**QLAB - Junio 2023**

**Práctica Calificada - Módulo de IV: Replicando a David Card, Premio Nobel de Economía 2021**

Docente: Dr. Juan Manuel del Pozo Segura

Para la realización de este ejercicio se requiere utilizar el software STATA, pero se permite el uso de otro programa estadístico siempre que pueda realizar lo solicitado. Se requiere la entrega de un informe en word o pdf (latex) con las respuestas a los ejercicios. Este informe deberá ser elaborado de forma grupal en grupos de hasta 5 personas. La calificación es sobre 20 puntos

Usted cuenta con la base de datos card.dta, correspondiente al paper “ Using Geographic Variation in College Proximity to Estimate the Return to Schooling” de David Card, 1993. Su primera tarea es leer el artículo, esto facilitará el desarrollo de esta tarea

Sobre la base de los datos y lo visto en clases, responda lo siguiente (2 puntos cada una):

1. Realice una descripción de las características de los individuos de la muestra. ¿Qué información nos provee el promedio de la dependiente en la interpretación de los coeficientes?¿Cómo se interpretan los coeficientes en este caso?

**Rpta:**

En el artículo, se utiliza una muestra obtenida de Young Men Cohort of the National Longitudinal Survey, en la que desde 1966 se encuestó a 5525 varones entre 14-24 años hasta 1981. Para trabajar dicha muestra, se tomó en cuenta lo siguiente:

* Porcentaje de la distribución de la edad en 1966: 14-15, 16-17, 18-20 y 21-24
* Porcentaje de la distribución regional en 1966: Northeast, Midwest, South y West
* Porcentaje de las personas que vivían en Standard Metropolitan Statistical Area
* Porcentaje de personas que vivían cerca al colegio
* Porcentaje de la estructura familiar de las personas de 14 años: madre y padre o solo madre
* Promedio educativo de los padres en años
* Porcentaje de personas fenotípicamente morenas
* Promedio de puntaje en el *Knowledge of the World of Work* Test
* Porcentaje de entrevistados en 1976
* La media educativa en 1976
* Porcentaje de personas que vivían en el sur en 1976

Respecto a la variable dependiente, se expresa que sería el logaritmo de los salarios por hora en 1976. Asimismo, el coeficiente de educación estimada es estable entre las especificaciones e implica una ventaja de ingresos del 7.3 – 7.5% por cada año adicional de educación, control por experiencia y otros factores

1. Regresione usando OLS el siguiente modelo en Stata

reg lwage educ c.exper##c.exper age black south smsa, smsa66, reg661-reg668

¿Qué ocurre con edad? ¿Por qué? Ahora saque edad del modelo anterior y regresione de nuevo. Interprete el estimado obtenido para educación, y compárelos con la Tabla 2, Columna 2, del paper

**Rpta:**

Sobre la variable “edad”:

Al correr el modelo, Stata establece que “edad” ha sido omitida debido a la colinealidad, pues existe una alta correlación entre las variables predictoras, lo cual podría generar inestabilidad o estimaciones poco confiables en el análisis de regresión.

Sobre la variable “educ”:

Al correr nuevamente el modelo, se observa que el coeficiente de “educ” es 0.0746933, similar al que aparece en la Tabla 2 del paper, donde se redondea el valor (0.075). Asimismo, el error estándar se mantiene en ambos 0.0034983. Basándonos en ello, existe una variabilidad esperada en torno al estimado del coeficiente de la variable “educ”. El intervalo de confianza del 95% para el coeficiente se encuentra entre los valores 0.0678338 y 0.0815527

1. Vea la Tabla 2 de Card (1993). ¿Qué se encuentra al inspeccionar el coeficiente estimado de educación? ¿Es creíble este coeficiente? ¿Por qué? ¿Qué 2 posibles fuentes de inconsistencia podrían ocurrir como *resultado de la omisión de variables*? Use el marco visto en clase (las expresiones que forman parte de A picture containing logo

   Description automatically generated) para responder y mencione la dirección esperada del sesgo.

