).

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

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

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

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?

Este modelo no es posible estimar por mínimos cuadrados ordinarios (MCO) porque violaría el primer supuesto de linealidad, ya que no es lineal en los parámetros. La función de producción es una función no lineal en los parámetros α y β, lo que significa que no se puede expresar como una combinación lineal de los parámetros.

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

Para transformar el modelo de función de producción a una forma lineal, podemos tomar el logaritmo natural de ambos lados de la ecuación. Esto nos dará una forma lineal en los parámetros:

$$\log(Y_i) = \log(A_i) + \alpha \log(K_i) + \beta \log(L_i)$$

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.

Table 6: Promedio de la tasa de participación y contribución

Medida	prate	mrate
Promedio	87.36	0.73

ii) Ahora, estime el modelo

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

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

Table 7: Resultados del modelo de regresión 10

	Estimación	Error estándar	Valor t	Valor p
β_0	83.08	0.56	147.48	0
β_1	5.86	0.53	11.12	0

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

El intercepto de la ecuación representa el valor esperado de la tasa de participación (`prate`) cuando la tasa de contribución (`mrate`) es cero. En este caso, el intercepto es 83.08, lo que indica que si no hay contribución por parte de la empresa, la tasa de participación esperada sería de aproximadamente 83.08.

El coeficiente de `mrate` es 5.86, lo que indica que por cada aumento de una unidad en la tasa de contribución (`mrate`), la tasa de participación (`prate`) aumenta en aproximadamente 5.86 puntos porcentuales, manteniendo todo lo demás constante. Esto sugiere que una mayor generosidad del plan de pensión está asociada con una mayor participación en el plan.

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

$$prate = 83.08 + 5.86\ (3.5)$$

Cuando mrate = 3.5, la tasa de participación (`prate`) aproximada es 103.59. Este valor de 103.59 no es razonable, ya que la tasa de participación no puede ser mayor al 100 %. En este caso, la relación entre `prate` y `mrate` puede no ser lineal o puede haber un límite superior en la tasa de participación que el modelo no captura adecuadamente.

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

El valor de R-cuadrada del modelo es 0.075, lo que indica que aproximadamente el 7.47 % de la variación en la tasa de participación (`prate`) es explicada por la tasa de contribución (`mrate`).

Un 7.6 % de la variación en `prate` explicada por `mrate` puede considerarse relativamente bajo, lo que sugiere que hay otros factores no incluidos en el modelo que también influyen en la tasa de participación. En mi opinión, esto indica que la generosidad del plan de pensión es un factor importante, pero no el único determinante de la participación en el plan.

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 encuación que reconce estos factores es:

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

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

El signo más probable para β_2 es positivo. Esto se debe a que un nivel de ingresos más alto generalmente está asociado con un mejor acceso a cuidados prenatales y una mejor nutrición, lo que puede resultar en un mayor peso al nacer del bebé. Por lo tanto, se espera que un aumento en el ingreso familiar (`faminic`) esté asociado con un aumento en el peso al nacer (`bwght`).

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

Es probable que `cigs` (número de cigarrillos fumados durante el embarazo) y `faminic` (ingreso familiar) estén correlacionados. La correlación puede ser negativa, ya que las mujeres con ingresos más altos pueden tener un mayor acceso a información sobre los riesgos del fumar durante el embarazo y, por lo tanto, pueden fumar menos.

Por otro lado, también podría haber una correlación positiva si las mujeres con mayores ingresos tienen más recursos para comprar cigarrillos. Sin embargo, en general, se espera que la correlación sea negativa, ya que los ingresos más altos suelen estar asociados con una mayor conciencia de la salud y un menor consumo de tabaco.

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.

Resultados de la regresión de bwght sobre cigs y faminc

	Dependent variable:	
	Peso al nacer (bwght)	
	(1)	(2)
Cigarrillos (cigs)	-0.514***	-0.463***
	(0.090)	(0.092)
Ingreso familiar (faminc)		0.093***
		(0.029)
Constant	119.772***	116.974***
	(0.572)	(1.049)
Observations	1,388	1,388
R2	0.023	0.030
Adjusted R2	0.022	0.028
Note:	*p<0.1; **p<0.05; ***p<0.01	

Modelos estimado:

$$bwght = 119.772 \ - 0.514 \ cigs$$

$$bwght = 116.974 \ - 0.463 \ cigs \ + \ 0.093 \ faminic$$

Al agregar `faminic` *al modelo, el coeficiente de* `cigs` *cambia de -0.514 a -0.463, lo que indica que el efecto negativo del fumar durante el embarazo sobre el peso al nacer se reduce ligeramente al controlar por el ingreso familiar. Esto sugiere que parte del efecto negativo de* `cigs` *sobre* `bwght` *puede estar mediado por el nivel de ingresos, ya que las mujeres con mayores ingresos pueden tener un mejor acceso a cuidados prenatales y una mejor nutrición, lo que podría mitigar el impacto negativo del fumar.*

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?

El coeficiente β_1 representa el cambio esperado en el porcentaje de votos recibidos por el candidato A (`voteA`) por cada unidad adicional gastada en la campaña por el candidato A (`expendA`), manteniendo constantes los demás factores en el modelo. En otras palabras, indica cuánto se espera que aumente el porcentaje de votos para A por cada unidad adicional de gasto en su campaña.

