MÉTODOS ESTADÍSTICOS APLICADOS A LOS NEGOCIOS

TRABAJO PRÁCTICO 2 - 20/09/2017

GRUPO
ANTÓN MARÍA PAZ
FIGLIOLIA JULIETA ALDANA
PONCIO FEDERICO
SABATER ANNA
WANG JIA QI

NOTA: Por limitación de espacio las salidas de STATA comentadas se encuentran adjuntas en el archivo logfile y dofile.

PUNTO 1

Ejercicio 1a

En la muestra, hay 501 distritos escolares observados a lo largo de 7 años, desde 1995 a 2001. El análisis descriptivo de las variables indica que el panel está fuertemente balanceado. Por lo que el panel tiene n*T observaciones, es decir 501*7=3507 observaciones en total.

La base incluye las siguiente variables: un identificador para cada estado, las dummies correspondiente para cada año del período 1995 a 2001, la fracción elegible de almuerzo libre, el gasto educativo por alumno, los fondos recibidos durante 1995-2001, el enrolamiento o matriculación del distrito y transformaciones de algunas variables ya mencionadas como el logaritmo de la experiencia promedio o el logaritmo de los fondos recibidos.

La variable dependiente a explicar es el desempeño en matemática de 4to grado y está medida en una fracción satisfactoria, es decir es un puntaje que representa el desempeño de matemática de 4to grado de cada distrito escolar durante los años 1995 a 2001.

Ejercicio 1b

Se estima el siguiente modelo por POLS:

$$math4it = \beta_0 + \beta_1 lavgrexpit + \beta_2 lunchit + \beta_3 lenrollit + u_{it}$$

De acuerdo a la salida de STATA todos los coeficientes son estadísticamente significativos, tanto de manera individual como conjunta. Esto se observa en los p-valores correspondientes a cada variable y el test F, los cuales computan 0.000. Teniendo en cuenta que el modelo es lineal-log, se estima que:

- Manteniendo constante el resto de las variables, un aumento de un 1% en el gasto educativo por alumno, deflactado por IPC, estás asociado con un aumento en el desempeño en matemática de 4to grado 0.2828*0.01=0.0028 puntos.
- Manteniendo constante el resto de las variables, un aumento de una unidad de fracción elegible para almuerzo libre reduce el desempeño en matemática de 4to grado en un 0.4121 puntos.
- Manteniendo constante el resto de las variables, un aumento de un 1% en el logaritmo del número de estudiantes enrolados o matriculados en el distrito produce una baja en el puntaje de desempeño en matemática de 4to grado de 0.000095 puntos.
- La constante predice que, en el caso hipotético de que las variables *lavgrexpit*, *lunchit* y *lenrollit* fueran cero, el desempeño promedio en matemática de 4to de un distrito sería de -1.59 puntos.

Ejercicio 1c

Se estima el siguiente modelo por efectos fijos y sin ajustar la varianza:

$$math4it = \alpha_i + \beta_1 lavgrexpit + \beta_2 lunchit + \beta_3 lenrollit + u_{it}$$

La constante correspondiente a esta salida es el término α_i, el cual representa los efectos fijos individuales. Es decir, es un intercepto específico para cada uno de los distritos escolares que absorbe las influencias de todas las variables omitidas que difieren de una distrito escolar a otro pero que son constantes en el tiempo. Por ejemplo, se podría mencionar el nivel socioeconómico de un distrito, las condiciones de aprendizaje y acceso a materiales, las características demográficas y culturales de la comunidad del distrito, si los estudiantes viven(o no) en un distrito de zona rural o marginal, la extensión territorial del distrito.

En comparación el inciso anterior, el modelo con efectos fijos estima que:

- Manteniendo constante el resto de las variables, un aumento de un 1% en el logaritmo de la experiencia promedio de un alumno está asociado con un aumento en el desempeño en matemática de 4to grado de 0.8061*0.01=0.008061 puntos. Antes, el efecto de la experiencia promedio en el desempeño era de 0.0028 puntos.

