EL COLEGIO DE MÉXICO

CEE 60 AÑOS 1964-2024 CENTRO DE ESTUDIOS ECONÓMICOS

CENTRO DE ESTUDIOS ECONÓMICOS

Maestría en Economía 2024–2026

Microeconometría para la Evaluación de Programas Sociales

TAREA 1

PRESENTA: José Daniel Fuentes García

PROFESORA: Aurora Ramírez

LABORATORISTA: Mario Lechuga

El Colegio de México
Microeconometria para Programas Sociales
Jose Daniel Fuentes García
Github : danifuentesga

Índice

El Colegio de México
Microeconometria para Programas Sociales
Jose Daniel Fuentes García
Github : danifuentesga

Instrucciones

Entregables:

1. **Archivo 1:** Un PDF con **todas** las respuestas a las preguntas que aparecen aquí. Únicamente se revisarán las respuestas contenidas en este PDF. En caso de haber un error, se dará crédito parcial basado en el `log.file` y el archivo de Excel que se mencionan abajo.
2. **Archivo 2:** Un `log.file` de Stata que muestre el código utilizado para responder las preguntas 2 a 6 de los ejercicios Parte II.
3. **Archivo 3:** El archivo de Excel llamado `balanceo` debidamente llenado.

Todos estos archivos se enviarán vía el sitio de **Moodle**. No se calificarán tareas extemporáneas. Por favor asegúrense de subir **todos** los archivos antes de oprimir el botón de enviar en Moodle.

Parte I. Teoría

Pregunta 1.

El presidente de la nación A está pensando realizar una reforma educativa donde se evalúen a los maestros para mejorar su desempeño en clase. El presidente logra hacer una prueba piloto antes de lanzar la reforma y lo hace en la región norte, la región más rica del país y donde se encuentran los mejores maestros. Para ver si la evaluación tuvo efecto compara las calificaciones de los estudiantes de esta región rica **después del tratamiento** con el promedio de las calificaciones del resto de los estudiantes.

¿Habría un sesgo de selección en su estimación del efecto al realizar esta comparación? ¿Por qué? ¿Estaría sobreestimado o subestimado el efecto?

RESPUESTA:

Sea $d_i = 1$ si el alumno está en la **región norte** (piloto) y $d_i = 0$ si está en el resto (control); y_i es la calificación observada post-tratamiento. Siguiendo a Angrist & Pischke, la diferencia se descompone en:

$$\underbrace{E[y_i \mid d_i = 1] - E[y_i \mid d_i = 0]}_{\text{diferencia observada}} = \underbrace{\tau}_{\text{efecto causal}} + \underbrace{E[\varepsilon_i \mid d_i = 1] - E[\varepsilon_i \mid d_i = 0]}_{\text{sesgo de selección}},$$

donde ε_i recoge el componente asociado al resultado potencial sin programa (y_{0i}). En este caso, como el **norte ya era más rico y con mejores maestros antes** del piloto, razonablemente $E[y_{0i} \mid d_i = 1] > E[y_{0i} \mid d_i = 0]$, por lo que el término de selección es **positivo** y la comparación $E[y|d{=}1] - E[y|d{=}0]$ **sobreestima** τ (Angrist & Pischke, cap. 2).

En conclusión: Sí hay sesgo de selección y es positivo. La estimación está **sobreestimada** porque el grupo tratado ya tenía mejores condiciones iniciales. Esto ocurre porque los grupos no estaban balanceados antes del tratamiento, y la diferencia no refleja únicamente el efecto del programa (Ravallion, 2001)

Jose Daniel Fuentes García
Github : danifuentesga

Parte II. Aplicaciones

Las siguientes preguntas tienen como objetivo 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. Para realizar esta tarea utiliza la primera línea basal del programa **PROGRESA** (`encaseh97`) que está disponible en este sitio.

Para que te familiarices con el programa PROGRESA y comprendas el ejercicio práctico, comienza leyendo la nota metodológica del programa (*General Rural Methodology.pdf*). 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.

Pregunta 1.

