**TPº10: ESTIMACIÓN CON VARIABLES PROXY**

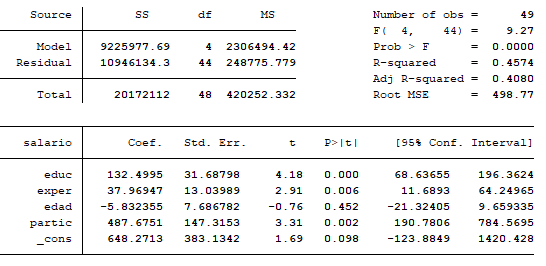
**EJERCICIO Nº 1:**

A partir de los datos del archivo Impacto.dta, se desea estimar el impacto de un curso de capacitación del Ministerio de Trabajo, a través del siguiente modelo:

Donde SALARIO es el monto salarial percibido a partir del 6º mes de finalizado el curso, EDUC los años de instrucción del individuo, EXPER los años de experiencia laboral y PARTIC una variable binaria que indica si participó o no del curso.

**a) Estime el modelo por MCO y mencione cuál es el impacto promedio de la participación en el curso en el salario. ¿En qué casos podría sospechar que existe un sesgo?**

Se propone el siguiente modelo para evaluar el impacto promedio de la participación en el curso en el nivel de salario, sin omitir variables:



A partir del modelo y los indicadores resultantes, podemos observar lo siguiente:

Coeficientes:

* Por cada aumento unitario de educación, manteniendo el resto constante, el salario se incrementa en promedio 132.50 unidades monetarias. Siendo esta una variable significativa al 5% de confianza.
* Por cada incremento unitario en la variable 'exper', ceteris paribus, la variable explicada salario se ve afectada positivamente en promedio en 38 unidades monetarias. Siendo la variable experiencia también significativa con un p-valor de 0.006 (<0.05)
* Por cada año que se suma a la variable edad, para el resto de variables fijas, el salario disminuye en promedio 5.8 unidades monetarias, siendo este el único coeficiente negativo y mostrando insignificancia para su pvalor (. Dada la magnitud de la insignificatividad, podemos considerar eventualmente quitarla del modelo.
* El coeficiente de la variable ficticia "Participación", en referencia al programa de capacitación, nos muestra que en el caso de las observaciones que sí participaron (=1) su salario aumenta en promedio unas 487.6 unidades monetarias, para el resto de variables constantes. La variable muestra relevancia estadística a un nivel de significatividad del 5% (para un pvalor )

Eso podemos interpretarlo en el gráfico de la regresión como un desplazamiento superior por sobre la constante, para aquellas personas que tengan valor 1 en la variable cualitativa.

En conclusión, el impacto promedio sobre el salario es de aproximadamente casi unas 500 unidades monetarias, alrededor de 75% de lo que representa el intercepto en sí mismo.

En relación al sesgo, verificamos correlación lineal entre las variables pero no podemos inferir nada.

**b) Proponga un modelo que incluya la variable no observada CAPACIDAD. ¿Es posible estimarlo? ¿Qué variables proxy propone para controlar el efecto de la capacidad individual sobre el salario?**

Si proponemos un modelo formal que incluya en la ecuación general la variable NO observada capacidad individual, tenemos que:

De todas maneras, este modelo no es posible estimarlo directamente, ya que no hay forma de medir u observar la variable capacidad individual.

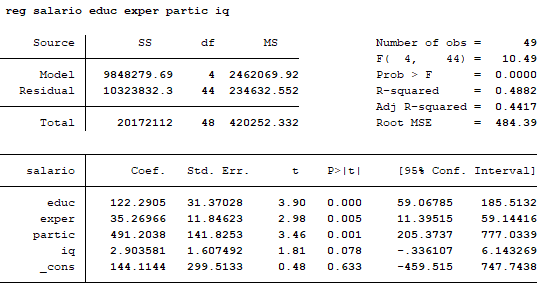
A modo de solución “suplente” al problema de variables omitidas, podríamos utilizar una variable proxy con el objetivo de controlar el efecto de la capacidad sobre el salario. Dado que los requisitos para que la variable proxy pueda ser incluida son:  
  
 - El error u no debe estar correlacionado con las variables independientes ni la proxy.   
 - Que sea “buena” proxy: su error v no puede estar correlacionado con las variables independientes.

