**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.**

Se tiene data sobre los individuos en dos momentos: 1966 y 1976. En 1976, los individuos evaluados (n=3010), en promedio, tienen una edad de 28 años, 13 años de educación, con padres y madres con 10 años de educación, con un IQ de 102, con experiencia laboral de 8 años y con un sueldo por hora de 577 centavos. Asimismo, 23% son de raza negra, 71% son casados, 9% están inscritos en la escuela, 78% vivían con sus padres a los 14 años. En cuanto a la distancia con la universidad en 1966, 68% vivió cerca de su universidad durante los 4 años

**¿Qué información nos provee el promedio de la dependiente en la interpretación de los coeficientes?¿Cómo se interpretan los coeficientes en este caso?**

La dependiente "lwage" es el logaritmo del pago por hora en centavos.

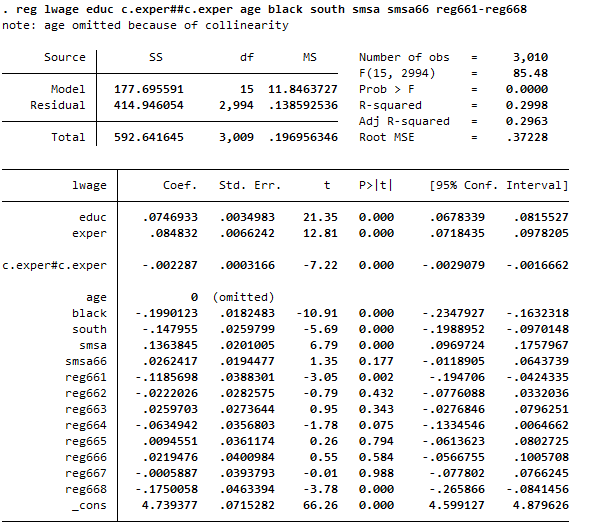
La media de wage es 577, lo que implica que en promedio los evaluados perciben 577 centavos por hora. Sin embargo, para efectos del modelo, se emplea una transformación de esta variable (lwage: logaritmo de wage). La variación del promedio de lwage indicará el efecto de la variable independiente sobre el salario por hora.

1. **Regresione usando OLS el siguiente modelo en Stata**

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

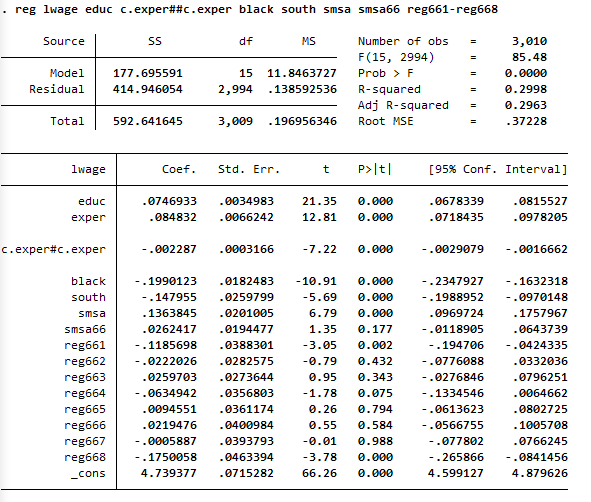
**¿Qué ocurre con edad? ¿Por qué?**

La variable edad aparece omitida, debido a la presencia de multicolinealidad con otras variables. Una de ellas puede ser "exper", ya que es calculada a partir de la edad



**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**

Se identificó, en cuanto a la variable educación, el mismo puntaje que el presentado en la tabla 2 columna 2; cabe recordar que la tabla 2 muestra diferentes puntajes ya que se realizan varias estimaciones. El coeficiente obtenido fue de 0.075, con un p valor menor a 0.05, lo cual indica una influencia positiva significativa de la variable educación sobre la variable dependiente (lwage).

