

Modelación en Ciencia de Datos

Práctica 2

Variables Instrumentales

Hay muchos estudios en economía laboral que tratan el tema de la estimación de las funciones de ingresos del capital humano y que establecen cómo los ingresos salariales están determinados por la educación y la experiencia laboral. Un ejemplo destacado es Card (1993), que investiga el rendimiento económico de la educación y utiliza la proximidad a la universidad como variable instrumental.

Esta práctica utiliza el conjunto de datos **CollegeDistance**, que es similar a los datos utilizados por Card (1993). Proviene de una encuesta de graduados con variables codificadas como salario, educación, matrícula promedio y una serie de medidas socioeconómicas. El conjunto de datos también incluye la distancia desde una universidad mientras los participantes de la encuesta estaban en la escuela. **CollegeDistance** viene con el paquete **AER**.

La regresión del salario en la educación y las variables de control para estimar la función de ingresos del capital humano es problemática porque la educación no se asigna al azar entre los encuestados: las personas toman sus propias decisiones de educación y, por lo tanto, las diferencias medidas en los ingresos entre las personas con diferentes niveles de educación dependen de cómo estas elecciones son hechos.

En la literatura, esto se conoce como un problema de selección e implica que la educación es endógena, por lo que la estimación de MCO estará sesgada y no podemos hacer una inferencia válida con respecto al coeficiente verdadero.

Instrucciones

1. Adjunte el paquete **AER** y cargue los datos de **CollegeDistance**.
2. Obtenga una descripción general del conjunto de datos. Específicamente, tome en cuenta que la variable distancia (la distancia a la escuela más cercana en 10 millas) servirá como instrumento en las estimaciones. Utilice un histograma para visualizar la distribución de la distancia.
3. Estime dos regresiones cuyos resultados no son confiables debido al problema de selección. Guarde estos resultados para compararlos posteriormente

con los obtenidos mediante el enfoque de variables instrumentales aplicado por Card (1993). Estime por ejemplo el modelo:

$$\log|wage| = \beta_0 + \beta_1 \log|education| + u$$

y luego incluya otros regresores como: desigual, hispano, af-am, femenino y urbano. Interprete sus resultados.

4. ¿Por qué se puede utilizar la distancia a la escuela como instrumento? Justifique el motivo por el cual esta variable podría ser utilizada como un instrumento.
5. Calcule las correlaciones de la distancia del instrumento con la educación regresora endógena y la variable dependiente salario. ¿Qué parte de la variación en la educación se explica por la regresión de la primera etapa que utiliza la distancia como regresor?
6. Repita las estimaciones anteriores utilizando IV, es decir, utilice la distancia como un instrumento para la educación en ambas regresiones mediante la función `ivreg()`. Guarde sus resultados y obtenga los errores estándar robustos para ambos modelos.
7. Verifique que sus resultados coinciden con los obtenidos cuando utiliza el procedimiento de estimación mediante las dos regresiones de MC2E para ambos modelos.
8. Interprete sus resultados en el contexto del problema, esto es, cuáles son los rendimientos de la educación estimados y el efecto del sesgo de selección en el modelo.
9. Realice las pruebas de verificación de variables instrumentales (endogeneidad de la regresora, relevancia del instrumento y exogeneidad del instrumento).
10. Pruebe si su instrumento es débil. Tome en cuenta que las familias con un fuerte interés por la educación pueden mudarse a vivir cerca de las universidades. Además, los vecindarios cercanos a las universidades pueden tener mercados laborales más fuertes reflejados en ingresos más altos. Tales características pueden invalidar el instrumento, ya que introducen variables no observadas que influyen en los ingresos, pero que no pueden ser capturadas por años de escolaridad, la medida de educación.