**QLAB - Junio 2023**

**Práctica Dirigida - Módulo de IV**

Docente: Dr. Juan Manuel del Pozo Segura

# Efecto de college sobre register

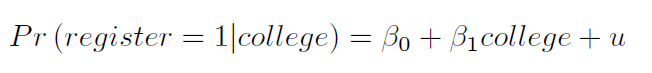
1. A partir de la base de datos “murnane10.dta”, halle el efecto la variable College (que indica si la persona asistió a la universidad junior, comunitaria o de 4 años en 1984?) sobre Register (si la persona está actualmente registrado para votar) a través de OLS. ¿Considera que el efecto estimado es grande? ¿hay motivos para suponer que el regresor College es un regresor endógeno? Si sí: ¿qué puede estar causando un sesgo de variable omitida? Dibuje el DAG.

*Buscamos estudiar el efecto de haber ido a un college sobre el registrarse para votar. Antes de embarcarnos en la regresión, siempre es necesario crear estadísticos descriptivos de las variables de interés. ¿Por qué? Pues porque la magnitud del efecto de x sobre y, sobre todo el caso donde x es binaria, se juzga en términos de cuánto cambia el promedio de y al pasar de x=0 a x=1. Esto es particularmente importante este ejemplo, pues la variable x (college) es binaria.*

**

*Vemos que la proporción de aquellos que han ido a un college representa 54% del total de observaciones, mientras que 67% de individuos se han registrado como votantes. Vemos que entre quienes no votan (3037), la mayoría no han ido a un college, mientras que lo opuesto se encuentra entre los 6190 que sí votan .*

*¿Cuál es el efecto casual de haber ido a un college en la probabilidad de que la persona se registre para votar? Como se mencionó en clase, es necesario primero pensar qué queremos estimar en términos de la población. La ecuación estructural de interés viene dada por*

**

*Nótese la diferencia con lo que se ha visto en clase, donde la dependiente era una variable continua. En cambio, en esta PD, la dependiente es una variable binaria, que toma valores 0 (no se registra) y 1 (sí se registra). Por tanto, lo que estamos estimando es el efecto de haber ido a un college en la probabilidad condicional de que la persona se registre para votar (register=1). En tanto estamos usando un modelo lineal para describir una probabilidad, la ecuación anterior se le conoce como Linear Probability Model (LPM). El procedimiento de estimación es el mismo que lo visto en clase: la variable register entra en las fórmulas vistas en la clase como si fuera una variable continua cualquiera. Nada cambia, y podemos definir la ecuación de interés, en error-form, como*

**

*(Uno podría preguntarse, ¿por qué no estimamos la ecuación usando un modelo no lineal como Probit/Logit en vez de LPM si la dependiente es binaria? Hay 2 razones que abonan al uso de modelos lineales como LPM: i) método IV con modelos no lineales como Probit/Logit [en vez de IV con modelos lineales] es conceptualmente más complicado e implica el uso de supuestos distribucionales que pueden no ser válidos; y más importante ii) El estimador MCO provee coeficientes tan válidos como los de Probit/Logit cuando la variable dependiente es binaria. Es por eso que varios papers toman modelos lineales estimados por MCO y por IV como punto de partida de las estimaciones, ¡y muchas veces los resultados se basan en estos estimadores sencillos!)*

*Estimemos este LPM por MCO*

**

*Encontramos que, en promedio, la probabilidad de registrarse para votar es 17 puntos porcentuales más altos para aquellos que han ido a un college que aquellos que no. Este efecto es estadísticamente diferente de 0. Dado lo visto al inicio, ¿el coeficiente que acabamos de estimar, A picture containing font, typography, text, white

Description automatically generated, puede considerarse grande? El haber ido a college, en promedio, incrementa la probabilidad de que la persona se registre para votar de 0.67 a 0.67+0.17=0.84. ¡Este es un incremento de 25%! Por tanto, el efecto puede definitivamente considerarse grande.*

*La pregunta de oro: ¿este efecto puede considerarse causal? No existe absolutamente NADA en los datos que nos diga si este efecto puede considerarse causal o no. Como se mencionó en la clase, esto depende del juicio del autor y de los argumentos que pueda esgrimir. Un punto de partida es pensar: ¿qué variable puede causar un sesgo de variable omitida? I.e. ¿qué determinante legítimo del hecho que la persona se registre para votar no se observa (y por tanto está en el error) y* ***a la vez*** *está correlacionado con la probabilidad de que la persona tenga estudios superiores? No existe una respuesta única (uds. pueden proponer otro factor). Pero podemos afirmar que, por ejemplo, la importancia que le da una persona a la sociedad*

* *está correlacionado con la probabilidad de que la persona obtenga educación superior (y), pues toma en consideración la valoración social en su decisión de obtener educación superior o no*
* *está correlacionado con el hecho que la persona vote (x), pues tiene incentivos para contribuir con cambios que considera necesarios mediante su voto, por lo que el registro es necesario*

*Por tanto la ecuación que deberíamos estimar, en vez de* , es *la siguiente*

**

*Sin embargo, la variable* ***m*** *que mide la importancia que le da una persona a la sociedad (civismo) no se observa y por tanto esta ecuación (larga) no es estimable. Dicho de otro modo*

* *Queremos estimar* 
* *Sólo podemos estimar*

*Por tanto, en la ecuación (corta) que hemos estimado arriba por MCO,* *es un error compuesto, pues incluye la variable* ***m*** *que no observa y que está creando un problema de variables omitidas. Consecuentemente, el estimado obtenido por MCO, A picture containing font, typography, text, white

Description automatically generated, no refleja un efecto causal. Dicho de otro: , el resultado A picture containing font, typography, text, white

Description automatically generated= 0.17 no sólo refleja el efecto sobre la probabilidad de que la persona se registre que únicamente puede atribuirse a la educación (que la persona haya ido a un college), sino que está confundido, contaminado, distorsionado, por otros factores.*

*Esto implica que enfrentamos un problema de endogeneidad del regresor college, que puede representarse gráficamente usando DAG como*

*A picture containing text, font, line, diagram

Description automatically generated*

*¿Qué se necesita? Un instrumento que se relacione con el haber ido a un college y register de una forma especial. Esquemáticamente*

*A picture containing text, font, line, diagram

Description automatically generated*

1. Considere como instrumento la variable Distance (distancia a la escuela más cercana cuando la persona estaba en secundaria) y explique intuitivamente por qué cumpliría los supuestos del modelo de variables instrumentales (IV). Luego estime el modelo de IV usando como instrumento es la variable Distance.