Además de que la proxy debe estar relacionada con la variable en cuestión, podemos usar el coeficiente intelectual (IQ), el cual si se puede medir directamente.

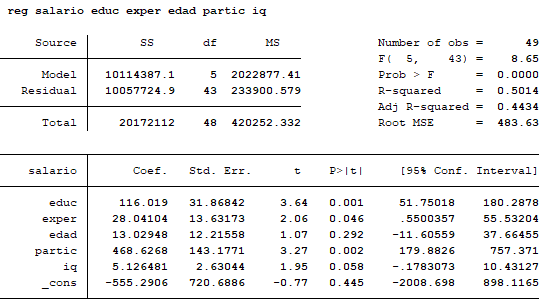
**c) Estime el modelo inicial incluyendo como variable proxy el coeficiente intelectual (IQ). Compare los resultados con el modelo inicial y concluya.**

Incluyendo al modelo inicial la variable proxy de coeficiente intelectual (IQ) tendríamos la siguiente ecuación (marcando con color aquellos términos que resultan diferentes al modelo original propuesto):

En stata correríamos el siguiente código:



Comparando el modelo inicial, que no incluye la variable proxy, podemos observar las siguientes diferencias:



Los coeficientes cambian significativamente en promedio, ya que los positivos disminuyen y edad, que presentaba un beta negativo, paso a ser positivo. ¿Podría decirse que había endogeneidad?

La variable proxy no es estadísticamente significativa para un pvalor de . Aunque por muy poco, prácticamente podríamos decir que lo es.

Pregunta; porqué al incluir edad en la regresión la constante nos pasa de 144 a -555?   
Por que omitis una variable relevante y tenes la varianza sesgada. La media de la edad y su efecto no puede ir en el termino de error por que por construccion tiene media cero, la que capta los efectos medios de los inobservables es la constante, por eso es que la constante cambia.

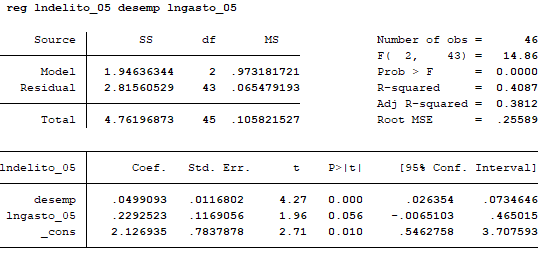
**d) ¿El estimador de PARTICIP es consistente? ¿y el de CAPACIDAD?**

**EJERCICIO Nº 2:**

Se desea evaluar el impacto del desempleo (DESEMP) y del gasto en seguridad (GASTOSEG) sobre la cantidad de delitos cada 1000 habitantes (DELITO) en 46 ciudades argentinas. Para ello se dispone de los datos del año 2005 en el archivo DELITOS.dta. Se pide:

**a) Estime el modelo log(𝑑𝑒𝑙𝑖𝑡𝑜)=𝛽0+𝛽1𝑑𝑒𝑠𝑒𝑚𝑝+𝛽2log(𝑔𝑎𝑠𝑡𝑜𝑠𝑒𝑔)+𝑢. Interprete los resultados.**

Para la salida de regresión del modelo estimado podemos inferir lo siguiente:

