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

Trabajo Práctico 4

Errores de especificación – Errores de medición – Causalidad

Contenidos: Errores de especificación: omisión de variables relevantes. Errores de medición. Análisis de causalidad.

Fecha de entrega: jueves 19 de noviembre. Ver reglas de formato y presentación en https://econometria1unlp.com/trabajos-practicos/.

PARTE I: Omisión de variables relevantes

1. Ecuaciones de Mincer: sesgo por habilidad

En el Trabajo Práctico N° 2 se presentaron, estimaron y discutieron las ecuaciones de ingresos o ecuaciones de Mincer. Aquí volvemos a este tema con el objetivo de estudiar las consecuencias de no disponer de cierta información que se considera relevante.

a) Considerar que los salarios se determinan de la siguiente forma:

$$log\, w = edu\, \beta + habilidad\, \gamma + Z\, \phi + \mu \quad (1)$$

donde log *w* (vector Nx1) es el logaritmo del salario mensual, *edu* (vector Nx1) son los años de educación formal y Z (matriz Nxq) contiene otras variables explicativas incluyendo la constante (primera columna de unos). Se supone que:

- El vector Nx1 de términos aleatorios μ satisface los supuestos clásicos.
- La variable *habilidad* no está correlacionada con ningún otro regresor del modelo una vez que se controla por *edu*. Es decir, puede escribirse $habilidad = \delta_1 + \delta_2 edu + r$ donde *r* no está correlacionado con ninguna de las variables incluidas en el vector *Z*.

Debido a la dificultad, o aún imposibilidad, de medir todas las dimensiones que determinan la *habilidad* de un individuo, es usual considerar que la *habilidad* no es observable para el analista. Suponer que este modelo se estima por MCO omitiendo la variable *habilidad*. Sabemos que el retorno a la educación estimado $\hat{\beta}$ podría ser sesgado para β. ¿Cuál es el sesgo? ¿De qué depende la magnitud y el signo del mismo?

b) Suponer ahora que entre las variables incluidas en el vector *Z* están la experiencia laboral, la antigüedad en la empresa, el estado civil, la región de residencia y la raza. De esta manera el modelo (1) toma la siguiente forma:

$$\log w = \varphi_1 + \varphi_2\, exp + \varphi_3 ant + \varphi_4 casado + \varphi_5 sur + \ldots$$
$$\ldots + \varphi_6 urbano + \varphi_7 negro + \beta\, edu + \gamma\, habilidad + \mu$$

donde las variables se definen como:

Exp	Experiencia en el mercado laboral (en años)
Ant	Antigüedad en la firma (en años)
Casado	=1 si casado, =0 si no es casado
Sur	=1 para la región sur, =0 en caso contrario
Urbano	=1 si vive en zona urbana, =0 en caso contrario
Negro	=1 si raza negra, =0 en caso contrario
habilidad	No observable

Se supone que se satisfacen los supuestos enunciados en el inciso anterior.

Si bien la habilidad es inobservable, es factible de ser aproximada por algunas medidas de inteligencia, como por ejemplo el coeficiente intelectual (*CI*). En este sentido decimos que el *CI* es una *proxy* de la variable *habilidad*.

b.1) La base de datos *omision_de_habilidad.dta* contiene información sobre todas estas variables para 925 trabajadores varones.[1] A partir de estos datos, estimar por MCO el modelo que omite *habilidad*.

b.2) Estimar el modelo que reemplaza *habilidad* por la proxy *CI*.

b.3) Comparar el retorno a la educación que surge de los dos incisos previos.

b.4) Supusimos que la variable *habilidad* no está correlacionada con ningún otro regresor del modelo una vez que se controla por *edu*. ¿La variable *proxy CI* también satisface este supuesto? Argumentar la respuesta usando nuevamente los resultados obtenidos en los incisos (b.1) y (b.2).

[1] Esta base es una versión modificada de los datos utilizados en el trabajo de Blackburn y Neumark (1992). "Unobserved Ability, Efficiency Wages, and Interindustry Wage Differentials", Quarterly Journal of Economics 107, 1421-1436.