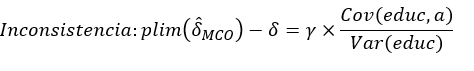
**

1. Vea la Tabla 2 de Card (1993). **¿Qué se encuentra al inspeccionar el coeficiente estimado de educación? ¿Es creíble este coeficiente? ¿Por qué?**

Se observa que la variable Educación tiene un coeficiente positivo mayor a 0.07 (un año de educación representa un incremento del 7% en los ingresos). Este coeficiente se mantiene en la misma magnitud a través de los diferentes modelos que propone el autor (en el cual añade controles progresivamente).

En cuanto a la credibilidad del coeficiente, no es creíble este coeficiente en la medida que la variable Educación **NO** es aleatoria. Depende mucho de la elección que tomen las personas y las variables que subyacen en dicha elección, normalmente hay variables no observables como la habilidad, motivación, educación del padre, etc. Dado que el coeficiente de Educación absorbe el efecto de variables subyacentes, se entiende que existe un problema de endogeneidad que puede sobreestimar el efecto de la variable.

**¿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 inconsistencia) para responder y mencione la dirección esperada del sesgo.

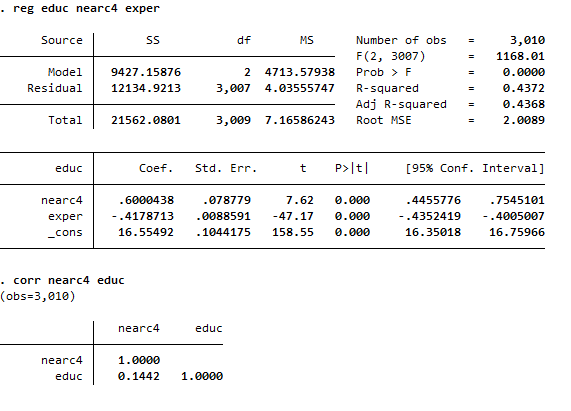


La inconsistencia puede generarse a partir de:

* Efecto parcial de “a” sobre *lwage ()*
* Efecto marginal de “educ” sobre “a”

Considerando que “a” podría ser habilidad o educación de la familia, es muy probable que la covarianza entre “educ” y “a” sea diferente de 0. No solo ello, sino que, además, podría decirse que esta variable “a” tiene un efecto sobre “lwage”, el cual no se puede observar a simple vista. Por lo que el sesgo es muy probablemente positivo, es decir, se sobreestima el efecto de “educ”.

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?

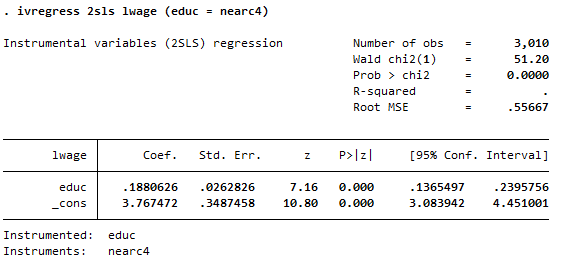


Debemos ver que la variable *nearc4* tiene un coeficiente positivo estadísticamente significativo, lo que implicaría que existe una relación causal positiva entre la variable *nearc4* y la variable *educ*. Respecto a la correlación, es relativamente alta y positiva, 0.14.

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.**



El DAG muestra la relación causal entre ***nearc4*** (instrumento), ***educ***(variable endógena) y ***Lwage***( variable dependiente). Mientras *nearc4* influye en *educ*, *educ* afecta a *lwage*