*Recordemos que la validez de una variable instrumental descansa en 2 supuestos*

* *: esto significa que el hecho de haber vivido cerca de una escuela cuando la persona era menor de edad está relacionado con el hecho que la persona haya asistido al college. Esto es creíble. Sin embargo, puede* ***y debe*** *testearse empíricamente*
* *: esto significa que el hecho de haber vivido cerca de una escuela cuando la persona era menor de edad* ***no*** *está relacionado con otros factores no observados que determinan el registro de voto, en particular sentido cívico. Esto* ***no puede*** *testearse empíricamente y sólo nos queda la argumentación para llevar a cabo esta verificación. ¿Es razonable asumir la 0 covarianza en este caso? Pensemos en qué casos fallaría: si aquellos que viven cerca de una escuela tienen un sentido cívico mayor. Esto es poco plausible (al menos para el conocimiento limitado que tenemos los economistas sobre fenómenos sociales). Recuérdese que el factor en el que nos estamos centrando es civismo, debido a que hemos elegido dicha variable como aquella que está creando un problema de variables omitidas. En caso de que elijamos otra variable, debemos argumentar la correlación con dicha variable.*

*Habiendo generado argumentos para pensar que el instrumento es razonablemente válido, podemos estimar el modelo usando variables instrumentales. En tanto tenemos un modelo exactamente identificado, basta usar el estimador de variables instrumentales. El comando ivregress 2sls, por el contrario, estima todos los modelos (exactamente identificados y sobreidentificados) por 2SLS, el cual es necesario solo para modelos sobreidentificados. Sin embargo, en un modelo exactamente identificado, como el que tenemos, el estimador de IV y el 2SLS provee los mismos resultados.*



*Luego de escribir ivregress, primero, colocamos cuál es el tipo de estimador a emplear. Nosotros escogemos que este será el de 2sls. Luego, indicamos cuál es la variable dependiente. En nuestro caso, esta es . Finalmente, entre paréntesis colocamos primero cuál es la variable a ser instrumentalizada, , y luego cuál es el instrumento ().*

*Los resultados muestran que el coeficiente estimado por el método de IV, , es 0.28. Es decir, la probabilidad de registrarse para votar es 28 puntos porcentuales más altos para aquellos que han ido a un college que aquellos que no. Este efecto es estadísticamente diferente de 0. Dado lo visto en la pregunta 1, ¿es este coeficiente grande? El haber ido a college, en promedio, incrementa la probabilidad de que la persona se registre para votar de 0.67 a 0.67+0.28=0.95. ¡Este es un incremento de 40%! Un efecto definitivamente grande.*

*Una parte importante de la estimación por IV corresponde a analizar la relación que tiene el instrumento con el regresor endógeno. Como se ha enfatizado en la clase: no basta con que el A black text on a white background

Description automatically generated with low confidence, más bien queremos que la relación entre ambos, condicional a las otras variables, sea fuerte. Como se mencionó, esto se mide por el F test sobre la variable instrumental. Esto puede evaluarse a partir del primer panel, titulado “First-stage regressions”. Vemos, sin embargo, que este muestra solo el test t. En este caso simple, el test F corresponde al cuadrado del estadístico t, igual a -10^2=100. Por tanto, nuestro estadístico F es aproximadamente 100. Si estamos dispuestos a aceptar como máximo un sesgo de nuestro estimador IV del orden del 10% del sesgo del estimador MCO, el test F debe ser mayor a 10. En nuestro caso, instrumento claramente NO sufre del problema de weak instruments.*

*Asimismo, vemos que la relación entre el instrumento y el regresor endógeno es negativa, tal como se espera. Una relación estimada con el signo contrario proveería indicios de que el instrumento es inválido, no porque es débil, sino porque no cumple la exclusión restriction.*

1. Digamos que tenemos motivos para creer que la distancia a la escuela está relacionada con la etnicidad de la persona (e.g. hispanos viven más lejos de una escuela) y la etnicidad impacta en la probabilidad de registrarse para votar. ¿Qué significa esto para nuestro DAG? Dibuje el nuevo DAG y estíme el nuevo modelo correspondiente

*Nótese que el modelo que hemos estimado*

**

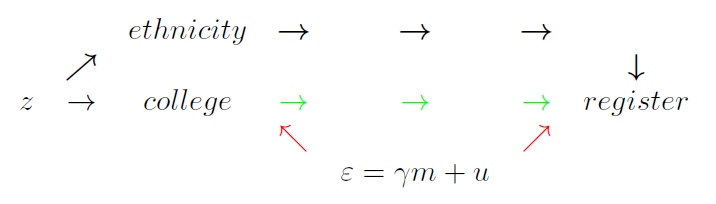
*no contiene controles. Después de todo, lo que nos interesa es el efecto de college sobre register. En virtud de que college es un regresor endógeno, nuestro modelo en realidad es*

*A picture containing text, font, line, diagram

Description automatically generated*

*por lo que hemos usado el estimador de IV.*

*En principio, el instrumento sólo afecta indirectamente a la dependiente a través de 1 canal: college. Pero ahora tenemos motivos para creer que hay otro canal por el cual afecta a la dependiente: por el canal de etnicidad. Entonces nuestro DAG es*

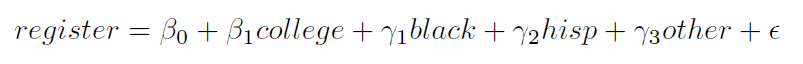
**

*Es decir, ¡tenemos motivos para creer que nuestro estimador de IV arriba, , sigue siendo inconsistente! Es decir, en tanto el instrumento no sólo afecta indirectamente a register por 1 solo canal sino por 2, el efecto de 0.28 sigue estando contaminado por el segundo canal (vía ethnicity).*

*¿Qué hacer en este caso? Debemos incluir este segundo canal en nuestro modelo. En nuestro caso, debemos incluir variables que capturan la etnicidad en el modelo. Afortunadamente, en nuestra base de datos tenemos las variables de control*

* *White*
* *Black*
* *Hispanic*
* *Otherrace*

*En tanto hay 4 categorías, podemos incluir sólo 3 dummies como variables de control. Elegimos omitir White y, por tanto, los coeficientes de las dummies nos dicen cuánto mayor es la probabilidad de que la persona se registre para votar para aquellos que son, por ejemplo, negros en relación a blancos. El modelo que estimamos es entonces*

**

*Usaremos un comando más versátil: ivreg2, con la opción first. En tanto la variable college sigue siendo un regresor endógeno, usamos el mismo instrumento que antes: distance*

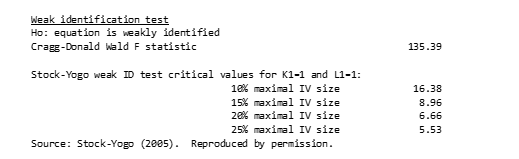


*Centrémonos en el panel inferior, bajo “IV (2SLS) estimation”. El efecto estimado en este caso es de 0.25, y es estadísticamente significativo. Asimismo, los coeficientes de etnicidad, black, Hispanic y otherrace, son estadísticamente significativos. Ciertamente, era necesario cerrar el Segundo canal cuya existencia se sospechaba.*

*Nuevamente, toda estimación creíble por IV pasa por verificar si nuestros instrumentos no son débiles. Vemos bajo el encabezado de “First Stage Regressions” que el test F de instrumentos excluidos es 135.39, el cual es definitivamente mayor que 10. Por tanto, el problema de instrumentos débiles no está presente en este caso. El siguiente test de instrumentos débiles implica usar los valores críticos del estimador de Stock y Yogo (que sólo ivreg2 produce), los cuales sólo aplican cuando nuestro modelo es homoskedástico (como en el presente caso). Estos valores críticos de prueba varían con*

* + *el número de regresores endógenos*
  + *el número de instrumentos excluidos*

*Debemos comprobar que nuestro test F es mayor que el 10% maximal IV size.*

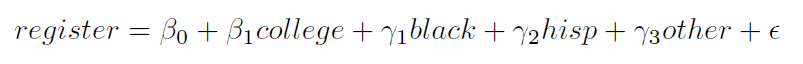


*¡Nuestro instrumento pasa la prueba! Esto provee evidencia de que nuestros instrumentos NO son débiles*