PARTE II: Errores de medición

2. Un ejercicio numérico

Utilizando los datos del archivo *gasto-ingresos.dta* sobre gasto e ingresos de los hogares, estudiaremos el comportamiento de los estimadores de mínimos cuadrados ordinarios con una serie de ejercicios en donde crearemos artificialmente un error de medición.

a) Como primer paso, regresar el gasto en alimentos en función del ingreso por MCO, utilizando los datos verdaderos. Interpretar estadística y económicamente los coeficientes.

b) Con el objetivo de imputar artificialmente un error de medición en las variables de regresión, generar computacionalmente una variable aleatoria normal con media cero y desvío estándar 100. Mostrar estadísticas descriptivas del error.

c) Generar una variable de gasto artificial definida como la suma de la variable *gasto_alimentos* y el error de medición aleatorio creado en el inciso (b). Utilizarla como variable dependiente para correr una regresión por MCO en función del ingreso. Comparar el resultado con las estimaciones del inciso (a) y explicar los cambios observados.

d) Imputar ahora el error de medición del inciso (b) a la variable de ingreso original y estimar un modelo de MCO que utilice al gasto original en función de este ingreso generado. Nuevamente, explicar los cambios con respecto a la regresión del inciso (a).

e) Repetir (c) y (d) suponiendo dos valores alternativos para la variabilidad de los errores de medición. En particular, considerar un desvío estándar de 500 y otro de 1000. Explicar los resultados observados sobre los estimadores de MCO.

PARTE III: Regresiones y Causalidad

3. Promedios, proporciones y regresiones

Esta sección contiene una serie de ejercicios computacionales y analíticos que tiene como objetivo mostrar que varios de los procedimientos estándar utilizados para realizar inferencia estadística (cómputo de promedios y proporciones, test de medias, etc.) pueden ser reinterpretados como un modelo de regresión.

Para la parte computacional utilizaremos el archivo *brechas-genero.dta* que contiene datos pertenecientes al aglomerado La Plata, extraídos de la Encuesta Permanente de Hogares (EPH) realizada en 2019, semestre 2. La base contiene datos de salarios y género para individuos ocupados al momento de la encuesta.

a) Mostrar numéricamente que la proporción de hombres en la muestra coincide con el promedio muestral de la variable binaria “hombre”. Explicar analíticamente el resultado anterior.

b) Considerar un modelo lineal que no contiene ninguna variable explicativa, es decir solo consta de una ordenada al origen α más un término μ_i que es aleatorio y cumple con los supuestos clásicos. Mostrar formalmente que el estimador de MCO de este modelo coincide con la media muestral de *Y*. Corroborar computacionalmente dicho resultado utilizando la variable *salario*.

c) Calcular empíricamente la brecha salarial entre hombres y mujeres y realizar un test de diferencia de medias. Comparar estos valores con los que arroja una regresión en donde la variable dependiente es el salario y la única variable explicativa es el género.

d) Explicar analíticamente el resultado anterior. Notar que se trata de un caso particular del modelo lineal con dos variables, en donde la variable explicativa es binaria. Es decir:

$$Y_i = \alpha + \beta D_i + \mu_i \qquad \text{con } i = 1, \dots, n$$

donde Y_i es el salario y D_i es una variable binaria que indica con 1 si se trata de un hombre y 0 en caso contrario. Mostrar formalmente que en este caso particular el estimador de MCO de α coincide con el salario promedio de las mujeres y el de β con la diferencia salarial promedio entre hombres y mujeres.

4. Datos no experimentales

La Asignación Universal por Hijo (AUH) es un programa de protección social de Argentina vigente desde 2009, destinado a los hijos de las personas que están desocupadas o que trabajan en el sector informal[2]. El cobro de la AUH requiere la acreditación anual de escolarización y controles de salud de los niños. Se abona a los menores de 18 años, hasta un máximo de 5 hijos, priorizando a los hijos discapacitados y a los de menor edad[3].

El archivo *auh-engho-2012.dta* contiene una muestra extraída de la Encuesta Nacional de Gasto de los Hogares (ENGHO) realizada en 2012. Se trata de 1296 observaciones correspondientes a la región Noroeste de Argentina, de los cuales 194 son beneficiarios de la AUH y el resto son individuos que si bien no están dentro del programa cumplen con las condiciones del mismo.

a) Computar un test de diferencia de medias para estudiar el efecto de la AUH sobre las horas trabajadas. Luego, computar el mismo test utilizando una regresión de MCO. Interpretar los resultados desde el punto de vista económico y estadístico.

b) Suponer que un economista ve el resultado anterior y argumenta lo siguiente: "Además de la presencia de la AUH, existen otros factores que afectan la decisión de la cantidad de horas que los individuos deciden trabajar, tales como la edad, el nivel educativo y el género". Realizar nuevamente un test de diferencia de medias pero para

[2] En años posteriores se incorporaron como beneficiarios a los hijos de monotributistas sociales, empleados del servicio doméstico, trabajadores temporarios en el período de reserva del puesto o personas que perciban otros planes sociales, como Argentina Trabaja. Los trabajadores deben percibir ingresos mensuales inferiores al salario mínimo, vital y móvil. Sin embargo, en la práctica esta restricción es difícil de controlar por lo que resulta ser no operativa.

