Profesor: Manuel Fernández

- David Arboleda
- Valentina Castilla
- Valentina Daza
- Santiago Torres.

Taller 2

Econometría Avanzada, 2022-2
Fecha de Entrega: 14 de octubre.

Universidad de los Andes

Primer ejercicio

El Gobierno Nacional ha trabajado activamente para hacer seguimiento a la situación nutricional de los niños y adolescentes del país. Por ello, realizaron la tercera Encuesta Nacional de Situación Nutricional de Colombia (ENSIN), la cual le sigue el rastro a políticas públicas nacionales y territoriales en materia de salud, alimentación y nutrición. Esta encuesta fue dirigida a colegios puesto que varios estudios señalan que es durante la fase escolar donde se experimenta el mayor periodo de crecimiento corporal, consolidación de gustos, hábitos alimenticios y adaptación a la alimentación de adulto. Para el caso colombiano, según la encuesta ENSIN, 7 de cada 100 menores en edad escolar presentan desnutrición crónica. En los indígenas y comunidades afro, 30 de cada 100 menores presentan este problema, mientras que esta situación se extiende a 11 de cada 100 niños de los hogares más pobres del país.

En consecuencia, el Gobierno Nacional lanzó un programa de educación alimentaria en donde profesionales de nutrición otorgan sesiones de capacitación nutricional a las familias de estudiantes de tercer grado escolar. Durante el trabajo de campo realizado en 2015, se capacitaron a los padres de 50,639 estudiantes de 1,648 colegios en 295 municipios del país. Todos los padres que fueron elegidos recibieron el tratamiento. La recolección de información nutricional de los niños tuvo lugar tanto antes como después de recibir la capacitación.

Para analizar el impacto que tuvo el programa sobre el tallaje de los niños, se contrató a una consultora para que diseñara un experimento aleatorio controlado. Dado que la experimentadora vio la clase de Econometría Avanzada, ella conoce que una distribución aleatoria del tratamiento no es suficiente para recuperar un efecto causal en un RCT: ella es consciente de la presencia de posibles efectos psicológicos y sociales que pueden sesgar el impacto real de la capacitación nutricional. Aprovechando que cuenta con una muestra relativamente grande de la población, su diseño incorpora 5 experimentos diferentes que se llevan a cabo en subgrupos (disjuntos) adecuadamente aleatorizados de la muestra.

1. **Subgrupo 1** En este grupo los estudiantes fueron asignados de forma aleatoria a las capacitaciones nutricionales. Sin embargo, se hizo la asignación en dos etapas, las cuales usaron aleatorizaron en difeentes niveles:

    - **Etapa 1: Nivel colegio.** Se capacitaron a todos los estudiantes de colegios aleatoriamente seleccionados.
    - **Etapa 2: Nivel estudiante.** Se realizó una asignación aleatoria estratificada por colegio, es decir se seleccionaron aleatoriamente a colegios y posteriormente se realizo una asignación aleatoria de los estudiantes dentro de los colegios seleccionados.

2. **Subgrupo 2** A colegios elegidos de forma aleatoria se les otorgó una capacitación sobre bullying, mientras que a los otros no se les asignó ningún tipo de capacitación.

3. **Subgrupo 3** En este grupo todos los colegios recibirán la capacitación, pero con distintos grados de información. Por un lado, de forma aleatoria, a algunos de ellos se les informa desde el primer momento de la capacitación que serán evaluados con un test final. Por otro lado, aunque los otros colegios también reciben la capacitación, en ellos no se realiza ningún tipo de test de conocimientos y los padres son informados de ello.

4. **Subgrupo 4** De forma aleatoria se seleccionan algunos colegios, los cuales ninguno recibirá la capacitación. Sin embargo, a algunos de ellos se les informa de su estatus de tratamiento, es decir, se les informa alos padres que otros colegios recibieron la capacitación nutricional pero ellos no. A los otros colegios de este grupo no se les comenta acerca de su estatus como grupo de control, de manera que no saben que no se encuentran recibiendo la capacitación ni de la existencia del programa.

5. **Subgrupo 5** De forma aleatoria se seleccionan colegios donde se les comentarán mensajes distintos a los padres acerca de la importancia de la nutrición: a algunos padres se les dará la información A: “Aunque

varios estudios muestran que la alimentación es clave para la formación del niño, otros estudios muestran que otros factores como el número de profesores y la calidad del colegio son aún mas relevantes.” . En los demás colegios, los padres recibieran la información B: “Una deficiencia alimenticia a la edad preescolar, reduce en forma permanente la destreza de los niños.”