- Manteniendo constante el resto de las variables, un aumento de una unidad de fracción elegible para almuerzo libre incrementa el desempeño en matemática de 4to grado en un 0.0594 puntos. Antes este efecto era negativo y contra-intuitivo, dado que era de -0.004121 puntos.
- Manteniendo constante el resto de las variables, un aumento de un 1% en el logaritmo del número de estudiantes enrolados produce una suba en el puntaje de desempeño en matemática de 4to grado de 0.04292*0.01=0.0004292 puntos. Antes el efecto era negativo de -0.000095 puntos.
- La constante en este modelo es el término α_i que representa los efectos fijos individuales, los cuales estaban siendo omitidos en la regresión anterior y causaban sesgo e inconsistencia. Esta constante α_i estima que, en el caso hipotético de que las variables $lavgrexp_{it}$, $lunch_{it}$ y $lenroll_{it}$ fueran cero, el desempeño promedio en matemática de 4to de un distrito sería de -6.70 puntos, lo cual resulta contra-intuitivo porque las notas no pueden negativas.

Es decir que la estimación cambió tanto en los valores de los coeficientes como en su significatividad individual.

En particular, se observa que se modificaron los signos de los coeficientes de las variables *lunch* y *lenroll,* los cuales pasaron de tener signo negativo a positivo y ahora resultan ser más apropiados con la teoría económica. A mayor fracción elegible para almuerzo libre se espera un mayor desempeño académico. A mayor cantidad de alumnos enrolados o matriculados se estima un mayor desempeño académico. Estas variables *lunch* y *lenroll* podrían estar reflejando un factor socioeconómico del distrito escolar, por ejemplo, una mayor fracción elegible para almuerzo libre en un distrito puede estar correlacionado con un bajo nivel socioeconómico de los alumnos de ese distrito.

El modelo en su conjunto es estadísticamente significativo a un nivel del 1%, lo cual indica que las variables en su conjunto sirven para explicar la variable dependiente $math4_{it}$. No obstante, *lunch* y *lenroll* no son variables estadísticamente significativas a un nivel de 10% de significancia. Las únicas dos variables estadísticamente significativas a un nivel del 1% son *lavgrexp* (el logaritmo de la media de la experiencia por distrito) y la constante α_i que representa los efectos fijos individuales. Esto es un indicio que la inclusión de los efectos fijos en la regresión permite evitar el sesgo de variable omitida.

Ejercicio 1d

Por un lado, el coeficiente *rho* que reporta STATA es un ratio entre los residuos de la varianza within del modelo y la suma de la varianza de los residuos within con la varianza de residuos total. Es decir, *rho* es un indicador de la participación que tienen los efectos fijos en la medición y representa el porcentaje de varianza total capturada por la varianza del error within.

Un *rho* alto o cercano a 1 indica que un alto porcentaje de la varianza del error de la estimación provenía de las diferencias por efectos fijos, es decir que mayor es la participación de los efectos fijos y mayor es su peso en la explicación de la varianza del error.

La salida del ejercicio 1c computa un rho de 0.7460, es decir que un 74,60% de la fracción de varianza se debe a los errores que están dentro del término u_i que contiene a los efectos fijos. Esto es un indicio que los efectos fijos explican en un 74.60% el desempeño en matemática de 4to grado.

Por otro lado, STATA reporta tres medidas de R^2 diferentes en el modelo de efectos fijos:

El R^2 *overall* es la fracción de variabilidad en la variable dependiente explicada por la predicción de los coeficientes del modelo, es como el R^2 del modelo de regresión lineal y resulta ser un promedio ponderado de los dos R^2 restantes. En la salida se reporta un R^2 *overall* de 0.0443.

El R^2 *within* indica cuánto de la variación <u>*dentro*</u> de cada unidad es explicada por el modelo. Es decir, fijando un individuo *i* se observa la variación explicada en la dimensión *t*. La salida computa un R^2 *within de 0.3233*, es decir que en un 32,33% es explicada la variación dentro de cada distrito escolar en el desempeño en matemática de 4to grado.

De forma análoga, el R^2 *between* indica la variación <u>*entre*</u> los individuos explicada por el modelo. Es decir que se fija en un tiempo *t* y se observa la variación en la dimensión de los individuos *i*. La salida registra un R^2 *between de 0.0037*.