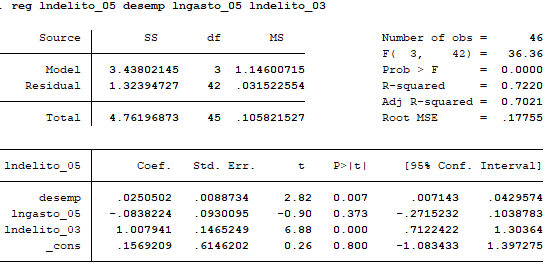


* Por cada incremento porcentual (1%) en la tasa de desempleo, manteniendo el resto de variables constantes, la tasa de delitos cada mil habitantes, se incrementa en promedio 0.049%. Siendo esta una variable significativa para un pvalor de 0<0.05
* Ante un aumento porcentual (1%) en el gasto público sobre seguridad, ceteris paribus, la tasa de delitos cada mil habitantes se incrementa en promedio 0.23%. Observando insignificancia estadística en la variable para un pvalor de 0.056>0.05

Para la muestra de 46 observaciones, el modelo posee una bondad de ajuste de aproximada de 41%, con una R-cuadrada de 0.4087. Hay significancia estadística conjunta, a partir de un test F con pvalor 0<0.05.

**b) Incorpore como variable independiente la tasa de delincuencia del año 2003 a fin de controlar los factores no observables urbanos que afectan la delincuencia y que pueden estar correlacionados con el gasto corriente en seguridad pública.**

Adherimos dicha variable dependiente rezagada en 2 años como variable independiente explicativa, dejando la ecuación del modelo de la siguiente manera:



* Por cada incremento porcentual (1%) en la tasa de desempleo, manteniendo el resto de variables constantes, la elasticidad de la tasa de delitos cada mil habitantes respecto el desempleo, se incrementa en promedio 0.025%. Siendo esta una variable significativa para un pvalor de 0.007<0.05
* Ante un aumento porcentual (1%) en el gasto público sobre seguridad, ceteris paribus, la tasa de delitos cada mil habitantes, disminuye en promedio 0.08%. Observando una insignificancia estadística en la variable para un pvalor de 0.373>0.05

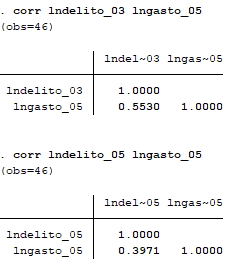
Para la muestra de 46 observaciones, el modelo posee una bondad de ajuste de aproximada de 72%, con una R-cuadrada de 0.722. Hay significancia estadística conjunta, a partir de un test F con pvalor 0<0.05.

**c) Compare ambos modelos y extraiga conclusiones.**

A partir de las estimaciones realizadas para ambos modelos y las salidas de regresión correspondientes podemos inferir las siguientes conclusiones:

* Incorporando la variable explicada rezagada en dos períodos como una variable independiente explicativa () vemos que el modelo cobra mayor significancia estadística por los siguientes indicadores:   
  - El coeficiente de la tasa de gasto público paso a ser negativo, teniendo una disminución de 0.312 (), lo que intuitivamente nos explica mejor la tasa de delitos, aunque su insignificancia estadística se incrementó considerablemente (para un pvalor de 0.373).  
  - El modelo posee una mayor bondad de ajuste (comparando ) lo que resulta más lógico para las variables explicativas que poseemos, aunque esto ya era de esperarse.

Claramente, la variable tasa de delitos cada mil habitantes tiene componentes inerciales en el tiempo que afectan a si misma en el presente. Incluir el rezago de esta variable de dos años atrás como explicativa proporciona una manera de dar cuenta de estos factores históricos que ocasionan diferencias actuales en la variable dependiente que, de otra forma, serían difíciles de explicar. Si obtenemos las correlaciones entre el gasto público en seguridad y ambas variables en el tiempo, obtenemos los siguientes valores:

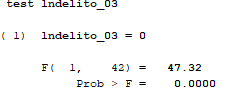
Podemos observar que el gasto público en seguridad esta más correlacionado con la tasa de delitos del año 2003, lo que puede ser una posible idea de lo que nos generó mayor insignificancia estadística en la variable lngasto.

Por el otro lado, la tasa de gasto público tiene una menor correlación con la tasa de delitos cada mil habitantes del año 2005, por lo que podemos inferir que resulta adecuada la incorporación de lndelito\_03.

