Problem Set 1

January 28, 2013

El objetivo de esta tarea es que aprendas a comprobar la aleatorización de una muestra dividida en un grupo control y otro de tratamiento utilizando los distintos comandos que STATA tiene para hacerlo. Dos objetivos secundarios son: (i) que aprecies la utilidad del supuesto de ignorabilidad de tratamiento y (ii) que aprendas a aleatorizar una muestra utilizando STATA. Para llevar a cabo esta tarea se utilizará la primera línea basal del programa PROGRESA (encaseh97) que está disponible en el sitio del curso dentro de la sección "Bases de datos."

Para que te familiarices un poco con PROGRESA y puedas entender el ejercicio práctico de una mejor manera, empieza leyendo la nota metodológica del programa que también se encuentra dentro de la sección "Bases de datos" del sitio del curso. En esta nota metodológica se describe el diseño de evaluación, el procedimiento de selección de la muestra, los instrumentos de recolección de los datos, así como la estructura de las bases de datos de la muestra de evaluación del programa.

Por favor entrega tus respuestas a las preguntas que aparecen aquí y una impresión de las tablas y preguntas que se hacen en el archivo de Excel llamado "Balanceo y aleatorización."

1. Familiarización con PROGRESA y la muestra para la evaluación del programa

Lee la nota metodológica y responde las siguientes preguntas sin exceder el límite de extensión para cada respuesta.

a) ¿Cuál es el propósito de PROGRESA? (Describir en sólo una frase.)

b) ¿Cuál fue el proceso que se siguió para identificar a los beneficiarios del programa? (La respuesta no debe exceder 250 palabras y no debe incluir diagramas.)

c) ¿Cuál fue el proceso de aleatorización y cuál es su principal desventaja? (La respuesta no debe exceder 100 palabras.)

d) ¿Información sobre qué tipo de localidades (rurales/urbanas) y sobre cuántas localidades hay disponible en la muestra de evaluación (en particular cuántas pertenecen al grupo control y cuántas al grupo de beneficiarias), adicionalmente, en qué entidades federativas se encuentran estas localidades? (Describir en sólo una frase.)

e) ¿Dentro de una misma localidad qué tipo de hogares es posible encontrar?

   (i) Hogares que reciben los beneficios del programa y hogares que nunca los recibirán.

   (ii) Hogares que reciben los beneficios del programa y hogares control.

   (iii) Hogares control y hogares que nunca recibirán los beneficios del programa.

   (iv) Únicamente hogares que reciben beneficios o que son del grupo control.

   (v) Únicamente hogares que nunca recibirán los beneficios del programa.

   (vi) Alguna combinación de las anteriores. (De ser este el caso, ¿qué combinación?)

   (vii) Ninguna de las anteriores. (De ser este el caso, ¿qué tipo de hogares es posible encontrar?)

2. Limpieza de la base de datos

Carga la base de datos encaseh97 disponible en el sitio del curso en tu computadora.

a) ¿Cuántas observaciones tiene la base de datos?

b) La base de datos tiene información a nivel individual. Dado que a nosotros nos interesa corroborar la aleatorización a nivel hogar y localidad, quédate únicamente con una observación por hogar. (Pista: la variable "número" identifica a cada miembro de cada hogar. Para quedarte con una observación por hogar te sugiero fuertemente que te quedes con la información del jefe del hogar, el miembro para el cual la variable "número" toma el valor 1.) ¿Cuántas observaciones tiene la base de datos ahora?

c) La base de datos es muy grande (199 variables). Para esta tarea se van a analizar únicamente 10 variables (Behrman y Todd (1999) analizan las 199 variables). Por lo tanto, quédate sólo con los códigos que identifican a los hogares (folio), a las localidades (claveofi) y a los tratados/controles (contba_1) y las 10 variables a analizar: p07, p08, p10, p11, p17, p18, p19, p20, p24 y p25. Describe la etiqueta que tiene cada variable (variable label).

d) La variable contba_1 es una dummy que toma el valor 1 si la observación pertenece a un hogar que vive en una localidad con tratamiento y 2 si corresponde a un hogar que vive en una localidad control. Para utilizar esta variable en las regresiones lineales, es conveniente recodificarla para que tome el valor 0 para denotar a los hogares que viven en localidades control. Una vez habiéndola recodificado, reporta la media de la variable.