Ejercicio 1e

La transformación within surge como respuesta al siguiente problema: se quiere estimar un modelo con variación en dimensión temporal y espacial pero se quiere evitar el sesgo producido por variables que solamente varían en dimensión espacial (entre individuos, pero no en el tiempo). Si se estimara sin tener en cuenta la transformación within, el término de error contendría a estas variables unidimensionales.

La transformación within parte del modelo poblacional propuesto en la ecuación (1):

$$y_{it} = \alpha_i + \beta_1 X_{1it} + .. + \beta_k X_{kit} + u_{it}$$

Donde el α_i representa el conjunto de variables que sólo varían espacialmente y el modelo contiene k variables explicativas de variación bidimensional.

A partir de este modelo, se suman las observaciones para los T momentos de tiempo y luego divide por T. De esta manera se obtiene la *media temporal* de las variables en la ecuación (2). Luego se restan ambos modelos haciendo ecuación(1)-ecuación(2) y esto elimina el efecto constante de la variable espacial α_i, permitiendo así estimar los coeficientes beta:

$$y_{it} = \alpha_i + \beta_1 X_{1it} + .. + \beta_k X_{kit} + u_{it} \quad \textbf{(1)}$$

$$\frac{1}{T}\sum_{t=1}^{T} y_{it} = \overline{y_i} = \frac{1}{T}\sum_{t=1}^{T} \alpha_i + \frac{1}{T}\sum_{t=1}^{T}(\beta_1 X_{1it} + .. + \beta_k X_{kit}) + \frac{1}{T}\sum_{t=1}^{T} u_{it} = \alpha_i + \beta_1 \overline{X}_{1i} + .. + \beta_k \overline{X}_{ki} + \overline{u}_i \quad \textbf{(2)}$$

$$\textbf{(1)-(2)} = (y_{it} - \overline{y}_i) = \beta_1(X_{1it} - \overline{X}_{1i}) + ... + \beta_k(X_{kit} - \overline{X}_{ki}) + (u_{it} - \overline{u}_i)$$

En este último modelo, se observa que los coeficientes *βeta* son los mismos que en el modelo original pero con la diferencia que no está más el conjunto de variables de dimensión espacial.

Notar que al estimar por $(y_{it} - \overline{y}_i)$ se está controlando por los efectos fijos.

Cuando se realizan estos pasos de manera manual en STATA, se observa que los coeficientes estimados son idénticos al modelo por efectos fijos. No así los errores estándares ya que éstos son corregidos al estimar efectos fijos con la opción de errores robusto o agrupados por individuos denominada *vce(cluster distid)*. (referirse al Do y Log files).

Ejercicio 1f

Se estima el mismo modelo que en el inciso 1c pero ahora se ajusta la estimación de la varianza con la opción de errores robustos o agrupados por individuos denominada *vce(cluster distid)*.

Se mantienen las conclusiones e interpretaciones mencionadas anteriormente ya que en la salida se reportan los mismos coeficientes de cada variable e idénticos R^2, *rho* y test F de significatividad conjunta.

No obstante, se observa que los errores estándares robustos aumentaron con respecto a la estimación anterior. Por lo que empeoraron los p-valores de significatividad individual de las variables *lunch y lenroll.* Esto sería un indicio que el efecto individual de estas variables no resulta estadísticamente relevante para explicar el desempeño en matemática de 4to grado pero sí lo son de manera conjunta en el modelo. Entonces *lunch y lenroll* serían variables de control.

Por este motivo, se concluye no descartarlas de la estimación, a pesar de que no cumplen con el test t de significatividad. Algunas razones de esta no significatividad podría ser un error en la medición de estas variables y el sesgo por variables omitidas como el nivel socioeconómico de cada distrito escolar, las condiciones de aprendizaje y acceso a materiales, las características demográficas y culturas de la comunidad de cada distrito.

Ejercicio 1g

Se incorporan 6-1 dummies temporales, es decir que se omitió la dummy del año 1995 y su efecto está siendo capturado por la constante.

En esta salida se observan cambios en los signos y magnitudes de los coeficientes. En particular, el coeficiente de *lunch* pasó de tener un signo positivo a uno negativo de -0.04194 y la variable no es significativa individualmente. Es decir que un aumento de una unidad de fracción elegible de almuerzo libre reduciría el desempeño en matemática de 4to grado en un 0.04194 puntos. Esto pareciera ir en contra de la intuición de la teoría económica. A su vez, empeoró el p-valor de la variable *lenroll* y tampoco es estadísticamente significativa.