*Nótese cómo en la regresión de la primera etapa los controles, las dummies de raza, han sido incluidos. Esto es algo que siempre debemos recordar: la primera etapa regresiona el regresor endógeno contra todos los instrumentos: los incluidos (las 3 dummies) y el excluido (distance).*

*Vale la pena una reflexión: ¿debemos incluir variables de control sólo si es que hay más de un canal bajo el cual el instrumento afecta a la dependiente? Muchos estudios incluyen variables de control aun cuando no hay otro canal hacia la dependiente generado por el instrumento. ¿Por qué? Para reducir los errores estándar de la estimación. Específicamente, si incluimos en la ecuación variables que son exógenas, reduciremos la varianza residual de nuestro modelo, lo que resulta en errores estándar menores. Por tanto, el uso de variables de control casi siempre ocurre, por una razón u otra.*

1. Queremos ahora examinar cómo el efecto de haber ido a un college sobre el registro para votar difiere según la etnicidad del individuo. ¿Cómo solventaría esto? ¿Es necesario

*Acabamos de estimar el modelo, el cual nos da el efecto de college sobre register descontando el efecto de la etnicidad en la dependiente. Pero ahora queremos desentrañar el efecto de college y ver cómo varía a través de las diferentes etnicidades. Es decir, queremos estimar ahora*

**

*La forma de estimar este modelo implica el uso de interacciones. Es decir, ahora debemos incluir diferentes términos para college, donde cada uno representa su producto con una etnicidad específica. En tanto tenemos 4 etnicidades, sólo podemos incluir 3 interacciones. Siguiendo lo visto en la pregunta anterior, incluimos*

* *college x black*
* *college x hisp*
* *college x other*

*y por tanto, los efectos estimados se interpretan en relación al efecto de college para blancos sobre la dependiente. Asimismo, siempre que usamos interacciones debemos incluir las variables en niveles. Esto significa que el modelo que estimamos es*

**

*Nótese que este modelo difiere de lo visto anteriormente en términos del número de regresores endógenos*

* *en las preguntas anteriores teníamos sólo 1 regresor endógeno, college, y 1 instrumento*
* *pero ahora tenemos 4 regresores endógenos*
  + *college*
  + *college x black*
  + *college x hisp*
  + *college x other*

*¿por qué los 3 últimos son endógenos? Pues porque college sigue siendo endógeno, y por tanto su producto con cada una de las 3 dummies de raza (que son exógenas) resulta en un regresor endógeno.*

*En tanto necesitamos al menos 1 instrumento por cada regresor endógeno, y debido a que tenemos 4 regresores endógenos, necesitamos ahora (al menos) 4 instrumentos. ¿Qué hacer? La respuesta es inmediata: los nuevos instrumentos vienen de la interacción del instrumento, , con cada una de las 3 dummies. Es decir*

* *el regresor endógeno se instrumentaliza por*
* *el regresor endógeno se instrumentaliza por*
* *el regresor endógeno se instrumentaliza por*
* *el regresor endógeno se instrumentaliza por*

*Es decir, tenemos ahora 4 regresores endógenos y 4 instrumentos, lo que permite la estimación del nuevo modelo. Antes de estimar, debemos crear las variables necesarias para ejecutar el comando*

//creación de los regresores endógenos

gen collegexblack = college\*black

gen collegexhispanic = college\*hispanic

gen collegexotherrace = college\*otherrace

//creación de los instrumentos

gen distxblack = distance\*black

gen distxhispanic = distance\*hispanic

gen distxotherrace = distance\*otherrace

*Ahora usamos el mismo comando siguiendo la misma lógica de la sintaxis. Nótese cómo ponemos en paréntesis los 4 regresores endógenos y los 4 instrumentos*

ivreg2 register black hispanic otherrace ///

(college collegexblack collegexhispanic collegexotherrace = ///

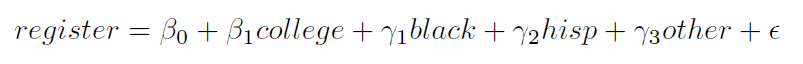
distance distxblack distxhispanic distxotherrace), first

*Veamos primero los resultados de la instrumentalización*



*Los resultados muestran que las interacciones de etnicidad con college no son estadísticamente significativos: ninguno delos coeficientes estimados son estadísticamente diferente de 0. Esto puede corroborarse con un F test*

**

*Esto significa que no es necesario pasar del modelo básico al modelo expandido*

*¿Qué nos dice la primera etapa de la estimación? En tanto tenemos 4 regresores endógenos, es necesario estimar 4 primeras etapas, una para cada regresor endógeno. Nótese cómo, correctamente, cada primera etapa regresiona el regresor endógeno contra todos los instrumentos: los incluidos (las 3 dummies de raza) y el excluido (). En cada una de estas regresiones, vemos que los tests F pasan el valor de 10. Igual que antes: si estamos dispuestos a aceptar como máximo un sesgo de nuestro estimador IV del orden del 10% del sesgo del estimador MCO, nuestros instrumentos sí sirven.*



# Los retornos a la educación

1. A partir de la base de datos “card.dta”, halle el efecto los años de educación (Educ) sobre los salarios en logaritmos (lwage) a través de OLS ¿hay motivos para suponer que el regresor de interés es endógeno?

*La regresión por MCO de logwage sobre educación provee el siguiente output*

*El coeficiente estimado, estadísticamente significativo, nos dice que un año más de educación genera un aumento del 5% en los salarios. Este efecto es estadísticamente significativo al 1%. Pero, por razones que ya han sido mencionadas en la clase, en realidad la relación entre educación y salario es del siguiente tipo*

*A picture containing text, font, line, diagram

Description automatically generated*

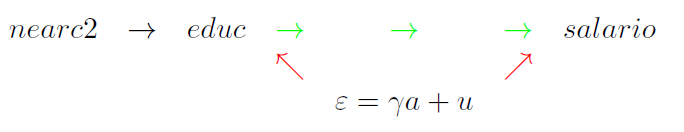
*Como ya sabemos, la habilidad es una variable omitida que está creando un sesgo de variables omitidas. ¿Por qué?*

* *Habilidad está relacionada con educación*
* *Habilidad impacta en el salario*

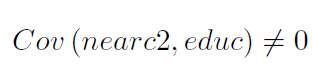
*Esto significa que el estimado de 5% hallado por MCO está contaminado por el efecto que otras variables, en especial habilidad, tiene sobre los salarios. Una solución a estos son el método de variables instrumentales.*

1. Un instrumento posible es la distancia a un 2-year college en 1966 (10 años antes de que se recogieron los datos que componen la encuesta). ¿Es razonable esta variable como instrumento? Dibuje el DAG y estime el modelo correspondiente. ¿Qué está pasando?