Después de efectuados los experimentos, el Gobierno les pasa la base de datos “Base_ENSIN.dta” que fue entregada por el contratista para que ustedes hagan la evaluación de impacto. Esta base contiene información a nivel de estudiante con las siguientes variables:

- *ID_col:* Código ENSIN de la sede educativa.
- *ID_estudiante:* Código ENSIN del estudiante.
- *Talla_niño:* Variable que mide el tallaje estandarizado de niño posterior a la implementación de la capacitación.
- *Grupo_1:* Dicótoma que toma el valor de uno si el estudiante/colegio recibió el tratamiento y cero de lo contrario. Esta incorpora tanto los tratados y controles de la primera y segunda etapa.
- *Grupo_1_E1:* Dicótoma que toma el valor de uno si el estudiante recibió el tratamiento en la etapa 1 y cero de lo contrario.
- *Grupo_1_E2:* Dicótoma que toma el valor de uno si el estudiante recibió el tratamiento en la etapa 2 y cero de lo contrario.
- *Grupo_2:* 1 si recibió capacitación bullying, 0 de lo contrario
- *Grupo_3:* 1 si fue evaluado posterior a la capacitación, 0 de lo contrario
- *Grupo_4:* 1 si recibió información de estado de control, 0 de lo contrario
- *Grupo_5:* 1 si recibió información A, y 0 si recibió información B.
- *Nivel_aleatorizacion:* Categórica, 1 si fue aleatorizado a nivel colegio y 2 si fue asignado a nivel estudiante.
- Controles en linea base: *Educacion_jefe*, *Ingreso_jefe*, *Personas_hogar*, *Ocupado_jefe*, *Raza_afro_indig* y *Sexo*

  En la siguiente tabla se encuentra información más detallada sobre los diferentes subgrupos de la muestra usados para evaluar diferentes tratamientos.

Tabla 1: Información de los grupos de aleatorización

Muestra	Submuestra	% de la muestra	Grupo de efecto	Etapa/nivel de aleatorización	Submuestra de etapas	% de la submuestra de etapas
50639 [1649]	23039 [542]	45.49 % [32.82 %]	Subgrupo 1	Etapa 1	15388 [402]	66.79 % [74.16 %]
				Etapa 2	7651 [140]	33.20 % [25.83 %]
	6900 [277]	13.62 % [16.79 %]	Subgrupo 2	Nivel colegio		
	6900 [277]	13.62 % [16.79 %]	Subgrupo 3	Nivel colegio		
	6900 [277]	13.62 % [16.79 %]	Subgrupo 4	Nivel colegio		
	6900 [276]	13.62 % [16.73 %]	Subgrupo 5	Nivel colegio		

- Nota: Entre corchetes se encuentra la información de nivel colegio y sin corchetes representa la información a nivel estudiante. Esto es clave para entender e interpretar los diferentes niveles de aleatorización.

Respondan las siguientes preguntas teniendo en cuenta el contexto y los datos disponibles.[1]

a) i) Propongan uno o varios modelos de regresión que recuperen el efecto de recibir el tratamiento sobre el tallaje. Asegúrense de describir los términos que componen sus ecuaciones a estimar (variables y parámetros), así como el o los subgrupos que usarán para estimar el efecto. Mencionen y discutan los supuestos necesarios para que su estimador sea consistente. ¿Son los supuestos de identificación del efecto causal teóricamente plausibles en este contexto?

ii) Suponiendo que los efectos de identificación se cumplen, estimen los modelos propuestos. Presenten sus resultados en una tabla. Interpreten el efecto encontrado, y discutan su significancia económica y estadística.

[1] El experimento descrito y los datos asociados a este son ficticios. Los resultados obtenidos en este taller no proveen información alguna sobre si capacitar o no a los padres incrementa el tallaje de los niños.

b) Para verificar la validez de la estimación, se les solicita evidencia que soporte la correcta aleatorización de los colegios en el grupo de tratamiento a nivel colegio y a nivel estudiante. Para esto presenten:

I. Una tabla de balance muestral para el subgrupo 1.
II. Una tabla de balance muestral para la etapa 2 del subgrupo 1.
III. Una tabla de balance muestral para el grupo de aleatorización a nivel estudiante para la etapa 1 del subgrupo 1.

Algunas consideraciones para verificar el balance:

- Deben tener en cuenta el tipo de aleatorización que se hizo (i.e., si fue estratificada o no).
- Piensen en la forma correcta de calcular los errores estándar.

Finalmente, a partir de las tablas, argumenten: ¿hay evidencia de que la aleatorización se hizo de forma correcta? ¿Para qué grupos parece cumplirse el balance en línea base?

Pista: Para hacer pruebas de balance muestral condicionales, pueden recurrir al concepto de *partialling out* o pruebas de diferencias de medias condicionales.

c) Los funcionarios del Gobierno les pidan ahora que realicen un procedimiento que les permita abogar a favor o en contra de la validez del SUTVA. Para ello requieren que:

I. Establezcan una relación matemática (en términos de los resultados potenciales) que represente el cumplimiento del supuesto SUTVA en la muestra.
II. Describan una comparación o ecuación a estimar basada en los datos y subgrupos disponibles que les permita verificar el supuesto de SUTVA.
III. Implementen y presenten en una tabla la comparación/metodología planteada en el inciso II.