No obstante, el p-valor del test F del modelo y del test de las dummies es 0.000, por lo que las variables temporales son conjuntamente significativas para explicar $math4_{it}$. En detalle, todas las dummies del período son significativas individualmente a un nivel de 1% de significancia, excepto, la dummy correspondiente al año 1996, la cual es significativa a un nivel del 8%. Esto indicaría que los efectos fijos temporales son significativos.

Además, el R^2 *within* estimado es 0.4713, es decir que en un 47,13% es explicada la variación dentro de cada distrito escolar en el desempeño en matemática de 4to grado. El *rho* es de 0.6488, con lo cual un 64,88% de la fracción de varianza se debe a los errores que estaban dentro del término u_i que contiene a los efectos fijos.

Por lo tanto, se concluye que no es correcto estimar por POLS y se deben considerar tanto los efectos fijos como temporales. A causa de la no significatividad individual de *lunch y lenroll*, conviene revisar si las variables utilizadas en el modelo son útiles para explicar el desempeño en matemática de 4to grado.

ANEXO: Ahondando en el estudio de la significatividad de *lunch* y *lenroll*, pensamos en estudiar dentro del modelo de efectos fijos, dummies temporales de lunch y lenroll. La motivación detrás de esto es que, si se asume que *lunch* y *lenroll* son teóricamente relevantes para explicar *math4*, entonces su no significatividad en el modelo puede provenir de una gran variación temporal en sus efectos y no sólo en las constantes del modelo (ya contadas por las dummies anuales).

Corrimos dos pares de regresiones: en el primero, una regresión incluye, en vez de *lunch*, sus interacciones temporales; y la otra las interacciones de *lenroll*. En el segundo par, se corrieron los modelos anteriores, pero incorporando también las dummies anuales por constantes (quitando las dummies necesarias para evitar la colinealidad).

Corridos estos modelos, siguen sin ser significativas las variables que interactúan anualmente, pero sí son significativas las dummies anuales. Los test F de las cuatro regresiones son significativos, pero estos pueden estar inflados por la alta multicolinealidad inducida por introducir tantas dummies. Los comandos correspondientes se encuentran en la sección "ANEXO" del Do y Log Files.

Luego de estos intentos, quedaría explicar por qué no serían teóricamente relevantes las variables ya que estadísticamente no lo son y, si se rehace la regresión original pero sin incluir *lunch* y *lenroll*, sigue siendo significativa, con el mismo R^2 ajustado.

Entonces, la variable *lunch*, incluida como indicador socioeconómico, puede no ser relevante teóricamente para explicar el rendimiento de los alumnos de 4to grado. Esto podría ocurrir porque dichos alumnos son niños pequeños como para que realicen aportes económicos en su hogar en el caso de tener nivel socioeconómico bajo. Siendo tan pequeños, la

presión por mantenerlos en escolaridad sería todavía alta. Quizás esta variable sería más relevante para explicar el rendimiento de alumnos en cursos mayores. Por este motivo, en cursos de menor edad se podría optar por medir el impacto del nivel socioeconómico en el desempeño matemático mediante la variable *calidad de nutrición de los alumnos* durante sus primeros años, en vez de la coyuntura económica del momento en 4to grado.

Con respecto a *lenroll*, dicha variable estaría en el modelo para indicar el grado de saturación de las escuelas y la relación esperada sería es que mayor estudiantes menor el rendimiento escolar. De todas formas, para capturar este efecto podría ser más adecuado usar variables como *cantidad de alumnos por clase* o *cantidad de profesores por alumno*.

PUNTO 2

Ejercicio 2a

Se estima con efectos fijos el siguiente modelo y sin ajustar la estimación de la varianza.

$$\ln(\text{WAGE}_{it}) = \beta_{1i} + \beta_2 \text{EXPER}_{it} + \beta_3 \text{EXPER}^2_{it} + \beta_4 \text{SOUTH}_{it} + \beta_5 \text{UNION}_{it} + \varepsilon_{it}$$

El coeficiente de la constante β_{1i} representa los efectos fijos individuales, es decir que capta las influencias de todas las variables omitidas que difieren de una mujer a otra pero que son constantes en el tiempo. Por ejemplo, algunos factores que no se estarían midiendo en este modelo son la habilidad de la mujer, la educación del padre y/o la madre, el nivel socioeconómico del hogar.