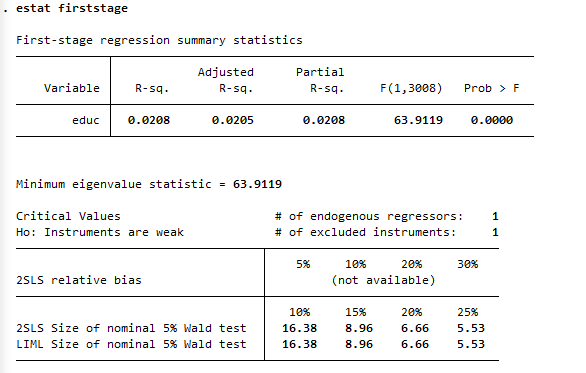


**Comente sus resultados y compárelos con los obtenidos en el punto 2.**

El coeficiente estimado de la variable educ es de 0.188 (es un valor mayor al obtenido en el punto 2) que indica que en promedio un aumento de una unidad del nivel de *educ* está asociado con el aumento de 0.188 en el logaritmo de los salarios por hora y además tiene un valor “p” que indica que estadísticamente significativo, después de controlar la endogeneidad con el instrumento *nearc4.*

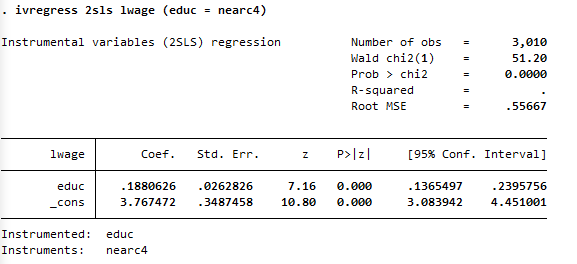
1. Testee la presencia de instrumentos débiles usando i) el estadístico F; ii) el estadístico de Cragg y Donald (1993); iii) las tablas de Stock y Yogo (2005) con respecto a la medida del test de Wald
   1. **el estadístico F**

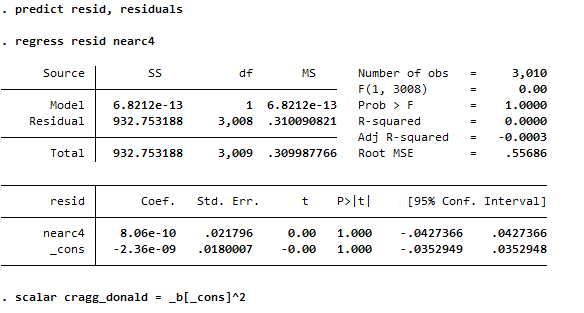
El valor de R2 ajustado para educ es de 2.05%, que implica la explicabilidad de la variable educ en 2.05% por el instrumento. Además el estadístico F es menor a 0,0001 que indica la significancia estadística del instrumento sobre la variable educ. Finalmente el valor estadístico es de 63.9 que indica la presencia de un instrumento fuerte.



* 1. **el estadístico de Cragg y Donald (1993)**

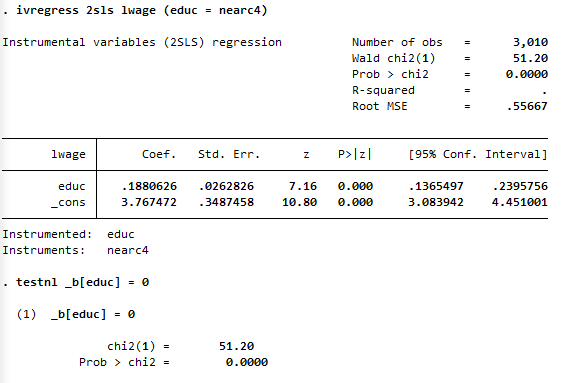
el valor del estadístico de Cragg y Donald es prácticamente cero o indefinido de acuerdo a los valores resultantes, por lo tanto no se puede aseverar la significancia del instrumento.