A partir de los resultados encontrados, ¿parece cumplirse el supuesto de SUTVA? En caso de que no se cumpla, ¿para qué grupos parece no cumplirse?

d) Sumado a lo anterior, pueden existir otros tipos de sesgos comportamentales producto del experimento que contaminarían los resultados de su estimación. El Gobierno quiere cerciorarse que ninguno de ellos esté presente. Por eso les solicita que estudie la presencia de los siguientes efectos comportamentales:

i) Efecto placebo.
ii) Efecto Hawthorne.
iii) Efecto John Henry.
iv) Efecto “experimenter demand”.

Utilizando los diferentes subgrupos, hagan los siguientes ejercicios para determinar la presencia de cada efecto en la muestra:

I. Establezcan una relación matemática para cada uno de los efectos comportamentales (en términos de los resultados potenciales) que representen la presencia del efecto.
II. A partir de su respuesta en el inciso I, describan el experimento ideal que les permitiría verificar el efecto existe.
III. Implementen y presenten en una tabla la comparación/metodología planteada en el inciso II. ¿Es razonable suponer o no la existencia de dicho efecto?

e) Con base en la información encontrada en los incisos $b)$, $c)$ y d, ¿la estimación realizada en el inciso $a)$ recupera el efecto causal del programa sobre el tallaje?. En caso de que su respuesta no sea afirmativa, ¿cómo rediseñarían el experimento de manera que se mitiguen las posibles fuentes de contaminación?

Segundo ejercicio

Card (2009) investiga la relación que existe entre la inmigración y la desigualdad salarial. Para lograr este objetivo, el ahora premio Nobel estudió cómo distintos flujos de inmigración que tuvieron lugar entre 1980 y 2000 incidieron en le estructura salarial de distintas ciudades en Estados Unidos. Estos flujos fueron heterogéneos en varias dimensiones, incluyendo el lugar de procedencia de los migrantes como su nivel de calificación. En el estudio, el autor hizo uso de la base de datos "Card.dta", la cual amablemente compartió con ustedes jóvenes promesas de la econometría. Así mismo, les envió dos archivos de ayuda: "Codebook.txt", el cual contiene el directorio de las variables incluidas en la base; y, "country_ids.xslx", el cual contiene la codificación de los países de procedencia de los migrantes. En esencia, esta base de datos contiene los flujos de inmigrantes a 124 ciudades de Estados Unidos durante el periodo 1980-2000, junto con algunas características de estos lugares, tales como la población de inmigrantes que residía en 1980 (antes del gran éxodo), su tamaño y su desarrollo industrial.

La relación que se explora en este artículo está dada por

$$w_{lj} = \beta_0 + \tau \ln(r_{lj}) + \beta_2 \boldsymbol{X}_l + \epsilon_{lj}$$

donde l es un indicador de la ciudad y j indexa un subpoblación caracterizada por su nivel educativo. Asimismo, w_{lj} es la brecha salarial (en logs) residual entre inmigrantes y nativos en el grupo j[2], r_{lj} es el ratio entre la cantidad de horas de trabajo de inmigrantes y nativos observadas en el grupo j en la ciudad l. Finalmente, $\boldsymbol{X}_l$ es una serie de controles a nivel de ciudad. En este trabajo, nos interesa determinar la influencia que tiene un incremento en el número de horas trabajadas por extranjeros relativo a aquellas provistas por los locales en la diferencias salariales en dos niveles: cuando los trabajadores cuentan sólo con un grado de educación secundaria, versus cuando tienen un grado equivalente a un título universitario. Así las cosas, los grupos de interés son $j \in \{hs, coll\}$, esto es, la población que cuenta a lo sumo con un grado de educación secundaria (hs) y aquellos con un título profesional ($coll$).

a) Expliquen qué tipo de relación captura τ entre w_{lj} y r_{lj}. ¿Creen ustedes que el estimador de MCO $\hat{\tau}$ es consistente para τ? Justifiquen con base en el contexto.

Enfrentado a un problema de endogeneidad, el autor opta por una estrategia de variables instrumentales. No obstante, no lo hace de la manera usual: Card usa una variación de variables instrumentales conocida como "Bartik IV" o como "Shift-share IV", el cual se llama en honor al paper seminal de Bartik (1991). En esencia, el Bartik IV es una variable instriumental que utiliza la exposición diferencial de las unidades de la muestra a un choque común. La estrategia shift-share tiene gran acogida en economía para estudiar temas de migración y de comercio internacional, aunque también ha sido utilizada en economía laboral, economía política, desarrollo, macroeconomía y finanzas.