Se evalúa la hipótesis nula de que la constante es idéntica para todas las mujeres de la muestra y se observa un p-valor de 0.000 del test F. Por lo que, son estadísticamente significativos las variaciones de los efectos fijos individuales de cada mujer en la estimación del salario.

Creemos que es recomendable re-estimar el modelo ajustando la varianza por cluster porque, al trabajar con datos en dimensión temporal, aparece el problema de autocorrelación y las variables pueden estar potencialmente correlacionadas en el tiempo. Como los errores heterocedásticos no son válidos, se sugiere trabajar con los errores estándar consistentes a heterocedasticidad y autocorrelación (HAC). Los errores estándar agrupados son un tipo de HAC y éstos permiten la presencia de una correlación arbitraria dentro de un conglomerado, agrupación o "cluster" (en este caso, dentro de cada mujer encuestada) pero se supone que los errores de la regresión no están correlacionados entre los grupos de mujeres. Es decir, utilizando los errores clusterizados se permite que exista una correlación intra-individuo a lo largo del tiempo pero que no haya correlación entre las distintas entidades individuales.

Ejercicio 2b

Se re-estima el modelo anterior ajustando la varianza por cluster. Para ello, se agrega el comando *vce(cluster id)*. No se observan cambios de gran magnitud en los errores estándar ni en los coeficientes estimados. El modelo en su conjunto sigue siendo estadísticamente significativo y se reportan idénticos valores de R^2 *within* y *rho* del modelo inicial.

Según la salida de STATA, el coeficiente de la experiencia *exper* es 0.0574. Esto significa que ante un aumento de un año de experiencia laboral el salario de la mujer se incrementa en un 5.74%. Este efecto es estadísticamente significativo a un nivel de 10% pero no al 5%.

Al mismo tiempo, se observa que el coeficiente de *exper2* no es significativo individualmente. Por esta razón, se concluye que la inclusión de esta variable no favorece la estimación. Una recomendación sería descartar *exper2* del modelo y correr la regresión para evaluar si mejora la significatividad individual de *exper*.

Ejercicio 2c

Se re-estima el modelo, ajustando la varianza y se incorpora como regresor una variable dicotómica llamada *d88* que toma valor 1 para el año 1988. Esta variable dicotómica está tratando de capturar el efecto temporal, es decir, las variaciones de 1987 a 1988 en las mujeres encuestadas en National Longitudinal Surveys.

Se observa que no son estadísticamente significativos los coeficientes de *d88* y de la constante, la cual refleja los efectos correspondiente al año 1987. En particular, el test de la variable *d88* arrojó un p-valor de 0.14 y no sería significativa a un nivel de 10%. La constante computó un p-valor de 0.225 y tampoco es significativa a un 10%.

Estos resultados parecen ser intuitivos ya que no debería haber un cambio importante de un año a otro en las encuestas realizadas a las mujeres.

A su vez, se observa que el modelo en su conjunto es significativo y el rho computado es de 0.8687, por lo que en un 86.7% la varianza está siendo explicada por los efectos fijos individual. Por esta razón, resultaría apropiado estimar por efectos fijos y no por POLS.

Ejercicio 2d

Partiendo de un modelo con efectos fijos, con el objetivo de evitar el sesgo por variable omitida, a continuación se detallan las razones a favor (en contra) de por qué incluiríamos (o no) cada una de las siguientes variables en la estimación del modelo:

- Hours. Creemos que las horas trabajadas podría ser una variable omitida que varié tanto entre las mujeres individuales como a lo largo del período 1987-1988. Por lo que, dejando de lado su efecto fijo

por colinealidad con la constante β_{1i}, sería adecuado incluir hours en la estimación para evaluar su efecto temporal en el salario. Identificamos los dos siguientes efecto opuestos. Por un lado, a mayor horas trabajadas se espera un mayor salario ya que se recibe una mayor compensación. Por otro lado, a mayor horas trabajadas se espera un menor salario dado que las mujeres consiguen empleo en sectores de peor pagos y más precarizados. En caso que *hours* varíe tanto entre las mujeres como a lo largo del tiempo, se podría utilizar el método de regresión por variables instrumentales).