3. Corroboración de la aleatorización a nivel HOGAR para toda la muestra

A partir de ahora, utiliza el archivo de Excel llamado "Balanceo y aleatorización" disponible en el sitio del curso dentro de la sección Tareas. Para esta parte de la tarea, llena la tabla "NIVEL HOGAR" de la pestaña "Sin ignorabilidad de tratamient0" y responde las preguntas que ahí aparecen.

Para poder comparar los resultados que se van a obtener, reporta siempre, en las celdas correspondientes, los valores p de las distintas pruebas que efectuarás. Rechaza la hipótesis nula de igualdad cuando el valor p sea igual o menor a 0.1.

a) Corrobora la igualdad de DISTRIBUCIONES de las variables p07, p08, p10, p11, p17, p18, p19, p20, p24 y p25.

(Pista: Si las variables son continuas, se aplica la prueba Kolmogorov-Smirnov; si las variables son discretas y toman más de dos valores, se aplica una prueba Ji cuadrada de Pearson; si las variables son binarias, se aplica una prueba t. Siguiendo a Behrman y Todd (1999) aplica la prueba Kolmogorov-Smirnov a las variables p07 y p08 y la prueba Ji cuadrada de Pearson a las variables restantes. La instrucción que debes darle a STATA para que realice la prueba Kolmogorov-Smirnov es ksmirnov varname, by(contba_1); para la prueba Ji cuadrada de Pearson, tab varname contba_1, chi2. Al aplicar la prueba Kolmogorov-Smirnov, reporta en la tabla de Excel el valor p ajustado; al aplicar la prueba Ji cuadrada de Pearson, reporta el valor Pr que aparece al fondo de cada tabla.)

b) Corrobora la igualdad de MEDIAS de las variables p07, p08, p10, p11, p17, p18, p19, p20, p24 y p25 utilizando una prueba t.

(Pista: La instrucción que debes darle a STATA para que realice la prueba t es ttest varname, by(contba_1). Al aplicar la prueba t reporta en la tabla de Excel el valor Pr($|T| > |t|$) que aparece al fondo de cada tabla en la columna central.)

c) Corrobora la igualdad de MEDIAS de las mismas variables del inciso anterior utilizando una regresión lineal simple.

(Pista: La instrucción que debes darle a STATA para que arroje los resultados de una regresión lineal simple es reg varname contba_1. Al correr la regresión lineal simple, reporta en la tabla de Excel el valor $P > |t|$ correspondiente a la variable contba_1 que aparece en la cuarta columna de la tabla con los resultados de la regresión.)

d) Corrobora la igualdad de MEDIAS de las mismas variables del inciso (b) utilizando una regresión lineal con errores estándar robustos.

(Pista: La instrucción que debes darle a STATA para que arroje los resultados de una regresión lineal con errores estándar robustos es reg varname contba_1, robust. Al correr la regresión lineal con errores estándar robustos, reporta en la tabla de Excel el valor P > |t| correspondiente a la variable contba_1 que aparece en la cuarta columna de la tabla con los resultados de la regresión.)

e) Corrobora la igualdad de MEDIAS de las mismas variables del inciso (b) utilizando una regresión lineal con errores estándar agrupados a nivel localidad para corregir por posibles correlaciones entre las variables de una misma localidad

(Pista: La instrucción que debes darle a STATA para que arroje los resultados de una regresión lineal con errores estándar robustos es reg varname contba_1, cluster(claveofi). Al correr la regresión lineal con errores estándar agrupados a nivel localidad, reporta en la tabla de Excel el valor P > |t| correspondiente a la variable contba_1 que aparece en la cuarta columna de la tabla con los resultados de la regresión.)

4. Corroboración de la aleatorización a nivel LOCALIDAD para toda la muestra

Para esta parte de la tarea, llena la tabla "NIVEL LOCALIDAD" de la pestaña "Sin ignorabilidad de tratamient" y responde las preguntas que ahí aparecen.

Hasta este momento, la base de datos tiene información a nivel hogar. Dado que ahora interesa corroborar la aleatorización a nivel localidad, saca los promedios de cada variable a nivel localidad y quédate únicamente con una observación por localidad. Para hacer esto, corre las siguientes instrucciones en STATA: foreach var in p07 p08 p10 p11 p17 p18 p19 p20 p24 p25 {

```
bysort claveofi: egen 'var'_local=mean('var')
}
bysort claveofi:gen num_local=_n
keep if num_local==1
```

a) Repite los incisos (a) – (d) del problema 3.