Para ahondar un poco más en el asunto, podemos pensar el problema de la siguiente manera: considere que la relación que se quiere evaluar es de la forma

$$y_i = \rho + \beta x_i + \epsilon_i$$

pero sospechamos que $\mathbb{E}[\epsilon_i | x_i] \neq 0$, por lo que el estimador de MCO sería inconsistente del parámetro de interés. Supongamos además que la relación de interés x_i se puede descomponer como

$$x_i = \sum_k z_{ik} g_{ik}$$

para algunos vectores z_i, g_i, donde i in. Para simplificar la exposición, trabajemos con un ejemplo en particular: Suppongamos que y_i es la tasa de crecimiento del salario en un lugar i mientras que x_i es la tasa de crecimiento del empleo. Note entonces que la tasa de crecimiento del empleo total en i es la suma ponderada de todos las tasas de crecimiento *en cada industria*. Así las cosas, z_{ik} representaría la proporción que la industria k ocupa en el lugar i, mientras que g_{ik} su tasa de crecimiento. Antes de seguir, trate de convencerse de que en este contexto particular efectivamente $x_i = \sum_k z_{ik} g_{ik}$.

[2]Más precisamente, es la diferencia entre el logaritmo del salario residual promedio de los inmigrantes y el logaritmo salario resiudal promedio de los nativos. Para obtener el componente residual, estimamos modelos de regresión lineal en cada población en aras de aplicar el método de "partialling-out" para separar el componente del salario que no depende de características del individuo tales como su edad, su sexo, etc.

Ahora bien, para resolver el problema de endogeneidad debemos recurrir a variables instrumentales. En estos casos podemos optar por un instrumento tipo shift-share. Para ello, suponemos que

$$g_{ik} = g_k + \tilde{g}_{ik}$$

donde g_k es un factor intrínseco al factor k, mientras que $\tilde{g}_{ik}$ es un componente idiosincrático al par $i-k$. Remitiéndonos al ejemplo nuevamente, g_k es la tasa de creciemiento promedio de la industria k (noten que no depende i), mientras que $\tilde{g}_{ik}$ refiere a características particulares de la industria k en el lugar i que la desvían del promedio. Entonces, el instrumento Bartik está dado por:

$$B_i = \sum_k z_{ik} g_k$$

En palabras sencillas, el instrumento resume cómo un choque común a todas las unidades g_k, el "shift", afecta diferencialmente las distintas unidades i de acuerdo a los pesos implícitos z_{ik}, conocidos como "shares". Volviendo al ejemplo, supongamos hay un choque de demanda común (para todos los i) en la industria manufacturera. Lo que dice el instrumento es que la tasa de empleo en el lugar i, x_i, debería reaccionar más si la industria manufacturera en ese sitio es muy fuerte (z_{ik} es grande). Análogamente, aquellos sitios donde la industria no juega un papel importante, deberían permanecer relativamente inalterados por el choque. De esta manera, la variación en la intensidad de los choques generadas por las características intrínsecas de i nos permite potencialmente identificar un efecto causal al estimar el modelo por Mínimos Cuadrados en 2 Etapas (MC2E).

Si quieren investigar más sobre shift-shares pueden consultar los siguientes enlaces: Link 1, Link 2 (Cap 7.8.3), Link 3.

Volviendo al trabajo de Card, el autor propone el siguiente instrumento tipo Bartik:

$$B_{lj} = \sum_c z_{lc,1980}\ g_{cj}$$

Donde, los "shares" están dados por

$$z_{lc,1980} = \frac{N_{lc,1980}}{N_{c,1980}} \times \frac{1}{P_{l,2000}}$$

donde $N_{c,1980}$ es el número de inmigrantes provenientes de cada uno de los 38 países (indexados por c) que vivían en Estados Unidos en 1980, mientras que $N_{lc,1980}$ representa el total de inmigrantes provenientes del país c pero que viven en la ciudad l. Finalmente, $P_{l,2000}$ es la población de la ciudad l en el 2000. La idea de incluir este factor es expresar el número de migrates que llega a una ciudad l como fracción de su población (simplemente es un ajuste de escala).

Por su parte, el "shift" está dado por g_{cj} y es sencillamente el número total de personas que ingresaron *a todo Estados Unidos* entre 1980 y el 2000 provenientes del país c y con educación alcanzada j. Así las cosas, siguiendo la idea de Bartik, B_{lj} es un potencial instrumento de $\ln(r_{lj})$.

b) Con base en la estrategia de Card:

   i) Expliquen intuitivamente por qué el instrumento tipo Bartik puede predecir cambios en los ratios de horas trabajadas entre inmigrantes y nativos.

   ii) Planteen la ecuación estructural, la primera etapa y la forma reducida asociada al problema.