*Este instrumento implica que el DAG corresponde a*

**

*Igual que como hemos hecho antes, debemos motivar que nuestro instrumento cumpla con las dos condiciones básicas*

* *: esto significa que el hecho de haber vivido cerca de un college de 2 años está relacionado con el nivel educativo de la persona. Esto es creíble. Sin embargo, puede* ***y debe*** *testearse empíricamente*
* *: esto significa que el hecho de haber vivido cerca de un college de 2 años* ***no*** *está relacionado con otros factores no observados que determinan el salario, en particular la habilidad. Esto* ***no puede*** *testearse empíricamente y sólo nos queda la argumentación para llevar a cabo esta verificación. ¿Es razonable asumir la 0 covarianza en este caso? Pensemos en qué casos fallaría: si aquellos que viven cerca de una escuela tienen una mayor habilidad.* ***Esto no es tan jalado de pelos****: si aquellos padres quienes eligen la localización del hogar cerca de un college de 2 años se interesan más por la educación del hijo, puede que estos hayan recibido mayor estimulación de niños. En este caso, la localización al college NO sería aleatoria, sino que estaría relacionado con la habilidad. Mayor cercanía al college (z), mayor habilidad (a). ¡En este caso el instrumento fallaría! Entonces, asumamos que este mecanismo está ausente. Igual que antes, el factor en el que nos estamos centrando es habilidad, debido a que hemos elegido dicha variable como aquella que está creando un problema de variables omitidas. En caso de que elijamos otra variable, debemos argumentar la correlación con dicha variable.*

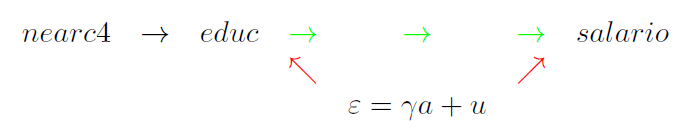
*Asumiendo que esta discusión nos satisface, estimemos el modelo por IV usando ivreg2. El output que se muestra abajo nos dice que ¡cada año adicional de educación genera un incremento del salario en 25%! Este estimado es demasiado grande como para tomarse seriamente. ¿Qué puede estar ocurriendo? Los estadísticos de la primera etapa proveen la respuesta. El test F del instrumento es 6.76, menor al valor de 10. Esto indica que tenemos un problema de instrumentos débiles. Esto se corrobora al ver el valor del estadístico de Stock y Yogo: 6.76 es sólo mayor al 20% maximal IV size. Es decir, con este F test estamos dispuestos a asumir un sesgo del estimador IV del orden del 20% del sesgo del estimador MCO. Nadie estaría dispuesto a asumir tremendo sesgo.*

*Esto ilustra un problema más común de lo que esperaríamos: nuestro instrumento cumple la ER pero la correlación con el regresor endógeno es débil. Por tanto, nuestro instrumento es inservible.*

**

1. Use ahora la distancia a un college de 4 años en 1966 (10 años antes de que se recogieron los datos que componen la encuesta). Dibuje el DAG y estime el modelo correspondiente ¿Qué encuentra ahora en términos de instrumentos débiles?

*Esta vez la estrategia de instrumentación es*

**

*En tanto el instrumento nearc4 es cualitativamente similar al nearc2, las condiciones mediante las cuales , explicadas arriba, se mantienen. Esto significa que la correlación entre el instrumento y el error compuesto podría no ser necesariamente 0, por lo que la violación del supuesto de ortogonalidad no es tan trillada.*

*¿Qué encontramos en términos de debilidad d el instrumento?*

**

*En este caso, un año adicional de educación implica un aumento del salario del 18%. ¡Este efecto mucho más creíble que el efecto de 35% hallado anteriormente! Una explicación de por qué, esta vez, este resultado es más razonable se encuentra al analizar la primera etapa*

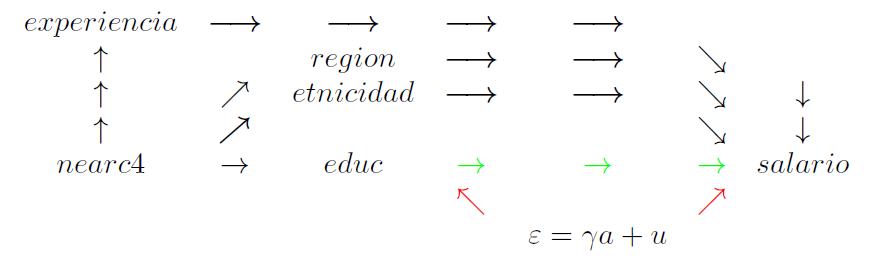
**

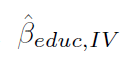
*A diferencia de lo encontrado arriba, esta vez el F test es 64, mayor al límite de 10. Asimismo, es mucho mayor a los valores de Stock y Yogo del 10% maximal IV size. La comparación de este resultado que resulta de usar instrumento no débil (nearc4), con lo que resulta de usar un instrumento débil (nearc2) ilustra que la suma importancia de que el instrumento no sea débil, i.e. que la correlación entre z y x no sólo sea diferente de 0 sino que sea mucho mayor que 0.*

1. Digamos que tenemos motivos para creer que la distancia a un 4 year college está relacionado con diferentes variables
   1. La etnicidad
   2. La localización
   3. La experiencia laboral de la persona

Y cada uno de estos afecta el salario de la persona ¿Qué significa esto para nuestro modelo? Dibuje el nuevo DAG y estime el modelo correspondiente

*En principio, el instrumento sólo afecta indirectamente a la dependiente a través de 1 canal: educación. Pero ahora tenemos motivos para creer que hay más de un otro canal por el cual nearc2 afecta a la dependiente. Entonces nuestro DAG es*

**

*Si esto es cierto, ¡tenemos motivos para creer que nuestro estimador de IV arriba, , sigue siendo inconsistente! Es decir, en tanto el instrumento no sólo afecta indirectamente al salario por 1 solo canal sino por 4, el efecto del 18% sigue estando contaminado. ¿Cuán plausible es este DAG? Ciertamente, la localización del hogar en el pasado, si bien no está relacionada con la habilidad, puede estar relacionada con la etnicidad (poblaciones negras viven más lejos de una escuela) y con la región (poblaciones que viven al sur tienden a vivir más lejos de la escuela). Tanto la etnicidad como la región afectan el salario que la persona obtiene. ¿Y qué hay de la experiencia laboral? Bien podría decirse que ese canal es falso, pues la experiencia depende de otros factores diferentes a la cercanía al colegio. Esto en principio implica que no tenemos que incluir esta variable en el modelo. Pero, como se explicó en la pregunta 1 de la PD, la incluimos para reducir la varianza residual*

*¿Qué hacer en este caso? Ya sabemos que debemos incluir estos canales en nuestro modelo. A diferencia de los visto en la pregunta 1, tenemos diferentes grupos de variables que capturan estos canales alternativos*

* *Black: incluiremos una dummy que captura si la persona es negra o no*
* *South: incluiremos una dummy que captura si la persona vive en el sur o no*
* *Incluiremos la experiencia en el modelo para aumentar la eficiencia. En tanto esta es una variable continua, la incluiremos como cuadrática (Exper y Exper2) usando factor variables: c.exper##c.exper*

*El modelo que estimamos es entonces*

**

*Veamos primero los resultados de la instrumentación. El efecto de educación sigue siendo estadísticamente significativo, pero esta vez cada año adicional*

**

*Veamos primero los resultados de la instrumentación. El efecto de educación sigue siendo estadísticamente significativo, pero esta vez cada año adicional aumenta el salario en 22%.*

*¿Qué ocurre en la primera etapa?*

**

*El estadístico F es de 35.20, mayor al corte de 10. Asimismo, este valor es mayor al 10% maximal IV size de Stock y Yogo, 16.38. Esto nos asegura que, nuevamente, el instrumento sigue siendo no débil. No obstante, el F es menor que antes cuando no habían controles en el modelo. ¿Por qué? Compárese las primeras etapas*