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

Table 8: Resultados del modelo de regresión de voteA sobre expendA, expendB y prtystrA

Variable	Parametro	Error estándar	Valor t	Valor p
(Intercept)	33.267	4.417	7.532	0
expendA	0.035	0.003	10.365	0
expendB	-0.035	0.003	-11.636	0
prtystrA	0.343	0.088	3.894	0

Los coeficientes del modelo indican lo siguiente:

- *El coeficiente de* ***`expendA`*** *es positivo, lo que sugiere que un aumento en los gastos de campaña del candidato A está asociado con un aumento en el porcentaje de votos recibidos por él. Por cada unidad adicional gastada por el partido A, se espera que su porcentaje de votos* ***aumente*** *en aproximadamente 0.034 puntos porcentuales, manteniendo constantes los demás factores.**
- *El coeficiente de* ***`expendB`*** *es negativo, lo que indica que un aumento en los gastos de campaña del candidato B está asociado con una disminución en el porcentaje de votos recibidos por A. Por cada unidad adicional gastada por B, se espera que el porcentaje de votos para A* ***disminuya*** *en aproximadamente 0.035 puntos porcentuales, manteniendo constantes los demás factores.*

- *El coeficiente de* **prtystrA** *es positivo, lo que sugiere que una mayor fortaleza del partido del candidato A está asociada con un mayor porcentaje de votos recibidos por él. Por cada punto porcentual adicional en la fortaleza del partido de A, se espera que su porcentaje de votos* ***aumente*** *en aproximadamente 0.34 puntos porcentuales, manteniendo constantes los demás factores.*

El valor de R-cuadrada del modelo es 0.569, lo que indica que aproximadamente el 56.87 % de la variabilidad en el porcentaje de votos recibidos por A es explicada por los gastos de campaña y la fortaleza del partido.

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.

β_0 es el salario promedio cuando **educ**, **exper** *y* **tenure** *son cero. β_1 es el cambio esperado en el salario por cada año adicional de educación, manteniendo constantes los demás factores. β_2 es el cambio esperado en el salario por cada año adicional de experiencia laboral, para los individuos con el mismo nivel de educación y antigüedad. β_3 es el cambio esperado en el salario por cada año adicional de antigüedad en la empresa, para los individuos con el mismo nivel de educación y experiencia.*

ii) Estime el modelo e interprete los coeficientes.

Table 9: Resultados del modelo de regresión de wage sobre educ, exper y tenure

Variable	Estimación	Error estándar	Valor t	Valor p
(Intercept)	-276.240	106.702	-2.589	0.010
educ	74.415	6.287	11.836	0.000
exper	14.892	3.253	4.578	0.000
tenure	8.257	2.498	3.306	0.001

$$wage = -276 + 74.42 educ + 14.89 exper + 8.26 ternure$$

Cuando **educ** *aumenta en 1 año, el salario (*wage*) aumenta en aproximadamente 74.42 unidades monetarias, manteniendo constantes la experiencia (*exper*) y la antigüedad (*tenure*). Si* **exper** *aumenta en 1 año, el salario (*wage*) aumenta en aproximadamente 14.89 unidades*

*monetarias, manteniendo constantes la educación (**educ**) y la antigüedad (**tenure**). Finalmente, si **tenure** aumenta en 1 año, el salario (**wage**) aumenta en aproximadamente 8.26 unidades monetarias, manteniendo constantes la educación (**educ**) y la experiencia (**exper**).*

Lo que resulta poco intuitivo es el valor del intercepto, que resulta negativo. Esto indica que, si una persona no tiene educación, experiencia ni antigüedad, el modelo predice un salario negativo, lo cual no tiene sentido en la práctica. Sin embargo, esto es común en modelos de regresión lineal y no debe interpretarse literalmente.

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

Hipótesis nula:

$$H_0 : \beta_2 = \beta_3$$

Hipotesis alternativa:

$$H_a : \beta_2 \neq \beta_3$$

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

```
Linear hypothesis test:
exper - tenure = 0

Model 1: restricted model
Model 2: wage ~ educ + exper + tenure

  Res.Df       RSS Df Sum of Sq      F Pr(>F)
1    932 130732100
2    931 130437974  1    294126 2.0993 0.1477
```

*Con un valor **p-value** de 0.1477, no podemos rechazar la hipótesis nula al nivel de significancia del 5%. Esto sugiere que no hay evidencia suficiente para afirmar que un año más de experiencia en la fuerza de trabajo general **tiene un efecto diferente** sobre el salario (**wage**) en comparación con un año más de antigüedad en el empleo actual.*

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

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

Table 10: Resultados del modelo de regresión de log(wage) sobre educ, exper y tenure

Variable	Estimación	Error estándar	Valor t	Valor p
(Intercept)	5.497	0.111	49.731	0
educ	0.075	0.007	11.495	0
exper	0.015	0.003	4.549	0
tenure	0.013	0.003	5.170	0

Interpretación de los coeficientes

El coeficiente de `educ` *es 0.075, lo que indica que un año adicional de educación está asociado con un aumento del 7.5 % en el salario, manteniendo constantes la experiencia y la antigüedad. El coeficiente de* `exper` *es 0.015, lo que sugiere que un año adicional de experiencia laboral está asociado con un aumento del 1.5% en el salario, manteniendo constantes la educación y la antigüedad. Finalmente, el coeficiente de* `tenure` *es 0.013, lo que indica que un año adicional de antigüedad en la empresa está asociado con un aumento del 0.13% en el salario, manteniendo constantes la educación y la experiencia.*

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.