Como hemos visto en la teoría de variables instrumentales, para que esta metodología funcione necesitamos algunos supuestos acerca de cómo se relaciona el instrumento con las demás variables del modelo. Por simplicidad, trabajaremos bajo el supuesto que los efectos de tratamiento son constantes. En este contexto, para que el estimador de MC2E sea consistente de τ requerimos de dos supuestos:

- **Relevancia:** Para todo $c = 1, \cdots, C$:

$$\ln(r_{lj}) = \zeta_0 + h_{cj} * \ z_{cl,1980} + \zeta_2 \boldsymbol{X}_l + \varepsilon_{lj}$$

  donde h_{cj} debe ser finito para todo par (c, j) y $\sum_c h_{cj} g_{cj} \neq 0$

- **Exogeneidad estricta:** Para todo c tal que $g_{cj} \neq 0$

$$\mathbb{E}[\epsilon_{lj}\, z_{cl,1980} | \boldsymbol{X}_l] = 0$$

Así las cosas, el supuesto de relevancia establece que, condicional en $\boldsymbol{X}_l$, debe existir al menos un k para el cual $z_{kl,1980}$ tenga poder predictivo sobre $\ln(r_{lj})$ y que esta influencia sea tal que no se cancele al agregar los choques. Por su parte, exogeneidad estricta es la restricción de exclusión tradicional, pero esta recae sobre los "shifts". Esencialmente, lo que esta condición plantea es que las poblaciones de inmigrantes en 1980 provenientes de un país c, *condicional en* $\boldsymbol{X}_l$, no deben estar correlacionados con el factor no observado de la brecha salarial observada en el 2000. Noten, sin embargo, que ello debe valer sólo para todos aquellos poblaciones donde hubo efectivamente flujo de migrantes. En palabras más sencillas, lo que establecen ambas condiciones es que **la única manera por la cual tener distintas proporciones de inmigrantes de diferentes nacionalidades en 1980 pudo afectar la brecha salarial en los 2000 en una ciudad con características $\boldsymbol{X}_l$ es porque esa distribución particular atrajo más inmigrantes de las flujos de inmigración observados entre 1980 y el 2000.**

c) Utilizando el contexto y los datos provistos

   i) Estime τ por MCO y por MC2E tanto para la población con sólo título de secundaria como para aquellos con título de universidad. Muestre sus resultados en **una sóla tabla**. Incluya como controles el residual de los salarios para inmigrantes y nativos en 1980, el logaritmo de la población en 1980 y en 1990, así como la proporción de estudiantes en la universidad y la fracción que ocupaba la industria manufacturera en estos mismos años. Además, utilice como pesos para las regresiones la población de 1990. ¿Por qué es importante considerar estos pesos?

      Finalmente, para cada estimación por MC2E, incluya el estadístico F en la tabla resultante. ¿Por qué es importante reportar e interpretar este estadístico cuando se hace variables instrumentales?

   ii) Suponiendo que los supuestos de identificación se cumplen, interpreten los resultados.

   iii) Argumenten si, desde su punto de vista, los supuestos de identificación son válidos en el contexto del problema.

Una ventaja importante de los instrumentos tipo Bartik es que se puede testear el supuesto de exogeneidad. Para lograrlo, se puede evaluar si los "shifts" (o aquellos más importantes) se correlacionan con otros factores que pueden explicar diferencias en la brecha salarial en el tiempo. Uno de los posibles ejercicios es el siguiente:

d) Utilizando los datos provistos:

   i) Completen la siguiente tabla:

	Fracción de inm. de México 1980	Fracción de inm. de Filipinas 1980	Fracción de inm. de El Salvador 1980	Fracción de inm. de China 1980	Fracción de inm. de Cuba 1980	Fracción de inmig. de Europa del Este + Otros 1980	Bartik para secundaria	Bartik para universidad
Log población 1980								
Fracción univ. 1980								
Salario res. nativo 1980								
Salario res. inm. 1980								
Fracción ind. manuf. 1980								
R^2								
N								

Errores estándar robustos a heterocedasticidad en parentésis

   donde cada columna son los coeficientes y errores estándar resultantes de estimar un modelo de regresión múltiple de la variable indicada en la columna contra todas las variables explicativas señaladas en las filas.

Nota 1: Por otros, se refiere a Australia, Chipre, Israel y Nueva Zelanda.

Nota 2: Para que los resultados sean legibles, reescale los coeficientes y errores estándar por 10,000,000.

Nota 3: Utilice la población en 1990 como pesos tal como se hizo en el inciso anterior.

ii) Interpreten los resultados a la luz de los supuestos de identificación. ¿La información encontrada refuta o apoya el cumplimiento de estos supuestos?