(Pista: Dado que promediaste todas las variables para trabajar a nivel localidad, ahora todas las variables son continuas. Por lo tanto, únicamente aplica la prueba Kolmogorov-Smirnov para checar la igualdad de distribuciones.)

5. Aplicación del supuesto de ignorabilidad de tratamiento y aleatorización

Para esta parte de la tarea, llena las tablas "NIVEL HOGAR" y "NIVEL LOCALIDAD" de la pestaña "Aleatorización original" y responde las preguntas que ahí aparecen

a) Vuelve a cargar la base de datos encaseh97 disponible en el sitio del curso en tu computadora y quédate nuevamente con una observación por hogar.

b) Para aleatorizar la muestra, la administración del programa utilizó el supuesto de ignorabilidad de tratamiento. Es decir, la administración aleatorizó la asignación del tratamiento únicamente entre los hogares pobres. Por lo tanto, para corroborar que la aleatorización de PROGRESA haya sido exitosa, debes eliminar a todas las observaciones de hogares no pobres. (Pista: la variable "pobre_1" identifica a los hogares pobres.) ¿Cuántas observaciones tiene la base de datos ahora?

c) Quédate únicamnte con las variables folio, claveofi, contba_1, p07, p08, p10, p11, p17, p18, p19, p20, p24 y p25 y recodifica la variable contba_1 para volverla 0-1 como en el incido (d) del problema 2.

d) Repite todos los incisos del problema 3 y del problema 4.

6. Nueva aleatorización

Para esta parte de la tarea, llena las tablas "NIVEL HOGAR" y "NIVEL LOCALIDAD" de la pestaña "Aleatorización nueva" y responde las preguntas que ahí aparecen.

Ahora vas a aleatorizar la muestra y comprobaras que tu aleatorización sea exitosa. Dado que supuestamente todas las localidades de la muestra experimental de PROGRESA son observacionalmente similares, una nueva aleatorizacion deberia arrojar tablas similares a las que has construido.

Para hacer una nueva aleatorización, empieza por cargar de nueva cuenta los datos originales de la encaseh97.

a) Dado que la aleatorización original de PROGRESA fue a nivel localidad, tú volverás a aleatorizar a nivel localidad. Para hacer esto, te vas a fijar únicamente en una observación por localidad. Una forma de lograr esto es numerando todas las observaciones que aparecen en cada localidad del 1 al Nl, donde Nl es el número máximo de observaciones en la localidad l. En particular, corre la siguiente instrucción en STATA: bysort claveofi: gen n=_n

b) Ahora, dado que todas las localidades tienen al menos una observación (si usas el comando "tab n" te daras cuenta que la variable n toma 506 veces el valor 1; es decir, eres capaz de identificar las 506 localidades de PROGRESA), genera un número aleatorio cada vez que la variable n tome el valor 1. Esto es, genera un número aleatorio para cada localidad. Para hacer esto y que todos obtengan el mismo resultado, corre las siguientes instrucciones en STATA: set seed 1 gen random=uniform() if n==1

c) Ordena los números aleatorios recién creados en orden ascendente. ¿Cuál es el menor número aleatorio que se generó y a qué número de localidad (claveofi) le corresponde ese número aleatorio?

d) Genera tu variable T de tal modo que tome el valor 1 para las primeras 320 localidades y el valor 0 para las siguientes 186. Es decir, asigna el mismo número de localidades a los grupos de tratamiento y de control como en el experimento original de PROGRESA. Para hacer esto, corre las siguientes instrucciones en STATA: gen T=1 if (_n<=320) replace T=0 if (_n>320 & _n<=506)

e) Ahora simplemente generaliza esta información para todas las observaciones dentro de cada localidad. Para hacer esto, corre la siguiente instrucción en STATA: bysort claveofi: egen tratamiento=max(T)

f) Quédate únicamente con una observación por hogar, con hogares pobres y con las variables folio, claveofi, tratamiento, p07, p08, p10, p11, p17, p18, p19, p20, p24 y p25

g) Repite todos los incisos del problema 3 y del problema 4 utilizando la variable T en lugar de la variable contba_1.