Basado en la información de la nota metodológica, responde si las siguientes afirmaciones son verdaderas o falsas. Si la afirmación es falsa, justifica tu respuesta.

a)

La ENCASEH fue un censo a través del cual se obtuvieron los datos socioeconómicos y demográficos pertinentes para identificar a los hogares que se beneficiarían del Progresa.
R:VERDADERO

b)

El PROGRESA focalizó su apoyo en tres etapas consecutivas: 1) Con base a información del Censo 1990 y del Conteo de 1995 se localizaron las localidades más marginadas del país y se verificó su acceso a drenaje y agua 2) Una vez identificadas se aplicó un censo a cada hogar 3) Dentro de cada localidad se llevaron a cabo asambleas en cada localidad para presentarles a los habitantes la lista de los hogares elegibles.
R:FALSO

Es falso porque, según la metodologia las tres fases fueron: 1) focalización geográfica para identificar las localidades con mayor marginación y verificar el acceso a servicios básicos de educación y salud, 2) levantamiento de un censo socioeconómico en los hogares seleccionados, y 3) presentación de resultados en asamblea comunitaria para recibir sugerencias y correcciones. No se trata del acceso a drenaje y agua en lugar de educación y salud, e indica erróneamente el año del conteo como 1995 cuando en realidad fue 1996.

c)

La aleatorización de PROGRESA se llevó a cabo a nivel de hogar.
R:FALSO

Es falso porquela aleatorización de Oportunidades (PROGRESA) en zonas rurales se realizó a nivel de localidad, no de hogar. Se seleccionaron un grupo de localidades que

cumplían con los criterios de inclusión, asignarlas aleatoriamente a los grupos de intervención o control y, posteriormente, dentro de las localidades de intervención, identificar los hogares elegibles mediante el ENCASEH 97. La unidad de aleatorización fue la localidad completa y no los hogares de manera individual donde si había diferencias relevantes.

d)

Del universo de localidades seleccionadas como marginales, 506 fueron escogidas para tomar parte en la muestra de evaluación del programa: 320 en el grupo de control y 186 en el grupo tratado.
R:FALSO

La muestra original de evaluación en áreas rurales incluyó 506 localidades, distribuidas en 320 localidades de tratadas y 186 localidades de control, no al revés. Por lo tanto, la afirmación intercambia las cifras de cada grupo y es falsa

e)

Para identificar el grado de pobreza en los hogares se utiliza una técnica estadística multivariable que incluye el ingreso del hogar y otras variables.
R:VERDADERO

f)

Hay dos tipos de localidades dentro del universo de localidades seleccionadas como marginales: 1) aquellas con hogares que reciben los beneficios del programa y hogares que nunca los recibirán. 2) aquellas con hogares que reciben los beneficios del programa y hogares control.
R:FALSO

Las localidades seleccionadas para la evaluación se clasificaban principalmente en localidades de intervención (donde los hogares reciben los beneficios) y localidades de control (donde los hogares no reciben beneficios inicialmente). No existía un tipo de localidad donde hubiera mezcla de hogares beneficiarios y hogares que "nunca" los recibirían, eventualmente el Programa llegaría a a todos.

Pregunta 2. Limpieza de bases de datos

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

Todo el proceso llevado a cabo también se puede encontrar en danifuentesga.

a)

¿Cuántas variables y cuántas observaciones tiene la base de datos?
R: **125,674** observaciones tiene la base y el número de variables es **199**.

b)

De acuerdo a la nota metodológica ¿Cuál es la variable que identifica el hogar? ¿Cuál es la variable que identifica al individuo?
R: De acuerdo con la nota metodológica, la variable que identifica el hogar es `folio` y la variable que identifica al individuo es `renglón`.

Como se menciona en la nota metodológica, la base de datos tiene información a nivel individual. Dado que nos interesa corroborar la aleatorización a nivel hogar y localidad, quédate únicamente con la observación correspondiente al jefe del cada hogar (Pista: la variable "numero" es el número de renglón o renglón que menciona la nota metodológica y "1" es el valor asignado al jefe del hogar.)

c)

¿Cuántos jefes de familia u hogares hay ahora en la base de datos?
R: Después de filtrar la variable `numero = 1` que corresponde al jefe de hogar, quedan un total de **24,077** observaciones, lo cual también corresponde al número de hogares.