Intuitivamente podemos pensar que las ciudades con un historial mayor de seguridad, tenderán a gastar más en la actualidad.

Contrariamente dado que estaban sesgados los coeficientes por la variable omitida, en el modelo original que corrimos el coeficiente era positivo, lo cual no tiene sentido.

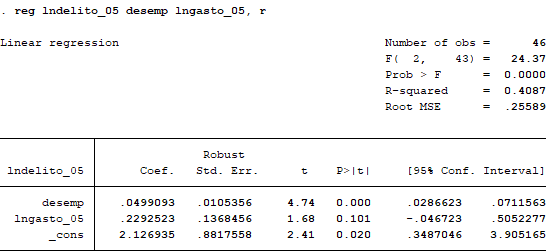
**d) Pruebe la hipótesis nula de que el parámetro de log(𝐷𝐸𝐿𝐼𝑇𝑂\_03) es igual a uno contra la alternativa de dos colas. Concluya.**

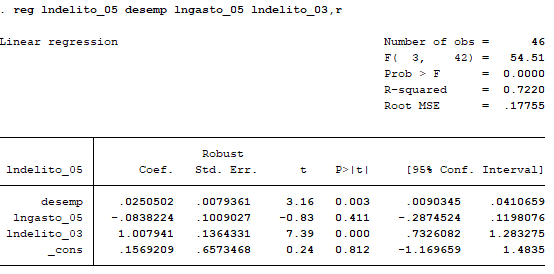


Suponiendo que el test lo puedo hacer así:

Para un p-valor de 0.0 (<0.05) podemos rechazar la hipótesis nula en favor de que la variable de logaritmo de la tasa de delitos del año 2003 es significativa y con un coeficiente positivo.

**e) Repita los incisos anteriores empleando los errores estándar robustos a heterocedasticidad y analice brevemente cualquier diferencia notable.**





En comparación con los modelos anteriores, podemos observar estadísticos t y F mayores.

Los errores estándar robustos y los estadísticos t robustos se justifican sólo si el tamaño de muestra se hace grande. Con tamaños de muestra pequeños, el estadístico t robusto puede tener distribuciones que no estén muy próximas a la distribución t y que podrían echar a perder la inferencia, lo que para este caso no podemos determinar con mucha certeza si podemos emplear robustez o no.

**ACTIVIDAD Nº 1:**

Responda si las siguientes afirmaciones son verdaderas o falsas:

a) Una variable binaria no puede utilizarse como variable proxy.

**Falso: Si se pueden emplear variables binarias o dummies como variables proxy.**

b) En fenómenos con un importante componente inercial no es conveniente incluir la variable dependiente rezagada como variable explicativa.

**Falso: por el contrario, cuando se tiene componentes inerciales en el tiempo que afectan a una variable en el presente, incluir la variable rezagada como explicativa proporciona una manera de dar cuenta de estos factores históricos que ocasionan diferencias actuales en la variable dependiente, que de otra manera, serían difíciles de explicar.**

c) Variable proxy es un método de estimación distinto a MCO.

**Falso: incluir una variable proxy es una solución alternativa al problema de omisión de variable, no está relacionado con estimaciones estadísticas de modelos, como MCO.**

**ACTIVIDAD Nº 2:**

a) ¿Qué es una variable proxy?

**Es una variable que incluimos dentro de un modelo, para poder explicar un efecto sobre la variable explicada que no está siendo captado por ninguna variable independiente, sea porque no la podemos medir, fue excluida, etc. Es decir, la proxy está relacionada con algo no observado y nos ayuda a solucionar el problema de variable omitida.**

b) ¿Qué requisitos debe cumplir una variable proxy para proporcionar estimadores consistentes? ¿Qué sucede si no los cumple?

1. **El error no debe estar correlacionado con las variables independientes ni la proxy.**
2. **Que sea “buena” proxy: su error no puede estar correlacionado con las variables independientes.**

**Si no cumple los supuestos, entonces tendremos estimadores sesgados.**

c) Busque un ejemplo de variable proxy en un paper aplicado de econometría. Comente.