* *Del modelo sin controles*

**

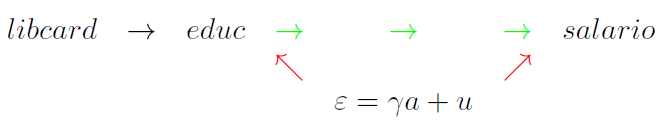
* *Del modelo con controles*

**

*En el modelo con controles la primera etapa también incluye los controles las dummies de black y south así como experiencia como polinomio de grado 2. Esto implica que el efecto de nearc4 sobre educación es menor, pues se está descontando el efecto de estas variables adicionales. Por el contrario, en el modelo sin controles, el efecto de nearc4 es bruto, pues incluye otros posibles canales. Esto reduce la relevancia relativa del instrumento excluido, nearc4, explicando los años de educación del individuo. Esto resulta en que el estadístico F de esta variable caiga de 63.9 a 35.2 en la primera etapa. De todos modos, es mayor a 10, lo que nos da cierta tranquilidad.*

1. En 1966, 10 años antes del año en el que se recogieron los datos, se llevó a cabo un programa **aleatorizado** en el cual se entregaba library cards a los hogares. Esto les permitiría acceder a las bibliotecas públicas y retirar libros para leerlos en casa. ¿Es este un instrumento válido? ¿Qué haría entonces, considerando lo que sabemos de nearc4? Corrija por heteroscedasticidad en la estimación de estos modelos

*¡En este caso podemos usar la tenencia de library card como un instrumento¡*

**

*¿Por qué? Nuevamente, pensemos las condiciones de un buen instrumento*

* *: esto significa que el hecho de tener una library card está relacionado con el nivel educativo de la persona. Esto es creíble, pues esto mejora la habilidad cognitiva del individuo, lo que mejora su nivel educativo. Pero, como ya sabemos* ***debemos*** *testearlo empíricamente*
* *: esto significa que el hecho de tener una library card* ***no*** *está relacionado con otros factores no observados que determinan el salario, en particular la habilidad. Esto* ***no puede*** *testearse empíricamente y sólo nos queda la argumentación para llevar a cabo esta verificación. ¿Es razonable asumir la 0 covarianza en este caso? Pensemos en qué casos fallaría: si aquellos con library card tienen una mayor habilidad. En principio, esto es obvio: aquellos con mayor habilidad cognitiva tienden a leer más, por lo que es probable que tengan su library card. Pero, como se ha enfatizado, la recepción de library cards fue aleatoria. En tanto fue aleatoria, no está relacionado sistemáticamente con ninguna característica de la persona, y esto incluye habilidad. Por tanto, podemos esperar que esta condición se cumpla en este caso en virtud de la aleatorización*

*Esto, junto con lo visto para nearc4, significa que ahora tenemos 2 instrumentos para 1 regresor endógeno (educación). Es decir, ¡tenemos un modelo sobreidentificado!. Tenemos 2 formas de proceder, una mejor que otra.*

*La primera es usar 2 veces el estimador de IV, 1 para cada instrumento. Esto implica ejecutar 2 comandos*

* *ivreg2 lwage black south c.exper##c.exper (educ=nearc4), first robust*
* *ivreg2 lwage black south c.exper##c.exper (educ=libcrd14), first robust*

*Nótese cómo hemos incluido la opción “robust”, que corrige por heteroscedasticidad del error condicional a las variables exógenas.*

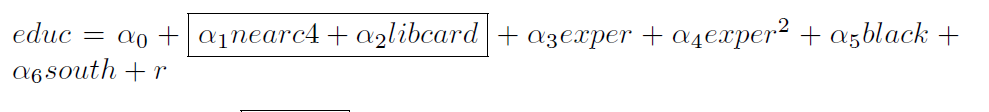
*Ahora obtenemos 2 efectos de educación, 1 para cada instrumento. Si ejecutamos estos comandos obtenemos que*

* *al usar como instrumento near4, el efecto de educación sobre salarios es 22% (que vimos anteriormente). El estadístico F relevante, en caso de corregir por heterosk., es el de Kleibergen-Paap Wald rk F statistic (36.96). El test de Cragg-Donald Wald F statistic ni los valores críticos de Stock y Yogo son válidos bajo heterosk.*
* *al usar como instrumento libcrd14, el efecto de educación sobre salarios es 13%. El estadístico F relevante es el de Kleibergen-Paap Wald rk F statistic (115.66).*

*En tanto ambos instrumentos son válidos y no sufren del problema de instrumentos débiles, ambos resultados si bien son diferentes (22% y 13%),, pueden considerarse causales. Pero hay un problema con este enfoque: el 1er estimado se calcula dejando de lado la información del 2do instrumento; y el 2do estimado se calcula dejando de lado la información del 1er instrumento.*

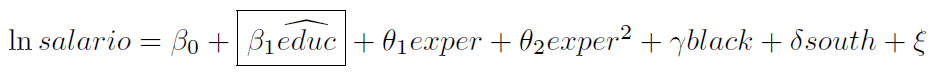
*La segunda forma de proceder solventa esto: usa ambos instrumentos a la vez. Esto implica usar el estimador en 2 etapas (Nótese que el comando ivreg2 y el ivregress usa siempre el estimador en 2 etapas aun con modelos exactamente identificados a pesar de que no es necesario en este caso. Sin embargo, hacer esto permite verificar si el instrumento es no débil. En el caso sobreidentificado, es necesario usar el estimador en 2 etapas). Lo que hacemos en este caso es*

1. *Estimar la primera etapa donde regresionamos el regresor endógeno contra los instrumentos*

**

*Y a partir de obtenemos la predicción de educación. Nótese cómo, nuevamente, incluimos las variables de control (instrumentos incluidos)*

1. *Estimar la ecuación de interés donde reemplazamos el regresor endógeno por su predicción de la 1era etapa*

**

*En este caso*

**

*Lo que nos dice que el efecto causal de educación, en promedio, sobre salarios es 15% con cada año de educación que la persona obtiene adicionalmente. En este caso el diagnóstico de la primera etapa es satisfactorio: los dos instrumentos tienen los signos esperados en la primera etapa (positivos) y el estadístico F Kleibergen-Paap Wald rk , 68.17, es mayor a 10. Nuevamente, en tanto los errores no son i.i.d. sino heterosk., los valores de Stock Yogo y el F stat. Cragg Donald no aplican.*

*Todo esto nos permite afirmar que el estimar el efecto de educación usando 2 instrumentos no débiles provee resultados satisfactorios, en términos de credibilidad del estimado y de la ausencia del problema de instrumentos débiles.*

**

1. Estime el mismo modelo, corrigiendo por heteroscedasticidad, pero usando el Método Generalizado de Momentos (GMM)

*El estimador de 2SLS es el estimador más eficiente si es que nuestro modelo se caracteriza por homoscedasticidad. Es decir, no hay otro estimador lineal consistente bajo endogeneidad que tenga menor varianza. En cambio, el estimador MCO es el estimador lineal consistente bajo exogeneidad con mayor varianza.*

*Pero en la mayoría de análisis aplicados la varianza condicional del error no es constante, es decir no tenemos que*

*Esto es algo que afrontamos en la pregunta anterior. Más bien, la varianza de los no observables (e.g. habilidad en el presente caso) error cambia a través de diferentes valores de las variables en el vector z. En este caso debemos estimar la matriz de varianzas y covarianzas robustas a heterosk, pues las matrices asumiendo homoskedasticidad son válidas.*