Únicamente analizaremos 10 variables (Behrman y Todd, 1999 analizan 199 variables). Por lo tanto, quédate sólo con los códigos que identifican a los hogares, a las localidades (`claveofi`) y a los tratados/controles (`contba_1`) y las siguientes 10 variables por analizar:

Variable	Etiqueta
p08	edad
p11	sexo
p17	habla también español
p18	alfabetismo
p19	asistencia a la escuela
p20	nivel de escolaridad
p24	condición de inactividad
p25	posición en la ocupación
p38	ingreso del hogar
p65b	tiene refrigerador

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.

d)

¿Qué porcentaje de los hogares pertenece al grupo de tratamiento en la muestra?
Pista: Para utilizar esta variable en las tabulaciones y regresiones lineales, recodifícala para que tome el valor 0 en los hogares que viven en localidades control.
R: En la muestra total, **38.30 %** de los hogares pertenecen al *grupo control*, mientras que el **61.70 %** restante pertenece al *grupo de tratamiento*. En términos absolutos, esto equivale a **9,221** hogares en el grupo control y **14,856** en el grupo tratado.

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

A partir de ahora, utiliza el archivo de Excel llamado "Balanceo" disponible en el sitio del curso. Para esta parte de la tarea, llena la tabla "NIVEL HOGAR" de la pestaña "Sin_ignorabilidad" y responde las preguntas que ahí aparecen (en la tabla, registra los valores p). **Además copia en tu pdf de respuestas todas las tablas debidamente llenadas y todas las preguntas y respuestas a las preguntas que ahí aparecen.**

Para poder comparar los resultados que se vas a obtener reporta siempre en las celdas correspondientes los valores p de las pruebas que realices, y rechaza la hipótesis nula si obtienes un nivel de significancia menor al 10 %.

a)

Corrobora la igualdad de DISTRIBUCIONES de hogares en el grupo de control y en el grupo de tratamiento de las 10 variables elegidas. (Pista: Si las variables son continuas, aplica la prueba Kolmogorov-Smirnov; si las variables son discretas y toman más de dos valores, aplica una prueba Ji cuadrada de Pearson. Siguiendo a Behrman y Todd (1999) aplica la misma prueba que ellos hayan aplicado a las variables en la tabla 6(f) columna 2.)

Tabla 1: **sin ignorabilidad de tratamiento: nivel hogar**

Variable	**Chi2 ó ksmirnov**
p08: ¿Cuántos años tiene el jefe del hogar?	0.074
p11: Sexo	0.845
p17: ¿Habla (nombre) también el español?	0.119
p18: ¿Sabe (nombre) leer y escribir?	0.004
p19: ¿Fue (nombre) a la escuela?	0.014
p20: Nivel de escolaridad	0.026
p24: ¿Por qué no estudió/trabajó (nombre) la semana pasada?	0.000
p25: ¿Qué hace (nombre) en su trabajo?	0.000
p38: ¿Tiene ingresos el hogar?	0.215
p65b: ¿Tiene refrigerador?	0.000

b)

Corrobora la igualdad de MEDIAS de hogares en el grupo de control y en el grupo de tratamiento de las 10 variables utilizando una prueba t. (Comando en STATA: ttest)

Tabla 2: **Sin ignorabilidad de tratamiento: Nivel Hogar**

Variable	**ttest**
p08: ¿Cuántos años tiene el jefe del hogar?	0.080
p11: Sexo	0.693
p17: ¿Habla (nombre) también el español?	0.243
p18: ¿Sabe (nombre) leer y escribir?	0.008
p19: ¿Fue (nombre) a la escuela?	0.011
p20: Nivel de escolaridad	0.522
p24: ¿Por qué no estudió/trabajó (nombre) la semana pasada?	0.004
p25: ¿Qué hace (nombre) en su trabajo?	0.001
p38: ¿Tiene ingresos el hogar?	0.214
p65b: ¿Tiene refrigerador?	0.000

c)

Corrobora la igualdad de MEDIAS de hogares en el grupo de control y en el grupo de tratamiento de las mismas variables del inciso anterior utilizando una regresión lineal simple. (Comando en STATA: reg variable contba_1)