**Rpta:**

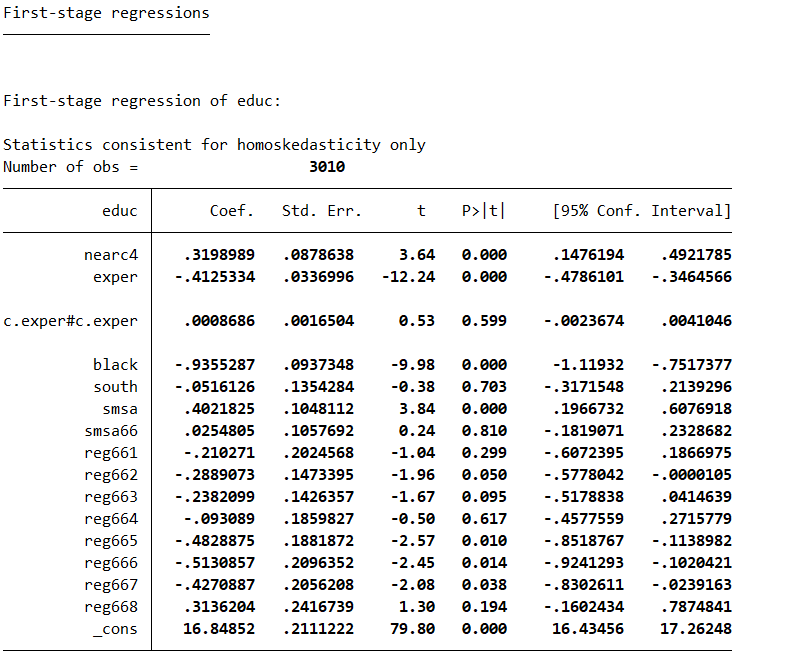
De acuerdo con la Tabla 2, específicamente la variable “educ”, el valor del coeficiente estimado es 0.075, lo que indica que se espera un incremento en ese valor por cada unidad de aumento. Debido a que el error estándar es de 0.003, es decir, se ha calculado la precisión del estimado con un valor de t que de 21.35 que indica que el coeficiente es altamente significativo y el valor p 0.000 expresa que la educación si genera un efecto sobre los salarios, entonces sí es creíble que pueda surgir este coeficiente (Para esto nos basamos en el modelo de regresión en Stata).

Asimismo, entre los sesgos posibles podríamos encontrar “sesgos de omisión de variables” o “sesgo de omisión de variables de interacción”. No obstante, la dirección del sesgo va a depender de la relación entre la variable omitida y la variable educación como la variable dependiente. Todo dependerá de si se encuentra positiva o negativamente relacionada

1. Estime la ecuación de primera etapa ¿Qué debemos ver en esta regresión? ¿Qué podría decir respecto de la correlación parcial existente entre educ y nearc4?

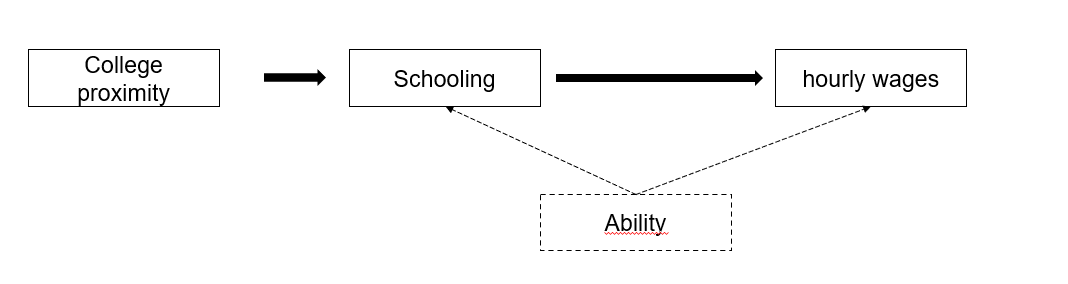
**Rpta:**

Debemos observar si el instrumento es un buen predictor de la variable independiente de interés. En este caso, se puede observar que nerac4 es muestra una correlación significativa con variable “educ” dado que el p value nos hace rechazar la hipótesis nula que esta correlación es 0.