Suponga ahora ustedes quieren replicar este ejercicio para Colombia en aras de entender el impacto que ha tenido la migración venezolana en las brechas salariales. No obstante, ustedes no cuentan con una información tan detallada como la que existe en Estados Unidos. Tan sólo cuentan con la siguiente información:

- Fracción de inmigrantes venezolanos en el 2000 en cada municipio. Noten que uno menos esa cantidad es la fracción de inmigrantes de otros países.
- Cantidad de inmigrantes que ingresaron entre el 2000 y el 2020 al país, discriminados **únicamente** por el hecho de si son venezolanos o no lo son.
- Sólo tienen información para la población con título universitario.
- r_{lj} y w_{lj} como arriba. (Pueden prescindir del j si quieren)

e) En este contexto:

i) Construyan el instrumento tipo Bartik análogo al de Card considerando los datos disponibles.

ii) Demuestren que en este escenario, usar como variable instrumental el Bartik hallado en $i)$ es equivalente a usar como variable instrumental la fracción de inmigrantes venezolanos en el 2000.

Ayuda: Reemplacen en la primera etapa del modelo la forma del Bartik encontrada en $i)$

iii) Recurriendo al resultado encontrado en $ii)$, bajo qué condiciones el efecto encontrado es el LATE. ¿Cuáles de estos supuestos se cumplen plausiblemente?

iv) En palabras describan quiénes son los compliers en esta situación.

Tercer ejercicio

En 2018, el Gobierno de Genovia implementó un programa de capacitación laboral enfocado hacia el sector agrícola. Esta iniciativa buscaba fomentar en los trabajadores de dicho sector una serie de habilidades blandas que el Gobierno consideró serían cruciales para la adopción de nuevas tecnologías y la sofisticación de los procesos productivos. Aunque la postulación al programa fue voluntaria y contó con $1,000,000$ de aplicantes, la inscripción efectiva fue restringida debido a los limitados recursos dispuestos para esta política.

Con el fin de hacer un filtrado "justo", los aplicantes debieron presentar un examen compuesto por una parte de razonamiento matemático y otra de comprensión lectora, ambas calificadas entre 0 y 100. Finalmente, se estableció un corte en el puntaje de ambas secciones del examen, de manera que **solo aquellos (y todos aquellos)** que obtuviesen puntajes superiores a los cortes de cada sección serían vinculados al programa de capacitación. Para atender preocupaciones de equidad, los cortes fueron definidos de manera diferencial entre los trabajadores de empresas productoras de aceite de oliva (a quienes se referirá como *grupo 1*) y el resto de empleados agrícolas (*grupo 0*), pues diversos estudios sugieren que los primeros cuentan en promedio con un mayor grado de capacitación y educación, lo que haría injusto evaluar bajo el mismo estándar a todos los trabajadores. Por ello, para el grupo 1 se estableció un umbral de 80 puntos en cada parte del examen para ser admitido en el programa, mientras que para el grupo 0 dicho umbral fue de 70 puntos.

Cuatro años después de la implementación inicial del programa, el Gobierno de Genovia desea conocer si su política tuvo impactos significativos sobre el salario de los trabajadores. Así, los han contratado a ustedes para efectuar un análisis cuidadoso sobre los efectos del programa de capacitación. Tras consultar con distintos académicos y trabajadores y efectuar una extensa revisión de literatura sobre el impacto de programas de este estilo, ustedes proponen el siguiente modelo teórico:

$$salario_i = 200 + 0.05 * matematicas_i + 1.2 * matematicas_i^2 - 0.012 * matematicas_i^3$$

$$+0.8 * lectura_i + 20 * G_i + 50 * T_i + 10 * G_i * T_i + \varepsilon_i$$

Donde $salario_i$ corresponde al salario del trabajador i, $matematicas_i$ es su puntaje obtenido en la sección de matemáticas, $lectura_i$ es su puntaje obtenido en la sección de comprensión lectora, G_i es la indicadora de si i pertenece al grupo 1, T_i indica si i participó efectivamente en el programa, y ε_i es un término de error idiosincrático.

a) Interprete los términos del anterior modelo.

b) Mientras se formaliza un acuerdo de confidencialidad que permita al Gobierno entregarles una versión más detallada de los datos, el Gobierno les provee la base de datos *init_anon.dta*, la cual cuenta con información anonimizada de los trabajadores agrícolas **no** productores de aceite de oliva. En esta base se encuentran únicamente los trabajadores que superaron el umbral de la prueba de lectura. Adicionalmente, los puntajes de la prueba de matemáticas han sido discretizados acorde a la siguiente codificación:

Clasificación	Puntaje obtenido
0	0-10
1	10-20
2	20-30
3	30-40
4	40-50
5	50-60
6	60-70
7	70-80
8	80-90
9	90-100

Finalmente, ustedes cuentan con información salarial de los trabajadores, codificada en la variable *wage*. Usando el comando *rdrobust/CCT*, estime el *efecto marginal*[3] del programa de capacitación sobre el salario. (*Pista:* use la variable discretizada como su *running variable* y establezca un punto de corte apropiado.). Interprete sus resultados. Discuta intuitivamente: ¿Por qué el hecho de que la *running variable* sea discreta puede producir problemas a la hora de estimar el parámetro de interés?