Tabla 3: **Sin ignorabilidad de tratamiento: Nivel Hogar**

Variable	**reg**
p08: ¿Cuántos años tiene el jefe del hogar?	0.080
p11: Sexo	0.693
p17: ¿Habla (nombre) también el español?	0.243
p18: ¿Sabe (nombre) leer y escribir?	0.008
p19: ¿Fue (nombre) a la escuela?	0.011
p20: Nivel de escolaridad	0.522
p24: ¿Por qué no estudió/trabajó (nombre) la semana pasada?	0.005
p25: ¿Qué hace (nombre) en su trabajo?	0.001
p38: ¿Tiene ingresos el hogar?	0.215
p65b: ¿Tiene refrigerador?	0.000

d)

Corrobora la igualdad de MEDIAS de las mismas variables del inciso (b) utilizando una regresión lineal con errores estándar robustos. (Comando en STATA: reg variable contba_1, robust)

Tabla 4: **Tabla 4:** Sin ignorabilidad de tratamiento: Nivel Hogar

Variable	**reg, robust**
p08: ¿Cuántos años tiene el jefe del hogar?	0.080
p11: Sexo	0.692
p17: ¿Habla (nombre) también el español?	0.240
p18: ¿Sabe (nombre) leer y escribir?	0.008
p19: ¿Fue (nombre) a la escuela?	0.011
p20: Nivel de escolaridad	0.519
p24: ¿Por qué no estudió/trabajó (nombre) la semana pasada?	0.005
p25: ¿Qué hace (nombre) en su trabajo?	0.001
p38: ¿Tiene ingresos el hogar?	0.205
p65b: ¿Tiene refrigerador?	0.000

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. (Comando en STATA: reg variable contba_1, vce(cluster claveofi))

Tabla 5: **Sin ignorabilidad de tratamiento: Nivel Hogar**

Variable	**reg, cluster**
p08: ¿Cuántos años tiene el jefe del hogar?	0.404
p11: Sexo	0.787
p17: ¿Habla (nombre) también el español?	0.849
p18: ¿Sabe (nombre) leer y escribir?	0.329
p19: ¿Fue (nombre) a la escuela?	0.267
p20: Nivel de escolaridad	0.815
p24: ¿Por qué no estudió/trabajó (nombre) la semana pasada?	0.115
p25: ¿Qué hace (nombre) en su trabajo?	0.107
p38: ¿Tiene ingresos el hogar?	0.479
p65b: ¿Tiene refrigerador?	0.232

Balanceo nivel HOGAR

¿Cuántas variables no están correctamente balanceadas entre los dos grupos de acuerdo con la prueba Kolmogorov-Smirnov?
RESPUESTA Usando Kolmogorov-Smirnov para variables continuas (**p08**, **p20**) y Chi² para categóricas (**p11**, **p17**, **p18**, **p19**, **p24**, **p25**, **p38**, **p65b**), se identificó que **6 variables** no están correctamente balanceadas: **p08**, **p18**, **p19**, **p20**, **p24** y **p25**. En otras palabras, los grupos difieren significativamente en **edad del jefe del hogar**, **alfabetismo**, **asistencia escolar**, **nivel de escolaridad**, **condición de inactividad** y **posición en la ocupación**.

¿Cuántas variables no están correctamente balanceadas entre los dos grupos de acuerdo con la prueba t?
RESPUESTA Según la prueba t (criterio **p <0.10**), **6 variables** no están correctamente balanceadas: **p08**, **p18**, **p19**, **p24**, **p25** y **p65b**. En términos descriptivos, los grupos difieren significativamente en **edad del jefe del hogar**, **alfabetismo**, **asistencia escolar**, **condición de inactividad**, **posición en la ocupación** y **tenencia de refrigerador**.

¿Cuántas variables no están correctamente balanceadas entre los dos grupos de acuerdo con la regresión lineal?
RESPUESTA Exactamente las mismas que la prueba t

¿Cuál es la relación de los resultados de la prueba t y de la regresión lineal?
RESPUESTA En este análisis, la **prueba t** y la **regresión lineal simple** dan los mismos **valores p** porque ambas comparan las **medias** de cada característica del hogar entre el **grupo control** y el **grupo basal (contba_1)**. Esto ocurre porque **contba_1** es una **variable binaria**, y en este caso ambos métodos son **matemáticamente equivalentes**. Por lo tanto, cualquier variable que no esté **balanceada** según la **prueba t** tampoco lo estará según la **regresión lineal**.