* 1. **las tablas de Stock y Yogo (2005) con respecto a la medida del test de Wald**

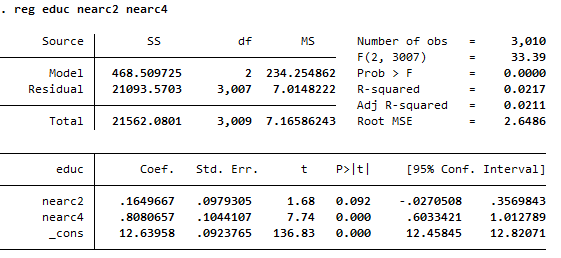
Se coloco la hipótesis nula que el coeficiente de la variable -educ- es 0 y los resultaron fueron ; un valor de p extremadamente pequeño (p < 0.001), hay evidencia estadística suficiente para rechazar la hipótesis nula de que el coeficiente de la variable "educ" sea igual a cero. Esto indica que la variable "educ" es estadísticamente significativa en el modelo y tiene un efecto diferente de cero en la variable dependiente "lwage".



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.

1. **Hacemos la regresión de la variable 'edu' con las variables 'nearc2' y 'nearc4' respectivamente.**

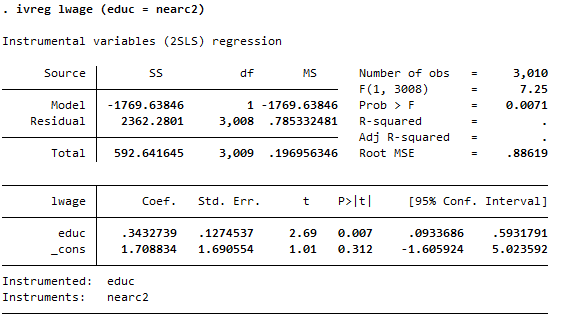
Al hacer la regresión de la variable ‘educ’ con ambas variables, observamos que está más fuertemente relacionada con ‘nearc4’ (0.80) que con ‘nearc2’ (0.16). Esto significa que la cercanía a un college de 04 años tiene una mayor influencia en los años de escolaridad.

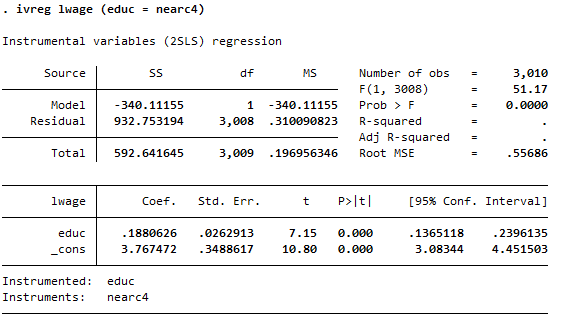


1. **Luego utilizamos el estimador IV por separado.**

Al hacer la regresión con el estimador IV obtenemos para ‘nearc2’ un valor de 0.34 y para ‘nearc4’ uno de 0.18.

La variable ‘nearc4’ al tener una relación más directa con la variable ‘educ’ ofrece mayot posibilidad de controlar el problema de endogeneidad por lo que un coeficiente inferior puede interpretarse como un resultado menos contaminado y nos presentaría la magnitud más real entre la influencia de los años de escolaridad en el salario por horas.

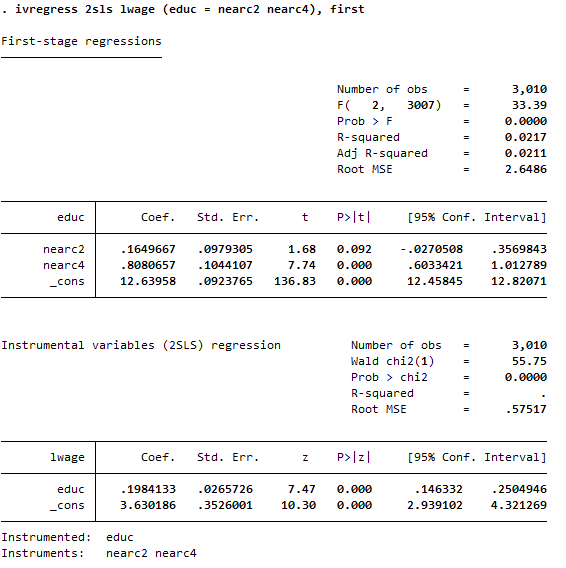
****

****

1. **Ahora usamos el estimador IV 2SLS para ambos instrumentos.**

Al hacer la regresión con el estimados 2sls con ambos instrumentos simultáneamente obtenemos un coeficiente de 0.1984, un valor muy ligeramente por encima del coeficiente cuando la regresión se hizo solamente con la variable ‘nerac4’.