**TPº11: V. INSTRUMENTAL Y MÍNIMOS CUADRADOS EN 2 ETAPAS**

**EJERCICIO Nº 1: salteado por ser regresión simple**

**EJERCICIO Nº 2: continuación del ej nro 1**

**EJERCICIO Nº 3:**

(1995) (CARD.dta) estimó el impacto de educación sobre el salario de un grupo de hombres en 1976 a partir del siguiente modelo teórico:

log(𝑠𝑎𝑙𝑎𝑟𝑖𝑜)=𝛽0+𝛽1𝑒𝑑𝑢𝑐+𝛽2𝑒𝑥𝑝𝑒𝑟+𝛽3𝑒𝑥𝑝𝑒𝑟2+𝛽4𝑟𝑎𝑧𝑎+𝛽5𝑐𝑎𝑝𝑖𝑡𝑎𝑙+𝛽6𝑠𝑢𝑟+𝑢

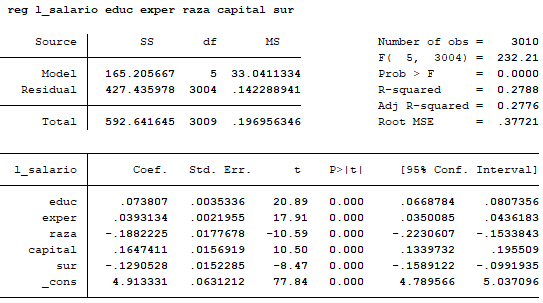
donde RAZA es una variable binaria para raza negra, CAPITAL una variable binaria por vivir en un área metropolitana y SUR una variable binaria por vivir en el sur. También se incluyen un conjunto completo de variables binarias regionales y una variable binaria de área metropolitana para el lugar donde el hombre vivía en 1966.

**a) ¿Por qué la variable EDUC podría ser endógena en este caso?**

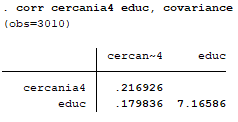
La variable de educación podría llegar a ser determinada endógenamente a partir de las variables exógenas socio-geográficas como capital y raza. Además, cuando trabajamos con diversos grupos sociales corremos el riesgo de incurrir en un problema de Heterogeneidad inobservable 🡪 los individuos en una población difieren en características que no son observables y por lo tanto omitidas en la especificación del modelo. Es similar al problema de variables omitidas.

**b) Estime la ecuación por MCO y explique cuáles son las consecuencias de la heterogeneidad inobservable sobre los EMCO.**

Como se mencionó anteriormente, un problema de heterogeneidad inobservable es posible causa de endogeneidad en una variable, que posee consecuencias similares al problema de variable omitida. Siguiendo este razonamiento, eso implica que tendremos estimadores sesgados y



**c) Card utilizó como variable instrumental para EDUC una variable binaria que era igual a 1 si la persona había crecido cerca de una universidad con carreras de 4 años (CERCANÍA4). Verifique el supuesto de relevancia del instrumento.**



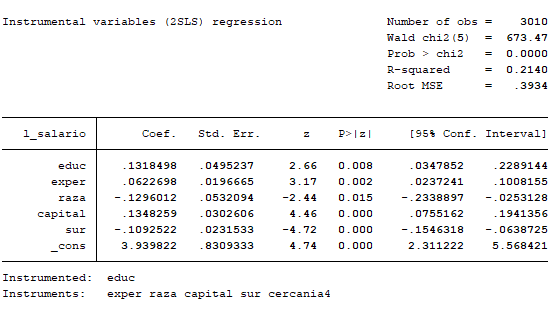
Recordando la condición de relevancia para el instrumento como , tenemos que:

Aplicando la matriz de covarianza entre la variable instrumentada y el instrumento podemos verificar el supuesto de relevancia para cercania4 (la variable instrumental). Dado que la covarianza entre estas es de y por ende, se cumple dicha condición de relevancia.