¿Cuántas variables no están correctamente balanceadas entre los dos grupos de acuerdo con la regresión lineal con errores estándar robustos?
RESPUESTA 5 variables no están correctamente balanceadas (p <0.10) con la regresión lineal con errores estándar robustos: p18, p19, p24, p25 y p65b

¿Cuántas variables no están correctamente balanceadas entre los dos grupos de acuerdo con la regresión lineal agrupando a nivel localidad?
RESPUESTA Ninguna variable presenta un **valor p** menor a **0.10**, por lo que ninguna está **incorrectamente balanceada** entre los grupos cuando se usa la **regresión lineal agrupando por localidad**. Esto sugiere que, al considerar la **correlación intralocalidad**, desaparecen las **diferencias** que antes parecían **significativas**.

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

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:

```
collapse (mean) p08 p11 p17 p18 p19 p20 p24 p25 p38 p65b (max) contba_1, by(claveofi)
```

a)

Repite los incisos (a) – (d) del problema 3. (Nota: ahora en lugar de comparar hogares en el grupo de control y tratamiento compararás localidades control versus localidades tratadas.) (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.)

Balanceo nivel LOCALIDAD

Tabla 6: **Sin ignorabilidad de tratamiento: Nivel Localidad**

Variable	KS	ttest	reg	reg, robust
p08	0.658	0.765	0.765	0.768
p11	0.913	0.860	0.860	0.863
p17	0.854	0.557	0.557	0.558
p18	0.857	0.667	0.667	0.671
p19	0.772	0.623	0.623	0.626
p20	0.665	0.967	0.967	0.967
p24	0.164	0.125	0.125	0.134
p25	0.758	0.479	0.479	0.475
p38	0.783	0.600	0.600	0.622
p65b	0.263	0.321	0.321	0.312

¿Cuántas variables no están correctamente balanceadas entre los dos grupos de acuerdo con la prueba Kolmogorov-Smirnov?
RESPUESTA Ninguna variable presenta diferencias significativas (todas con p >0.10), por lo que todas están balanceadas a nivel localidad.

¿Cuántas variables no están correctamente balanceadas entre los dos grupos de acuerdo con la prueba t / regresión lineal?
RESPUESTA Ninguna variable presenta diferencias significativas; los resultados son idénticos porque ambas pruebas comparan medias.

¿Cuántas variables no están correctamente balanceadas entre los dos grupos de acuerdo con la regresión lineal con errores estándar robustos?
RESPUESTA Tampoco hay variables con diferencias significativas; los errores robustos no cambian la conclusión porque la varianza entre localidades no genera sesgo.

¿Por qué no hace sentido correr una regresión lineal agrupando los errores a nivel localidad?
RESPUESTA No tiene sentido porque la base ya está çolapsada"por localidad; no existen más niveles dentro de cada cluster para corregir correlación interna.

¿Hace sentido hablar de una buena o mala aleatorización en este caso?
RESPUESTA Los resultados indican buena aleatorización, ya que no hay diferencias significativas entre grupo control y grupo tratamiento en ninguna de las variables

Pregunta 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. Para aleatorizar la muestra, la administración de PROGRESA 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?
RESPUESTA: Ahora la base tiene **12,519** observaciones

b)

Quédate únicamente con las variables folio, claveofi, contba_1, p08, p11, p17, p18, p19, p20, p24, p25, p38 y p65b y recodifica la variable contba_1 para volverla 0–1 como en el inciso (d) del problema 2. Repite todos los incisos del problema 3 y del problema 4.

Aleatorización Original : Nivel HOGAR

Tabla 7: Aleatorización Original: Nivel Hogar

Variable	KS / Chi2	ttest	reg	reg_robust	reg_cluster
p08	0.410	0.291	0.291	0.291	0.492
p11	0.516	0.392	0.392	0.397	0.489
p17	0.085	0.205	0.205	0.205	0.795
p18	0.807	0.722	0.722	0.724	0.884
p19	0.935	0.739	0.739	0.740	0.875
p20	0.405	0.073	0.073	0.071	0.483
p24	0.003	0.004	0.004	0.004	0.048
p25	0.000	0.052	0.052	0.079	0.326
p38	0.963	0.963	0.963	0.963	0.983
p65b	0.472	0.587	0.587	0.582	0.782