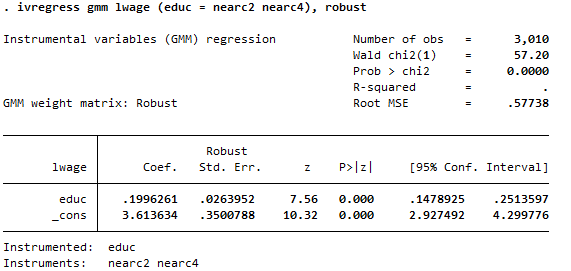
Esto puede interpretarse como una reafirmación del potencial de instrumentalización de la variable ‘nearc4’ para controlar el problema de endogeneidad de ‘educ’ ya que el coeficiente describe una magnitud similar para la influencia de los años de escolaridad en el salario por horas.



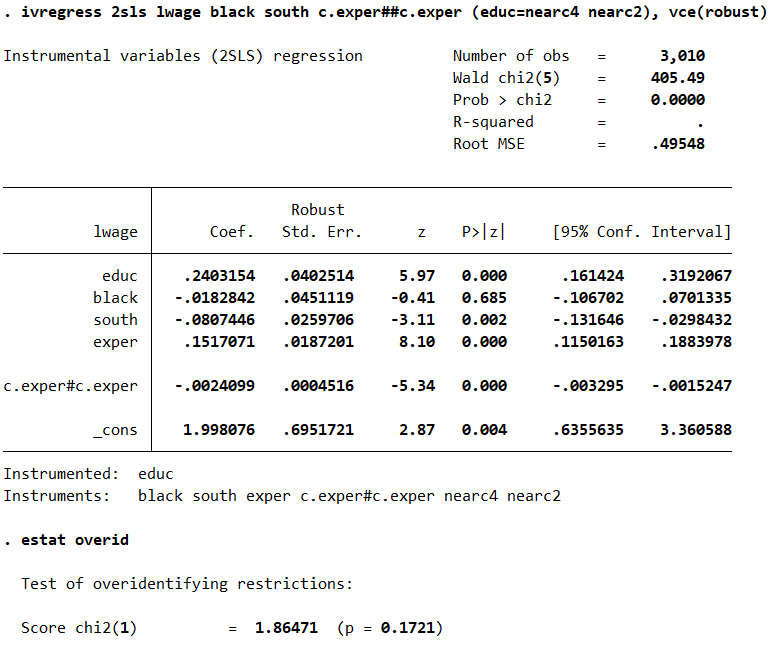
1. Hemos asumido hasta ahora homoscedasticidad ¿Es razonable esto?¿Qué permite el estimador de GMM que no permite el de IV en 2 etapas bajo heterosk. Compare los resultados (del coeficiente educación) del estimador 2SLS con el del estimador GMM bajo heterocedasticidad