*La ruta que toma el estimador 2SLS es estimar los parámetros del modelo y luego corregir sólo los SEs por heterosk. Pero heterosk., aun cuando es una violación del supuesto de homosk., es información que podemos incorporar directamente en la estimación. Esto es precisamente lo que el método GMM hace: incorporar directamente en la estimación de los parámetros la información que provee la heterosk. o la clusterización de los errores, i.e. incluir lo más de información que se pueda en la estimación de parámetros, y no sólo incluir esta información para corregir los SEs.*

*Nótese que el método GMM es un método (¡no hay tal cosa como el “modelo” GMM!) y como tal permite estimar modelos*

* *lineales (como el que vemos acá, A picture containing font, text, white, typography

  Description automatically generated) con y sin endogeneidad*
* *no lineales (e.g. modelo probit , logit, Poisson, Tobit, etc.), con y sin endogeneidad.*

*El estimador que usaremos es el de IV lineal*

*A riesgo de sobresimplificar, podemos explicar el método GMM de la siguiente forma. Cuando estimamos un modelo por MCO, , sabemos que debe cumplirse la condición de consistencia*

*, que a su vez implica que . Si seguimos operando, obtenemos la función poblacional del estimador MCO, A black text on a white background

Description automatically generated with low confidence. Y una vez que obtenemos esta expresión, usamos una muestra y por el principio de analogía obtenemos el estimador muestral MCO: A picture containing font, text, white, handwriting

Description automatically generated. Esto no es nada nuevo, ya se ha visto en clase*

*Pero si 1 de los regresores es endógeno y si tenemos 1 instrumento (o 2 regresores endógenos y 2 instrumentos, etc.), entonces por la ER, sabemos que A picture containing font, graphics, white, text

Description automatically generated donde, nuevamente, z es el vector de todas las exógenas. Esto a su vez implica que A picture containing font, text, line, handwriting

Description automatically generated . Si seguimos operando, obtenemos la función de poblacional del estimador IV A picture containing font, text, typography, white

Description automatically generated. Y una vez que obtenemos esta expresión, usamos una muestra y por el principio de analogía obtenemos el estimador muestral de IV A picture containing font, white, text, typography

Description automatically generated. Esto tampoco es nuevo, ya se ha visto en la clase.*

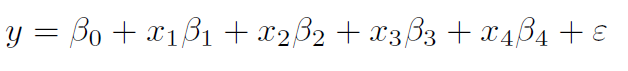
*Nótese cómo en estos ejemplos hemos partido de condiciones (i.e. de la restricción de que debe ser igual a 0) sobre la expectativa incondicional del*

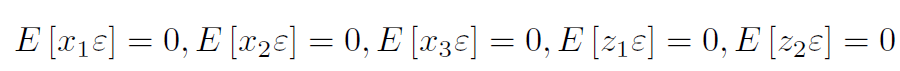
* *producto del error y los x para el estimador MCO*
* *producto del error y los z A picture containing text, watch

  Description automatically generated para el estimador IV*

*Estas condiciones sobre las expectativas se conocen como condiciones de momentos (en particular, condiciones respecto al primer momento, la expectativa). Estas condiciones son las que el método GMM explota. Así, hemos comenzado con condiciones de momentos y hemos terminado con los mismos estimadores que hemos visto antes en la clase. ¿Cómo así? Pues en los modelos como los que acabamos de ver, que son exactamente identificados, el método GMM nos provee los mismos estimadores que los hallados de forma “clásica”. Nada nuevo o interesante hasta ahora por parte del método GMM.*

*Pero como se dijo, la ganancia del método GMM está en los modelos sobreidentificados. En este caso tenemos más ecuaciones a resolver que incógnitas. E.g. si tenemos un modelo con 3 controles, x1, x2 y x3, y 1 regresor endógeno x4,*

**

*pero con M=2 (z1,z2), tenemos 5 condiciones de momentos* 

*Sin embargo, sólo necesitamos identificar 4 coeficientes: Beta1, Beta2, Beta 3 y Beta 4. Resulta que en estos modelos sobreidentificados no existe una solución única para los coeficientes (es el mismo problema que les enseñaron en el colegio: si tenemos 3 ecuaciones y 2 variables por hallar, no hay una única solución de estas 2 variables)*

*Para estos casos sobreidentificados, definamos la condición de momentos en términos de expectativas, vistas anteriormente, más generalmente como*

*A picture containing font, typography, text, handwriting

Description automatically generated*

*y sabemos que, entonces, podemos estimarlos usando una muestra*

*A picture containing font, text, typography, white

Description automatically generated*

*Lo que hace el método de GMM para solucionar el problema de la falta de una solución única para los coeficientes en casos sobreidentificados se puede expresar de la siguiente forma: “debido a que sé que no es posible que las condiciones de momentos en modelos sobreidentificados sean todas 0 (como sí ocurre en modelos exact. identificados), entonces vamos a hacerlas lo más cercanas posibles a 0”. Entonces, lo que hace el método GMM es minimizar las condiciones muestrales de momentos al cuadrado, ponderando cada condición de momento por una matriz W (que se asume positiva definida):Text

Description automatically generated with low confidence. Esta expresión puede pensarse como la minimización de A picture containing font, symbol, white, graphics

Description automatically generated ponderando cada término de esta suma de cuadrados por un elemento de la matriz W. Esta matriz W es clave y aquella que provee la menor varianza posible del estimador, la matriz W óptima, es la inversa de la matriz de varianza de los momentos:* A picture containing text, clock

Description automatically generated*. La matriz W, elegida de esta forma, hace que el método GMM alcance la varianza mínima, i.e. que provea el estimador más eficiente que el método 2SLS.*

*Esta matriz W es la que le da un gran poder al método GMM, pues incorpora información que otros estimadores no incluyen: información sobre el error. Diferentes matrices W tienen diferentes formas, reflejando diferentes características de la varianza del error (homosk., heterosk., errores clusterizados) que forma parte de esa matriz vía el término e\_i x e\_i’, que se puede ver más explícitamente en la expresión de la derecha.*