- *Collgrad* toma valor 1 si la trabajadora tiene un título universitario y podría ser una variable no observable, que varía de una mujer a otra, pero que no cambia en el tiempo. Por lo tanto su efecto fijo ya estaría siendo captado por la constante β_{1i} y no la incluiríamos en el modelo por efectos fijos por colinealidad.
- *Msp* toma valor 1 si la trabajadora está casada y podría ser una variable omitida que varía tanto entre las mujeres individuales pero no a lo largo del tiempo. Por lo tanto su efecto fijo ya estaría siendo captado por la constante β_{1i} y no la incluiríamos en el modelo por efectos fijos por colinealidad. A su vez, tampoco nos parece conveniente incluirla ya que no creemos que el estado civil tenga un sentido relevante desde la teoría económica para explicar el salario de las mujeres.
- *Black* toma valor 1 si la trabajadora es de raza negra y podría ser una variable omitida que varía entre las mujeres individuales pero no a lo largo del tiempo. Por lo tanto su efecto fijo ya estaría siendo captado por la constante β_{1i} y no la incluiríamos en el modelo por efectos fijos por colinealidad. De todas maneras, creemos que la cuestión racial podría ser un factor determinante en el salario de las mujeres.
- Age es la edad de la trabajadora y podría ser una variable omitida que varíe tanto entre las mujeres individuales como a lo largo del período 1987-1988. Por lo que, dejando de lado su efecto fijo por colinealidad con la constante β_{1i}, sería adecuado que esté incluida en la estimación para evaluar su efecto temporal en el salario. podría ser ser una variable omitida que varía tanto entre las mujeres individuales pero no a lo largo del tiempo. Identificamos el siguiente efecto: a mayor edad se espera un mayor salario pero hasta una cierta edad debido a que rendimiento laboral podría caer a partir de la jubilación. A su vez, hay tener en cuenta una posible colinealidad entre las variables *age, exper y exper2* al incluirlas simultáneamente en el modelo.

Se estima el siguiente modelo que incluye las variables *hours* y *age* para las que encontramos argumentos positivos:

$$\ln(WAGE_{it}) = \beta_{1i} + \beta_2 EXPER_{it} + \beta_3 EXPER^2_{it} + \beta_4 SOUTH_{it} + \beta_5 UNION_{it} + \beta_6 HOURS_{it} + \beta_7 AGE_{it} + \varepsilon_{it}$$

Los resultados a partir de los valores que hubiéramos esperado ex-ante que tendrían esos parámetros dicen que:

Por un lado, *hours,* la constante β_{1i}, *y union* son estadísticamente significativas para explicar el salario de la mujer. Por otro, *exper2, south y age* no resultan estadísticamente significativas para explicar el salario de la mujer.

El coeficiente de age es -0.0152 y el p-valor computado es 0.366. Por lo que esta variable no es estadísticamente significativa para explicar el salario de la mujer. Si su efecto fuera significativo la variable predice que un año extra de edad en la mujer reduce su salario un 1.52%. Esto no resultaría apropiado con la teoría económica. De hecho, en comparación a nuestra argumentación sobre *age*, el efecto de la estimación no coincidió con la relación que esperábamos y tuvo un efecto opuesto al que propusimos inicialmente (a mayor edad se espera mayor salario).

El coeficiente de *hours* es -0.0131 y el p-valor computado es de 0.000, por lo que esta variable es estadísticamente significativa a un nivel de 1% para explicar el salario de la mujer. Su efecto estima que una hora extra de hora trabajada reduce el salario un 1.31%. Esta relación negativa entre horas trabajadas y salario podría explicarse porque las mujeres consiguen empleo en sectores de peor pagos y más precarizados, por lo que, su salario es menor.

El modelo es conjuntamente significativo ya que el p-valor del test F es 0.000. A su vez, la constante que representa los efectos fijos es significativa a un nivel de 1% y el rho computado es 0.8758, es decir, que la participación de los efectos fijos para explicar el salario de la mujer es de un 87%.