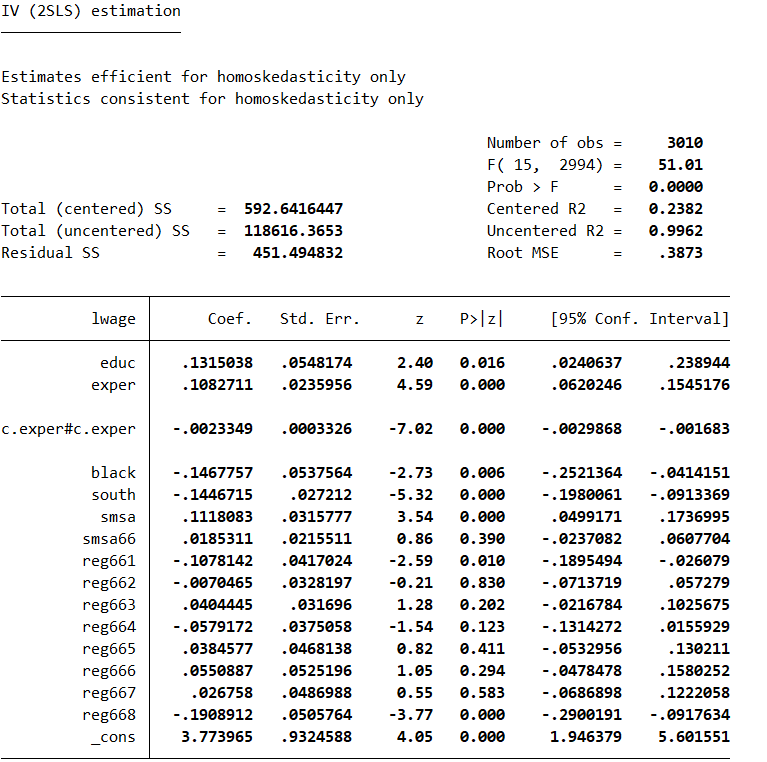
1. Estime el modelo usando el estimador de Variables Instrumentales, usando nearc4 como un instrumento para la variable *educ*. Dibuje el DAG correspondiente y explique por qué se espera que esta estrategia funcione. Comente sus resultados y compárelos con los obtenidos en el punto 2.

**Rpta:**



De acuerdo a la hipótesis del paper, las habilidades es una variable que introduce sesgo pues afecta tanto a los salarios por hora como al nivel de schooling alcanzado por la persona. Por ello, se utiliza la cercanía a colegios como una variable instrumental que no se ve afectada por la habilidad y explica el nivel de schooling para así determinar el efecto en el salario por hora.

A continuación, se muestran los resultados obtenidos usando el estimador de variables instrumentales. Estos resultados concuerdan con la tabla 3 del paper:



A diferencia de la regresión estimada en la pregunta 2, aquí se observa un incremento aproximado en 0.6 puntos porcentuales del efecto de la educación en el logaritmo de los salarios por hora. Asimismo, este efecto es significativo al 95%.

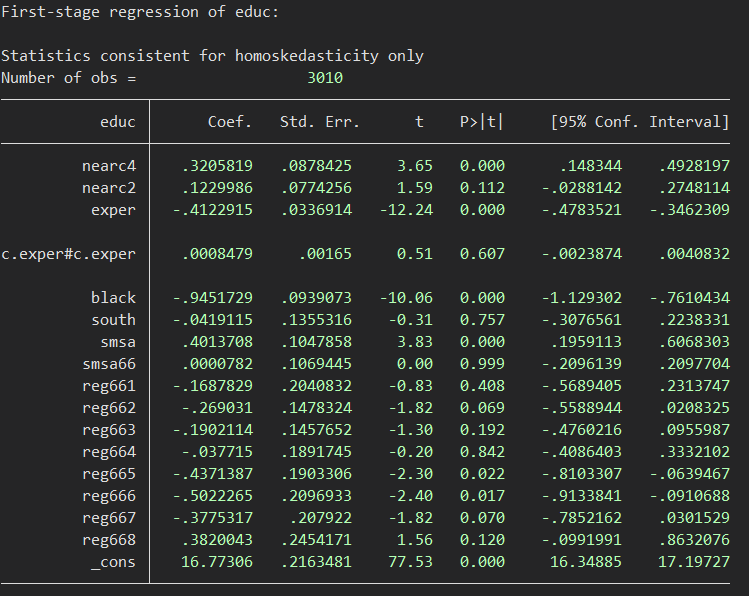
1. Testee la presencia de instrumentos débiles usando i) el estadístico F; ii) el estadístico de Cragg y Donald (1993); iii) las tablas de Stock y Yogo (2005) con respecto a la medida del test de Wald