¿Cuántas variables no están correctamente balanceadas entre los dos grupos de acuerdo con la prueba Kolmogorov-Smirnov?
RESPUESTA: 3 variables no están balanceadas (p17, p24, p25)

¿Cuántas variables no están correctamente balanceadas entre los dos grupos de acuerdo con la prueba t?
RESPUESTA: 2 variables no están balanceadas (p20, p24)

¿Cuántas variables no están correctamente balanceadas entre los dos grupos de acuerdo con la regresión lineal?
RESPUESTA: Las mismas 2 variables que en la prueba t (p20, p24)

¿Cuál es la relación de los resultados de la prueba t y de la regresión lineal?
RESPUESTA: Los p-values son idénticos porque ambas pruebas contrastan la misma hipótesis de igualdad de medias, solo que por métodos distintos.

¿Cuántas variables no están correctamente balanceadas entre los dos grupos de acuerdo con la regresión lineal con errores estándar robustos?
RESPUESTA: Las mismas 2 variables que en la prueba t y en la regresión (p20, p24)

¿Cuántas variables no están correctamente balanceadas entre los dos grupos de acuerdo con la regresión lineal agrupando a nivel localidad?
RESPUESTA: Solo 1 variable no está balanceada (p24), lo que indica que al controlar por correlación intra-localidad la evidencia de desequilibrio disminuye.

Aleatorización Original : Nivel LOCALIDAD

Tabla 8: Aleatorización Original: Nivel Localidad

Variable	ksmirnov	ttest	reg	reg_r
p08	0.480	0.799	0.799	0.801
p11	0.979	0.675	0.675	0.671
p17	0.750	0.419	0.419	0.422
p18	0.780	0.522	0.522	0.531
p19	0.927	0.871	0.871	0.868
p20	0.199	0.475	0.475	0.477
p24	0.062	0.316	0.316	0.302
p25	0.760	0.955	0.955	0.955
p38	0.961	0.410	0.410	0.404
p65b	0.522	0.866	0.866	0.863

¿Cuántas variables no están correctamente balanceadas entre los dos grupos de acuerdo con la prueba Kolmogorov-Smirnov?
RESPUESTA: 1 variable (p24) presenta p <0.10, indicando posible falta de balance.

¿Cuántas variables no están correctamente balanceadas entre los dos grupos de acuerdo con la prueba t / regresión lineal?
RESPUESTA: Ninguna variable presenta p <0.10, por lo que no hay evidencia de falta de balance.

¿Cuántas variables no están correctamente balanceadas entre los dos grupos de acuerdo con la regresión lineal con errores estándar robustos?
RESPUESTA: Ninguna variable presenta p <0.10, igual que en la prueba t y la regresión lineal.

¿Por qué no hace sentido correr una regresión lineal agrupando los errores a nivel localidad?
RESPUESTA: No tiene sentido porque ya estamos a nivel localidad, y el agrupamiento no aporta información adicional en esta unidad de análisis.

¿Hace sentido hablar de una buena o mala aleatorización en este caso?
RESPUESTA: Sí, parece razonable hablar de buena aleatorización porque la gran mayoría de las variables tienen p-valores mayores a 0.10 y no hay patrones sistemáticos de desequilibrio.

Pregunta 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 comprobarás que tu aleatorización sea exitosa. Dado que supuestamente todas las localidades de la muestra experimental de PROGRESA son

observacionalmente similares, una nueva aleatorización debería 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 aleatorizarás a nivel localidad. Para hacer esto, te vas a fijar únicamente en una observación por localidad: *(Comando en STATA:*

```
egen tag_local=tag(claveofi)
recode tag_local (0=.)
```

)

b)

Ahora, que eres capaz de identificar las 506 localidades de PROGRESA, genera un número aleatorio para cada localidad. Para hacer esto y que todos obtengan el mismo resultado, corre las siguientes instrucciones en STATA:
`(Comando en STATA: set seed 1`
`gen random=uniform() if tag_local==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?
RESPUESTA: El menor número aleatorio fue 0,000608 y corresponde a la localidad con claveofi $1{,}303 \times 10^8$.

d)