[3] Para más detalle dirigirse a la página web de la ANSES: https://www.anses.gob.ar/

cada una de las características mencionadas. ¿Considerarían que el grupo de tratados es comparable con el grupo de control? Comentar sobre las razones detrás de los resultados observados.

c) Computar una regresión que tenga como variable dependiente las horas trabajadas, y las siguientes variables explicativas: una dummy que indique si el individuo recibe AUH, la edad, dummies por nivel educativo y el género. Comentar los resultados de dicha estimación y compararlos con el efecto del programa estimado en el inciso (a). ¿A qué se deben las diferencias? ¿Consideraría que ahora los grupos son comparables?

d) Utilizando la fórmula del sesgo por omisión de variable relevante, replicar numéricamente la diferencia estimada en los puntos (a) y (c) para el efecto del programa sobre el promedio de las horas trabajadas.

e) Imaginar que existen factores inobservables que son relevantes para explicar las horas de trabajo y que no han sido incluidos en el modelo previo. ¿Bajo qué circunstancias la omisión de estos factores inobservables no generarían un sesgo en la estimación del efecto de la AUH?

5. Datos experimentales

Este ejercicio está basado en el trabajo de Attanasio et al. (2011), donde se utilizan datos experimentales para evaluar el efecto del programa colombiano *Jóvenes en Acción* sobre algunas variables laborales.[4] Se trata de un programa de entrenamiento laboral que consiste en una serie de cursos y pasantías en distintas tareas a unos 80 mil jóvenes entre 18 y 25 años de edad provenientes de hogares con ingresos bajos[5]. Una característica particular es que en su última etapa no alcanzaron los cupos disponibles para todos los cursos y se hizo asignación aleatoria de los candidatos a los cursos de entrenamiento.

El archivo *attanasio-et-al-2011.dta* contiene una sub-muestra de la base utilizada por los autores. La misma contiene datos de 2113 mujeres entre 18 y 26 años de las cuales 1078 recibieron entrenamiento laboral (grupo de tratamiento) y las otras 1035 quedaron fuera del programa (grupo control). Las variables utilizadas son:

- *tratado*: variable binaria que vale 1 si el individuo participó del programa y vale 0 en caso contrario.
- *salario*: ingreso mensual de asalariados, transcurrido un año luego de la finalización del programa. Vale cero para desocupados e inactivos.
- *beneficio*: ingreso mensual de cuentapropistas, transcurrido un año luego de la finalización del programa. Vale cero para desocupados e inactivos.
- *horas*: horas de trabajo semanales, transcurrido un año luego de la finalización del programa. Vale cero para desocupados e inactivos.

[4] Attanasio, O., A. Kugler & C. Meghir (2011) "Subsidizing Vocational Training for Disadvantaged Youth in Colombia: Evidence from a Randomized Trial," American Economic Journal: Applied Economics, American Economic Association, vol. 3(3), pages 188-220, July.

[5] Los beneficiarios del programa también reciben una transferencia monetaria condicionada. Para más información ver www.jovenesenaccion.com

- *edad*: años de edad
- *casado*: variable binaria que vale 1 si el individuo es casado y vale 0 en caso contrario
- *educ*: años de educación completados.

a) Computar un test de diferencia de medias entre los grupos de tratamiento y control para estudiar el efecto del Plan Jóvenes en Acción sobre los salarios. Interpretar los resultados.

b) Examinar el balance de las características mencionadas anteriormente entre el grupo de tratamiento y de control. Comentar si el resultado es coherente con lo que se espera en un diseño experimental.

c) Además del entrenamiento recibido en el programa, es posible que las diferencias en salarios se deban a otras características tales como la edad, educación o la situación marital. Evaluar el efecto del Plan Jóvenes en Acción, pero controlando por el efecto potencial que tienen las características mencionadas anteriormente sobre los salarios. Comentar los resultados.

d) Comparar este caso de evaluación de impacto de un plan social con la situación planteada en el Ejercicio 4. ¿Por qué se dice que el diseño experimental permite identificar el efecto causal del programa sobre las horas trabajadas? Dar argumentos formales e intuitivos.

e) Evaluar el impacto del programa en otras variables laborales. En particular, sobre las horas trabajadas e ingreso mensual de cuentapropistas.