1. **Usamos el estimador GMM para ambos instrumentos.**

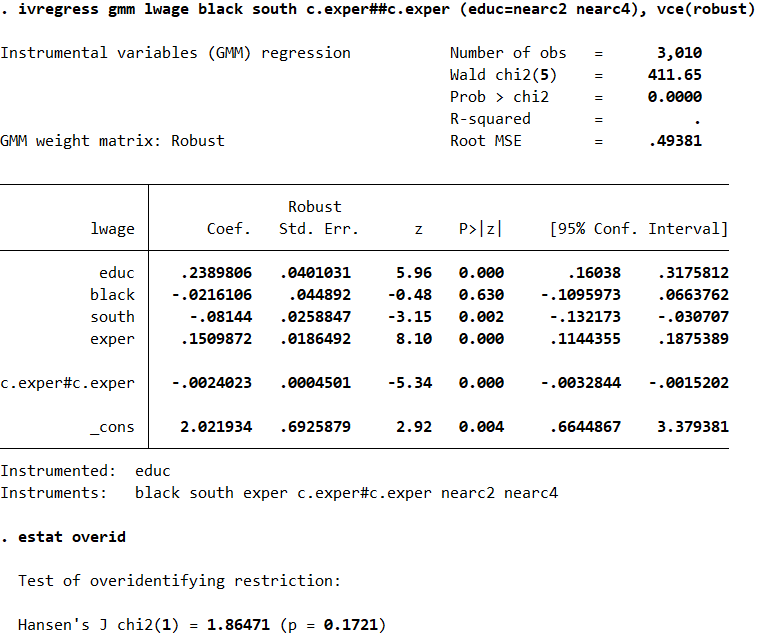
Al hacer la regresión con el estimador GMM (“relajamos” la suposición de homoscedastividad) y obtenemos un coeficiente de 0.1996, un valor casi insignificantemente diferente del coeficiente usando el estimador 2sls de 0.1984. Esto sugiere que la heteroscedasticidad no tendría un impacto sustancial en la estimación, lo que podemos interpretar como que nuestro modelo y las variables instrumentales seleccionadas estarían capturando adecuadamente el impacto causal de la variable explicativa (‘educ’) en la variable dependiente (‘lwage’)



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)
   1. A continuación se observa el test de sobreidentificación con 2SLS

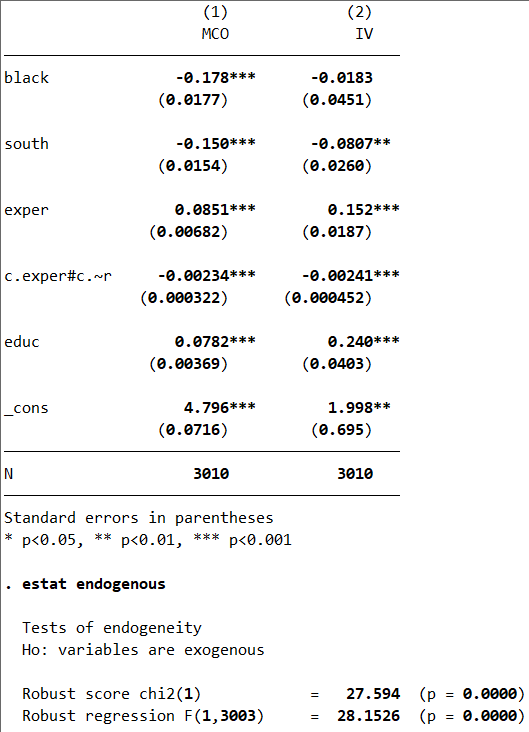


Y también se observa los resultados con el estimador GMM:



A partir de los resultados, se puede inferir que no podemos rechazar la hipótesis nula, es decir, ambos instrumentos son válidos para nuestro modelo.

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?
   1. A continuación se observa la tabla para evaluar la exogeneidad del modelo estimado.



A partir de los resultados obtenidos, se observa que el coeficiente de educ estimado por MCO es 0.078 y SE es 0.003, mientras que el coeficiente de educ estimado por IV es 0.240 y SE es 0.040. Si observamos el resultados del test de endogeneidad, el p-value del test (Robust regression F) es menor a 0.05, y se rechaza la H0, es decir, las variables no son endógenas. Por ello, debería usarse el modelo estimado por IV, pues, la inconsistencia penalizada es grande, y sólo se puede confiar en este estimado (IV); además, es necesaria la instrumentación para el modelo.