Genera una variable que indique quién recibirá el tratamiento llamada `trat` de tal modo que tome el valor 1 para las primeras 320 localidades y el valor 0 para las siguientes 186. Para hacer esto, corre las siguientes instrucciones en STATA:

```
egen n=group(random)
gen trat=1 if (n<=320)
replace trat=0 if (n>320 & n<=506)
sort claveofi n
replace trat=trat[_n-1] if trat[_n-1] != . & claveofi==claveofi[_n-1]
```

e)

Quédate únicamente con una observación por hogar, con hogares pobres y con las variables `folio`, `claveofi`, `trat`, `p08`, `p11`, `p17`, `p18`, `p19`, `p20`, `p24`, `p25`, `p38`, y `p65b`.

f)

Repite todos los incisos del problema 3 y del problema 4 utilizando la variable `trat` en lugar de la variable `contba_1`.

Aleatorización Nueva : Nivel HOGAR

Tabla 9: **Aleatorización Nueva: Nivel Hogar**

Variable	KS	ttest	reg	reg, robust	reg, cluster
p08	0.000	0.015	0.016	0.015	0.207
p11	0.028	0.039	0.039	0.056	0.130
p17	0.000	0.005	0.005	0.006	0.045
p18	0.000	0.002	0.002	0.002	0.349
p19	0.000	0.000	0.000	0.000	0.022
p20	0.000	0.003	0.003	0.003	0.151
p24	0.004	0.004	0.004	0.004	0.011
p25	0.000	0.000	0.000	0.000	0.006
p38	0.034	0.033	0.034	0.046	0.304
p65b	0.003	0.081	0.081	0.056	0.350

¿Cuántas variables no están correctamente balanceadas entre los dos grupos de acuerdo con la prueba Kolmogorov-Smirnov?
RESPUESTA: 10 variables no están balanceadas (todas tienen p <0.1).

¿Cuántas variables no están correctamente balanceadas entre los dos grupos de acuerdo con la prueba t / regresión lineal?
RESPUESTA: 9 variables no están balanceadas (todas excepto p65b).

¿Cuántas variables no están correctamente balanceadas entre los dos grupos de acuerdo con la regresión lineal con errores estándar robustos?
RESPUESTA: 9 variables no están balanceadas (todas excepto p65b).

¿Cuántas variables no están correctamente balanceadas entre los dos grupos de acuerdo con la regresión lineal agrupando a nivel localidad?
RESPUESTA: 9 variables no están balanceadas (todas excepto p65b).

Aleatorización Nueva : Nivel LOCALIDAD

Tabla 10: **Aleatorización Nueva: Nivel Localidad**

Variable	ksmirnov	ttest	reg	reg, robust
p08	0.255	0.210	0.210	0.217
p11	0.660	0.414	0.414	0.421
p17	0.938	0.453	0.453	0.445
p18	0.398	0.094	0.094	0.079
p19	0.151	0.013	0.013	0.008
p20	0.131	0.064	0.064	0.062
p24	0.142	0.132	0.132	0.132
p25	0.004	0.006	0.006	0.003
p38	1.000	0.968	0.968	0.969
p65b	0.607	0.706	0.706	0.692

¿Cuántas variables no están correctamente balanceadas entre los dos grupos de acuerdo con la prueba Kolmogorov-Smirnov?
RESPUESTA: 1 variable (p25) presenta p <0.1, lo que indica posible falta de balance según la prueba KS.

¿Cuántas variables no están correctamente balanceadas entre los dos grupos de acuerdo con la prueba t / regresión lineal?
RESPUESTA: 4 variables (p18, p19, p20 y p25) presentan p <0.1 en ambas pruebas, por lo que no están correctamente balanceadas.

¿Cuántas variables no están correctamente balanceadas entre los dos grupos de acuerdo con la regresión lineal con errores estándar robustos?
RESPUESTA: 4 variables (p18, p19, p20 y p25) no están balanceadas según la regresión robusta (p <0.1).

¿Qué tan exitosa fue nuestra nueva aleatorización y qué significa esto?
RESPUESTA: La nueva aleatorización a nivel localidad fue razonablemente exitosa, ya que la mayoría de las variables están balanceadas. Sin embargo, se identifican 4 variables con diferencias significativas según pruebas t y regresión, lo que sugiere que aún existen algunas diferencias entre tratamiento y control. Aun así, estas diferencias son limitadas y no invalidan el diseño experimental.