**Rpta:**

* Estadístico F: Observamos que el estadístico F muestra un p value menor a 0.05 lo que implica que hay evidencia suficiente concluir que existe una relación o efecto significativo.
* Estadístico de Cragg y Donald (1993): El estimador Cragg y Donald indica 13.26
* Tabla Stock y Yogo: Tenemos un problema de instrumento débil porque 13.26 es menor al umbral de 10%. Sin embargo, es mayor al 15% del *maximal IV size.* Es decir, estamos dispuestos a asumir un sesgo del estimador IV del orden del 15% del sesgo del estimador MCO

1. Use nearc2 con nearc4 como instrumentos para educ. Primero estime la primera etapa para educ y analice cuál de los dos instrumentos está más fuertemente relacionada parcialmente con educ. Después use el estimador de IV usando por separado nearc2 y nearc4 como instrumento para educ. Luego use el de 2SLS (incluyendo ambos instrumentos). Discuta sus resultados. Verifique que el estimador 2SLS es válido.

**Rpta:**



Como se puede observar, en la primera etapa de la estimación por 2SLS con las ambos instrumentos se identifica que solo la variable nearc4 es significativa al 95% de confianza. Por el contrario, el instrumento nearc2 no es significativo en esta etapa. Además, la variable nearc4 está más fuertemente relacionado con educ.

Una captura de pantalla de un celular

Descripción generada automáticamente con confianza media

Como era de esperarse, al utilizar solo nearc2 como instrumento para la educación los resultados no son significativos para la estimación. Utilizando únicamente nearc4 como instrumento para la educación, los resultados muestran que por cada año adicional de educación los ingresos aumentan aproximadamente en 13.2% significativos al 5% de significancia. Sin embargo, cuando se incluyen ambos instrumentos (nearc2 y nearc4) en la primera etapa, los resultados muestran que por un incremento de un año de educación tiene un efecto positivo del 15.7% en los salarios con un nivel de significancia del 1%.



En cuanto a la validez del modelo 2SLS con ambos instrumentos, observamos que el test de Sargan tiene un p-value de 0.2639 por lo que no se puede rechazar la hipótesis nula de que ambos instrumentos son válidos. Es decir, no existe evidencia suficiente para descartar la validez de los instrumentos.

1. Hemos asumido hasta ahora homoscedasticidad ¿Es razonable esto?¿Qué permite el estimador de GMM que no permite el de IV en 2 etapas bajo heterosk?. Compare los resultados (del coeficiente educación) del estimador 2SLS con el del estimador GMM bajo heterocedasticidad

**Rpta:**

**Pantalla de computadora con letras

Descripción generada automáticamente con confianza media**

El método GMM permite incorporar directamente la heterocedasticidad de los errores a la estimación de los parámetros. Esto permite que pueda obtener estimaciones más eficientes en presencia de homocedasticidad, al mismo tiempo que es un estimador robusto ante la presencia de endogeneidad o la no normalidad de los errores. En las estimaciones realizadas, se puede observar que existe una diferencia tanto en el valor del coeficiente entre el método 2SLS y método GMM, al mismo tiempo que observamos un mejor ajuste del modelo, pues el método GMM al ser más eficiente ante heterocedasticidad, permite un mejor ajuste de la estimación.

1. Sigamos con el caso donde se tienen dos instrumentos. Dado lo hallado en términos de si los instrumentos por separado son débiles o no en la pregunta 7, ¿qué concluye ud. a partir del resultado del test de sobreidentificación? ¿Con qué instrumento o instrumentos se quedaría al final? Realícelo asumiendo heterocedascidad para el estimador GMM y 2SLS (HINT: piense 2 veces)

**Rpta:**

1. ¿Deberíamos usar MCO y aceptar la inconsistencia o usar en vez IV a pesar de que es un estimador consistente pero ineficiente? ¿Hay forma de testear esto? Si sí, Llévela a cabo para el modelo sobre-identificado. ¿Qué concluye?

**Rpta:**

Gabriela Calvo, Mauricio Ibáñez, César Núñez, Rodrigo Soto

Junio 2023