El instrumento resulta ser válido.

**d) Argumente sobre el cumplimiento del supuesto de exogeneidad del instrumento. Realice el Ejercicio para computadora 15.3 de Wooldridge para profundizar sobre este punto.**

**e) Estime la ecuación por MC2E utilizando el instrumento propuesto por Card y compare los coeficientes estimados con los EMC del punto a). Interprete el coeficiente asociado a EDUC y realice una estimación por intervalos para ese parámetro con ambos métodos de estimación**.



Aplicando el método de Minimos Cuadrados en Dos Etapas obtenemos la anterior salida de regresión, de la cual podemos inferir lo siguiente:

* El coeficiente de educación resulta ser diferente al original, gracias a la instrumentación y reducción del sesgo, generando un efecto marginal promedio sobre el logaritmo del salario de 13.18%, para un incremento de 1 año en la educación, siendo esta una variable significativa para un pvalor de 0.008<0.05.
* El resto de coeficientes sufren cambios de intensificación en su magnitud con la misma dirección, aquellos positivos se incrementan y aquellos negativos disminuyen. Ninguna variable deja de ser significativa estadísticamente.

**f) Obtenga los residuales de la forma reducida y utilícelos para probar si EDUC es exógena. Es decir, determine si la diferencia entre MCO y VI es estadísticamente significativa.**

**ACTIVIDAD Nº 1:**

**Responda si las siguientes afirmaciones son verdaderas o falsas:**

**a) Cuando existen errores de medición en la variable a explicar los Estimadores de MCO son sesgados pero consistentes.**

FALSO: - Si tenemos que el error no está correlacionado con la variable observada mal medida, pero sí con la variable no observada (la correcta). Entonces, usar la variable medida con error proporciona una estimador **consistente** de pero un aumento de la varianza de los estimadores. No sabremos si el estimador será sesgado/insesgado, ya que es difícil de demostrar, pero nos recostamos sobre la ventaja de que cumple con la consistencia.

- Si en este caso tenemos que el error está correlacionado con la variable observada mal medida, pero incorrelacionado con la variable no observada (la correcta). Entonces, usar la variable medida con error resulta en estimadores **sesgados e inconsistentes** de y el efecto estimado por MCO estará **atenuado**. Una solución podrá ser con variables instrumentales. Generalmente, siempre que un estimador es inconsistente, termina siendo sesgado.

**b) Es posible comprobar empíricamente la condición de exogeneidad del instrumento, pero no la de relevancia.**

FALSO: Por el contrario, es más fácil comprobar la condición de relevancia del instrumento

**ACTIVIDAD Nº 2:**

**a) Enumere causas posibles para la existencia de endogeneidad.**

* **Heterogeneidad inobservable** 🡪 los individuos en una población difieren en características que no son observables y por lo tanto omitidas en la especificación del modelo. Es similar al problema de variables omitidas…
* **Errores de medida** 🡪 cuando el error de medida está correlacionado con la variable observada, contenido en el error.
* **Simultaneidad** 🡪 cuando la variable dependiente y alguna variable explicativa se determinan simultáneamente. Un ejemplo clásico de esto es cuando se estiman ecuaciones de demanda u oferta, el precio y la cantidad se ven en simultaneo. Otro ejemplo es tasa de inseguridad y gasto público, ¿cuál determina a cuál? El huevo o la gallina.
* **Error en forma funcional o Errores de medida en variables** que se van al término de error del modelo**.**

**b) ¿Cuál es la consecuencia sobre las propiedades de los EMC de la presencia de endogeneidad?**

**c) Explique el método de Mínimos Cuadrados en 2 Etapas.**

Se elige la combinación lineal que presente la mayor correlación la variable endógena, que vendrá dada por la ecuación de forma reducida. Entonces el valor estimado de la endógena en la ecuación reducida será el mejor instrumento, el proceso sería:

- En una primer etapa estimamos la forma reducida de la variable endógena.  
- En la segunda etapa estimamos la ecuación estructural.