c) Ustedes manifiestan sus preocupaciones a una colega, la cual les sugiere abandonar el enfoque de continuidad local y emplear en cambio un enfoque de aleatorización local. Bajo dicho enfoque, se considera que existe una ventana en torno al punto de corte al interior de la cual encontrarse a un lado u otro del umbral es aleatorio, en el sentido de que el salario potencial es independiente de la asignación al tratamiento. Formalmente, si c es el corte, suponemos que existen reales a, b tales que $a < c < b$ y $T_i \perp\!\!\!\perp (salario_i(1), salario_i(0))$ en $\mathcal{W} = (a, b)$. De esta forma, cobra sentido plantear un estimador del efecto de tratamiento dado por una diferencia de medias ponderada (por algunos pesos w_i):

$$\hat{\tau} = \bar{salario}^{+}_{\mathcal{W}} - \bar{salario}^{-}_{\mathcal{W}}, \quad \bar{salario}^{+}_{\mathcal{W}} = \frac{1}{N^{+}_{\mathcal{W}}} \sum_{i:X_i \in \mathcal{W}} w_i T_i Y_i, \quad \bar{salario}^{-}_{\mathcal{W}} = \frac{1}{N^{-}_{\mathcal{W}}} \sum_{i:X_i \in \mathcal{W}} w_i T_i Y_i$$

Donde $N^{+}_{\mathcal{W}} = \sum_{i=1}^{n} T_i \mathbb{1}\{X_i \in \mathcal{W}\}$ y $N^{-}_{\mathcal{W}} = \sum_{i=1}^{n} (1 - T_i) \mathbb{1}\{X_i \in \mathcal{W}\}$. En este orden de ideas, se está tomando un promedio ponderado del salario de las unidades tratadas y no tratadas que se encuentran en la ventana de aleatorización, y luego computamos la diferencia de dichos promedios.

Ahora bien, en aras de poder efectuar inferencia estadística de manera correcta, debemos suponer que se conoce el proceso de asignación de tratamiento al interior de la ventana (esto es, debemos poder hacer *randimization inference* como en el taller 1). En palabras sencillas, requerimos que la distribución del vector de tratamiento en la ventana $\mathbf{T}|\mathbf{X} \in \mathcal{W}$ sea conocida. Así las cosas, si se cuenta con n individuos, se hace necesario establecer una distribución de probabilidad sobre el conjunto Ω de posibles asignaciones de tratamiento, donde

$$\Omega = \{0, 1\}^n, \quad \text{supp}(\mathbf{T}|\mathbf{X} \in \mathcal{W}) = \Omega.$$ [4]

Existen dos aproximaciones comunes para elegir dicha distribución.

[3] Marginal en el sentido de que es el efecto de superar el umbral de la prueba de matemáticas dado que ya se superó el umbral de la prueba de lectura.

[4] El soporte asociado a la distribución de una variable aleatoria (discreta) es el conjunto de valores que puede tomar con una probabilidad estrictamente positiva.

- Por una parte, el enfoque de ***márgenes fijos*** sugiere tomar como dado el número de unidades tratadas observado en la ventana, k, y considerar que todos los vectores de tratamiento en los cuales exactamente k unidades son tratadas son equiprobables. Es decir, se impone una distribución uniforme sobre el subconjunto de vectores de tratamiento

$$\Omega' = \left\{ \boldsymbol{x} = (x_1, \cdots, x_n) \in \{0,1\}^n : \sum_{i=1}^{n} x_i = k \right\} \subset \Omega$$

  Donde la probabilidad de cada punto viene dada por $\frac{1}{\binom{n}{k}}$. En este escenario, para el cálculo del estimador puntual, los pesos $w_i = 1$, para todo i, producen un estimador insesgado del efecto local de tratamiento.

- Por otra parte, el enfoque de ***experimentos de Bernoulli*** considera que cada unidad en la ventana tiene exactamente la misma probabilidad de ser tratada, p, por lo que $T_i | X_i \in \mathcal{W} \sim \text{Bernoulli}(p)$. De esta forma, tomar pesos $w_i = \frac{N_{\mathcal{W}}^-}{p(N_{\mathcal{W}}^+ + N_{\mathcal{W}}^-)}$ para las unidades debajo del corte y $w_i = \frac{N_{\mathcal{W}}^+}{p(N_{\mathcal{W}}^+ + N_{\mathcal{W}}^-)}$ para las unidades encima del corte produce nuevamente un estimador insesgado.

Empleando el enfoque de aleatorización local en la ventana $(4.5, 7.5)$, estimen el efecto del programa sobre el salario utilizando

- Márgenes fijos.
- Experimentos de Bernoulli, con $p = 1/3$.