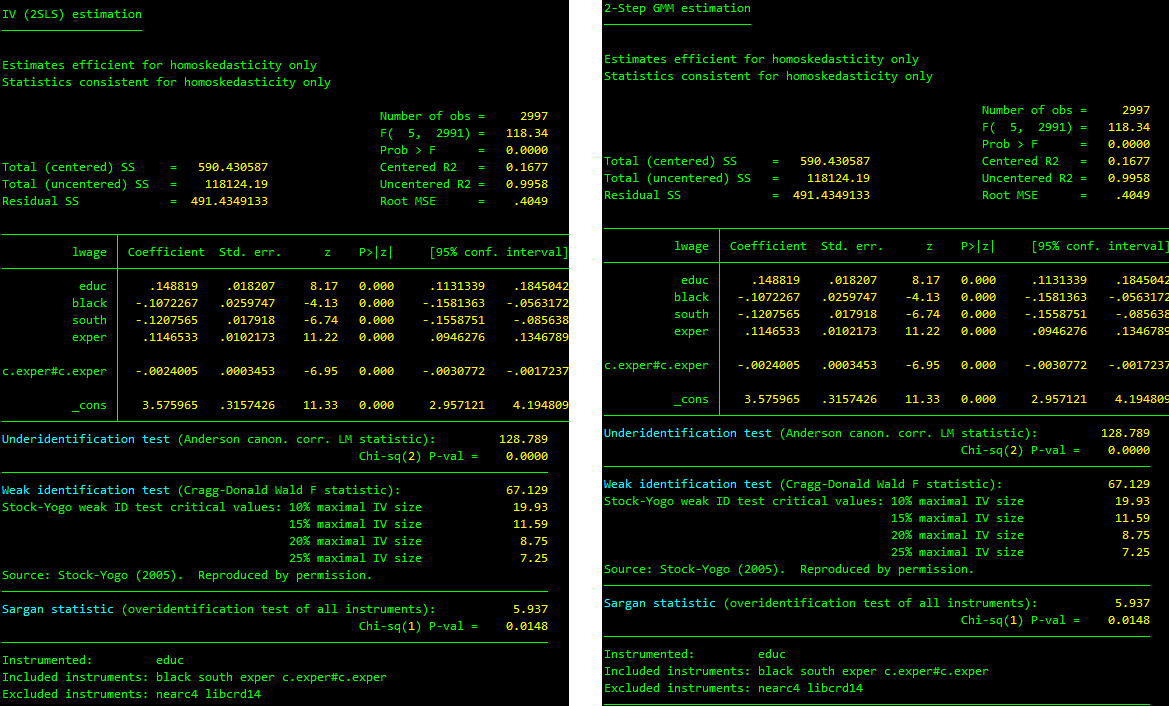
* *Bajo homosk.,*
* *Bajo heterosk,*
* *Bajo clusterización la matriz W toma otra forma.*

*A su vez, estas diferentes matrices W llevan a diferentes funciones a minimizar, lo que resulta en diferentes estimadores consistentes y diferentes matrices de varianzas de los parámetros. ¿Por qué la fórmula de la matriz W óptima es como es? La idea es que a las condiciones de momentos más ruidosas, con mayor varianza, se les da menos peso en la estimación; i.e. se da más relevancia a las condiciones de momentos con menos ruido.*

*Entonces, la única ganancia real que se obtiene usando este método, en el contexto de IV, es en el caso sobreidentificado y bajo desviaciones de homosk. (heterosk. o clustering de los errores). De hecho, podemos comprobar que GMM y 2SLS serán iguales para un modelo sobreidentificado bajo homosk.*

* 2SLS: lwage black south c.exper##c.exper (educ=nearc4 libcrd14)
* IV GMM: lwage black south c.exper##c.exper (educ=nearc4 libcrd14), **gmm2s**

Como se puede verificar, los resultados son exactamente iguales en términos de los coeficientes, SEs, y estadísticos auxiliares



Sin embargo, en el presente contexto, estamos estimando los modelos asumiendo heterosk. y por tanto el estimador GMM debería proveer resultados con menores SEs.



El coeficiente de educación es 0.147, muy similar al 0.149 bajo 2SLS. Nótese que el error estándar del coeficiente estimado de educación bajo GMM es ligeramente menor, 0.01820, que el que se obtiene usando 2SLS, 0.01821. En este caso las diferencias son muy pequeñas y usar GMM o 2SLS en este modelo sobreidentificado es indiferente. Pero en muchos casos las diferencias son mayores, tanto en coeficientes como en errores estándar. En este caso es razonable preferir los resultados bajo GMM. Pero recuérdese que siempre estamos asumiendo que ambos instrumentos son válidos, en términos de la condición de ortogonalidad, de relevancia, y que no son débiles.

# Test de exogeneidad y sobreidentificación

1. Aplique el test de exogeneidad de Durbin-Hausman al modelo estimado
   1. en la parte 1.iii, donde instrumentalizamos college por distance e incluimos controles de raza
   2. en la parte 2.v, donde instrumentalizamos education por nearc4 y libcrd14 e incluimos controles por raza, localización y experiencia

En ambos casos, estime el modelo corrigiendo por heterocedasticidad. ¿Qué concluimos: valió la pena instrumentalizar?

*Cuando estimamos un modelo con un regresor endógeno usando un instrumento válido corregimos el problema de inconsistencia. El estimador de IV o 2SLS resultante es consistente. En cambio, cuando estimamos el mismo modelo (con regresor endógeno) por MCO, el estimador que obtenemos es inconsistente. Pero, como mencionó en la clase, el IV o 2SLS implica problemas:*

* *debemos asegurarnos que el instrumento es válido y que está fuertemente correlacionado con el error (i.e. que no sea débil)*
* *el estimador IV sólo es válido con muestras grandes*

*Si estos 2 condiciones son ciertas, el estimador de IV o 2SLS fiable. Pero de todos modos, aun en este caso satisfactorio en términos de consistencia, el estimador de IV se caracteriza siempre por una mayor varianza que el estimador MCO. Es decir, siempre estimará el efecto causal con una mayor imprecisión. Por el contrario, el estimador MCO, aunque inconsistente en este caso, es el más eficiente entre todos los estimadores lineales, incluido IV y 2SLS, y por tanto estima el efecto, aunque inconsistente con una menor imprecisión.*

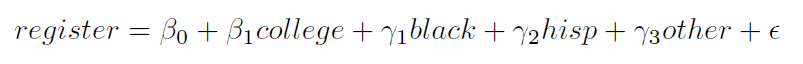
*Uno puede creer que si tenemos endogeneidad siempre tiene sentido tomar el estimador IV y podemos tirar a la basura el estimador MCO. Pero ¿qué pasaría si el coeficiente que obtenemos usando IV se parece mucho al coeficiente que obtenemos por MCO? En ese caso la inconsistencia generada por el sesgo de variables omitida es pequeña. Entonces en ese caso tiene sentido usar el estimador MCO pues, como se dijo, este estima un efecto (que si bien no es exactamente igual al verdadero efecto causal de interés que nos da IV, no es MUY diferente). Esta es la idea el test de Hausman Durbin Wu: compara el estimado IV con el estimado MCO y verifica si la diferencia penalizada entre ambos (penalizada, i.e. dividida, por la varianza de la diferencia entre los coeficientes estimados) es “grande”.*

* *Si la diferencia es grande, entonces la inconsistencia generada por omisión de una variable es grande, y por tanto debemos usar el estimador IV , aun cuando este estima el efecto causal con mayor varianza*
* *Si la diferencia no es grande, entonces la inconsistencia generada por omisión de una variable es casi 0, y por tanto preferimos al estimador OLS pues, si bien este es inconsistente, la inconsistencia es muy pequeña, y además este estimador estimar el efecto causal con mayor precisión (menor varianza del estimado)*

*Más formalmente, la diferencia entre ambos coeficentes se distribuye como una variable Chi cuadrado con 1 grado de libertad.*

* *H0: Variables son exógenas (i.e. la inconsistencia es pequeña y preferimos el estimador MCO pues su contaminación no es muy grande pero es más eficiente)*
* *Ha: Variables no son endógenas (i.e. la inconsistencia no es pequeña y preferimos el estimador IV a pesar de ser menos eficiente)*

*Para entender cómo generar este test usaremos dos modelos estimados anteriormente. El primero es*

**

*Comparemos el estimado de MCO con el de IV, corrigiendo por heterocedasticidad.*

**

*Como se ve, el coeficiente de college*

* *estimado por MCO es 0.18 y su SE es 0.009*
* *estimado por IV es 0.25 y su SE es 0.082*