En cada caso, calcule $N_{\mathcal{W}}^+,\ N_{\mathcal{W}}^-$ y los w_i que producen el estimador insesgado. Presenten sus resultados en una tabla como la siguiente, interprétenlos y compárenlos con lo obtenido en el numeral 2. De acuerdo al modelo teórico y sus resultados, ¿es mejor usar un enfoque de continuidad local o uno basado en aleatorización local cuando se tiene una variable de focalización discreta? Justifique (*Pista:* consulten el comando **rdrandinf** de Stata):

Enfoque	$N_{\mathcal{W}}^+$	$N_{\mathcal{W}}^+$	w_i si $X_i < c$	w_i si $X_i \geq c$	$\hat{\tau}$	p-valor exacto
Márgenes fijos						
Exp. de Bernoulli						

d) Finalmente, el Gobierno les facilita una base de datos más completa, *full_program.dta*, la cual dispone de la información salarial y de los puntajes obtenidos en cada prueba de **todos** los trabajadores que aplicaron al programa. A continuación se encuentra una breve descripción de las variables disponibles:

Variable	Descripción
id	Identificador del individuo
grupo	=1 si es productor de aceite de oliva
matematicas	Puntaje obtenido en la sección de matemáticas
lectura	Puntaje obtenido en la sección de lectura
wage	Salario

i) Para comenzar, vuelvan a estimar el efecto del programa solo para los trabajadores no productores de aceite de oliva empleando como *running variable* el puntaje de la prueba de matemáticas. Hagan esto:
   - Limitando la muestra a aquellos que superaron el umbral de la prueba de lenguaje.
   - Sin limitar la muestra.

   Interprete los resultados de las dos estimaciones y compárelos entre ellos y con los resultados obtenidos en los puntos 2 y 3. ¿Qué pueden decir sobre la magnitud del coeficiente obtenido con la muestra completa respecto al de la muestra limitada?

ii) En su nueva base, ustedes cuentan con observaciones de dos poblaciones, cada una de las cuales posee puntos de corte distintos. Para estimar un efecto de tratamiento que tome en consideración a ambas poblaciones, podemos construir un estimador *pooled* como sigue. Si X_i es una running variable, defina $\tilde{X}_i = X_i - c_i^X$ como la *running variable* recentrada, donde c_i^X es el corte de la variable X que aplica al individuo i. De esta forma, ahora ambas poblaciones tienen 0 como punto de corte en la variable recentrada, por lo que es posible estimar un efecto de tratamiento único empleando las herramientas usuales. Dicho efecto corresponde a un promedio ponderado de los efectos de tratamiento de cada población, donde los ponderadores vienen dados por qué tan probable es que una unidad en el punto de corte pertenezca a cada población.

- Construya las *running variables* recentradas.
- Limitando la muestra a las observaciones que superaron el umbral de la prueba de lectura, calculen e interpreten el estimador *pooled*. ¿Cómo se compara con los otros estimadores obtenidos hasta ahora? ¿Qué les permite concluir esto acerca el efecto de la intervención sobre los trabajadores productores de aceite de oliva?

Presenten e interpreten sus resultados. ¿Cómo se compara lo estimado en $ii)$ con lo obtenido en los anteriores numerales?

d') **Bono:** Para finalizar el análisis, ustedes deciden considerar la muestra completa. En este escenario, una unidad es tratada si y solo si supera ambos umbrales (i.e. sus dos variables recentradas son mayores o iguales a 0).

iii.1) Lo anterior implica que las dos *running variables* definen un área de tratamiento. Para visualizar este hecho, realicen un gráfico de dispersión de las dos *running variables*. Asegúrese de graficar en colores distintos las unidades tratadas y no tratadas.

iii.2). Podemos pensar en dos tipos de efecto de tratamiento de interés: (1) el efecto en un punto específico de la frontera del área que graficaron previamente, y (2) el efecto promedio en la frontera.

- Estimen el efecto del programa *en el punto (0,0)*. Es decir, el efecto de apenas superar ambos umbrales. Para ello, consulten el comando **rdms** de Stata. Dicho comando calcula automáticamente la distancia euclidia de cada unidad al punto de interés, y luego estima una regresión discontinua por polinomios locales empleando dicha distancia como la *running variable*.
- Estimen el efecto promedio en la frontera. Para esto, deben calcular la distancia euclidia de cada unidad a la frontera y emplearla como *running variable*.

  Pista: la fórmula que determina la distancia a la frontera tiene 4 casos:
  - Cuando ambos puntajes son mayores al corte.
  - Cuando el puntaje de matemáticas es mayor al corte, pero el de lectura es menor.
  - Cuando el puntaje de lectura es mayor al corte, pero el de matemáticas es menor.
  - Cuando ambos puntajes son menores al corte.

  Presenten e interpreten sus resultados. ¿Cómo se compara $iii.2)$ con lo obtenido en los anteriores numerales?