*El SE del estimado IV es 9 veces más grande que el de MCO. La forma de pensar el test es la siguiente: ¿la diferencia de puntos porcentuales entre ambos estimadores es lo suficientemente grande como para que elijamos definitivamente el estimador de IV? El test se calcula en Stata como*

*ivregress 2sls register black hispanic otherrace (college=distance), vce(robust)*

*estat endogenous*

**

*El resultado se muestra en la parte inferior: el p-value del test (Robust regression F) es 0.39, lo que implica que a cualquier nivel de significancia no rechazamos la H0 que la diferencia entre ambos es pequeña. Es decir, la inconsistencia (penalizada) no es tan grande, y podemos confiar en lo que reporta el estimador MCO. Entonces en caso la instrumentación no era necesaria. Para que quede claro, antes de proseguir, es necesario hacer 2 aclaraciones*

* *este resultado es fiable porque la instrumentación es fiable. Si el instrumento fuera no fiable, no valido, o débil, este resultado no es fiable. Hemos visto en la parte 1 que el instrumento distance, en esta aplicación, es válido, y por tanto los resultados de este test HDW es válido*
* *para llegar a la conclusión que podríamos confiar en el estimador MCO hemos tenido que realizar la instrumentación y el test. No podemos afirmar esto ex-ante, tenemos que seguir todos los pasos mencionados hasta ahora para llegar a esta conclusión*

*¿Qué hay del segundo modelo  instrumentalizado por nearc4 y libcard? Comparemos los resultados bajo MCO y 2SLS corrigiendo por heterosk.*

**

*Como se ve, el coeficiente de educación*

* *estimado por MCO es 0.078 y su SE es 0.003*
* *estimado por IV es 0.149 y su SE es 0.018*

*El SE del estimado IV es 6 veces más grande que el de MCO. La forma de pensar el test es la siguiente: ¿la diferencia de puntos porcentuales entre ambos estimadores es lo suficientemente grande como para que elijamos definitivamente el estimador de IV? El test se calcula en Stata como*

*ivregress 2sls lwage black south c.exper##c.exper (educ=nearc4 libcrd14), vce(robust)*

*estat endogenous*

**

*El resultado se muestra en la parte inferior: el p-value del test (Robust regression F) es 0.00, lo que implica que a cualquier nivel de significancia rechazamos la H0 que la diferencia entre ambos es pequeña. Es decir, la inconsistencia (penalizada) es grande, y sólo podemos confiar en lo que reporta el estimador IV. Entonces, a diferencia del caso anterior, la instrumentación no era necesaria.*

1. Aplique el test de sobreidentificación de Hansen para el modelo estimado en la parte 2.v, donde instrumentalizamos education por nearc4 y libcrd14 e incluimos controles por raza, localización y experiencia. Hágalo para la estimación bajo 2SLS y GMM. ¿Son estos instrumentos válidos?

*Como se mencionó en la clase, NUNCA podemos probar la validez de los instrumentos. Esto es importante de recordar porque muchos autores/docentes/papers dicen, erróneamente, que “si bien en un modelo exactamente identificado no podemos probar la validez de los instrumentos, en un modelo sobreidentificado sí podemos probarla”. Esto no es del todo correcto. Descompongamos esta frase*

* *“si bien en un modelo exactamente identificado no podemos probar la validez de los instrumentos”. Esto es cierto, no se puede probar nada en un modelo exactamente identificado*
* *“en un modelo sobreidentificado sí podemos probarla”. Esto es parcialmente cierto: en un modelo sobreidentificado no podemos probar la validez de los instrumentos, sólo podemos probar si los instrumentos* ***adicionales*** *son válidos (en el sentido de que no están correlacionados con el error).*

*Entonces, un modelo sobreidentificado permite verificar si, dado un instrumento inicial válido, los restantes adicionales son válidos. Para ver cómo podemos aplicarlo en nuestro caso, usemos un modelo sobreidentificado donde sabemos que los 2 instrumentos son válidos*

**

*En este caso el grado de sobreidentificación es 1: 2 instrumentos excluidos (nearc4 y libcard) y 1 regresor endógeno (educ). En este caso podemos ver si es que, dado que uno de los instrumentos es con certeza válido (libcard o nearc4) el otro (nearc4 o libcard, respectivamente) también es válido.*

*El test de sobreidentificación depende de si usamos el estimador de 2SLS o GMM y de si asumimos homosk. u heterosk. Si usamos 2SLS y asumiendo homosk. el test es el test de Sargan y si asumimos heterosk. el test es el Score test de Wooldridge. Si usamos GMM el test siempre será el test J de Hansen. En todos el estadístico se distribuye de la misma forma (chi cuadrado con g.l. igual al grado de sobreidentificacion) y la idea es la misma*

* *H0: Los instrumentos adicionales son válidos*
* *Ha: Los instrumentos adicionales no son válidos*

*Nótese que si tenemos 2 instrumentos cuya validez está en entredicho, y rechazamos la H0, ¿cuál de los 2 instrumentos es realmente inválido? Dicha pregunta es imposible de responder. Por tanto, para que este test sirva de algo, es necesario siempre tener un instrumento adicional válido; lo que hacemos es testear si otro instrumento, que permite la sobreidentificación del modelo, es válido.*

*En tanto el test de Wooldridge y J de Hansen son equivalentes, el resultado del test es el mismo en nuestro caso sea que estimamos el modelo por 2SLS y GMM, pues corregimos por heterosk. Veamos los resultados de GMM*

**

*En este caso, el test nos dice que rechazamos la hipótesis nula, y por tanto ¡NO tenemos indicios para creer que los instrumentos sean válidos! Es decir, el test nos está diciendo que*

* *Si nuestro instrumento inicial válido es nearc4, el instrumento adicional libcard es inválido; o*
* *Si nuestro instrumento inicial válido es libcard, el instrumento adicional nearc4 es inválido*

*Nosotros no podemos en base a los datos, saber cuál de los casos está ocurriendo/ Esto lleva a plantearse teóricamente cuál de nuestros instrumentos está fallando. En tanto el instrumento de libcard se asignó aleatoriamente, es más probable que este sea válido. Por el contrario, desde un momento primer sospechábamos que nearc4 no es válido (ver parte 2). ¿Hay alguna forma de probar que, efectivamente, libcard es valido y que nearc4 no lo es? ¡No! Eso requeriría que podemos probar la validez de un instrumento individual, lo cual es imposible*

*Es importante repetirlo una vez más: el test de sobreidentificación asume que tenemos 1 instrumento válido y queremos ver la validez de otro instrumento adicional. Si, por ejemplo, usamos como instrumentos en vez educación de la madre y del padre, bien podría ser que no rechacemos la H0 de que la correlación entre los instrumentos y el error es 0. De hecho, ¡¡¡esto ocurre en nuestro ejemplo!!!*

**

*El test de Hansen arroja un p-value mayor a cualquier nivel de significancia, lo que implica, en base al test, que los instrumentos son válidos ¡Pero no hay manera alguna en el universo de defender que esos instrumentos son válidos! Tanto motheduc como fatheduc son instrumento inválidos, y no importa cuál es el inicial y cuál el adicional. En ese caso el resultado del test es inválido porque sabemos por lo discutido en clase que ambos instrumentos son monumentalmente inválidos. Por ello, para que el test de sobreidentificación sea válido, no basta con ver el p-value: debemos asegurarnos que al menos 1 instrumento sea válido.*

Dr. Juan Manuel del Pozo Segura

Junio 2023