IN5244 Ciencia de Datos, Semestre Primavera 2024
Universidad de Chile - Departamento de Ingeniería Industrial
Profesores: Raimundo Undurraga y Richard Weber

Problem Set 2

Instrucciones Generales:

- La tarea es de carácter estrictamente individual
- Fecha de entrega: Viernes 4 de Octubre, hasta las 23.59hrs.
- Adjuntar PDF con el texto de las respuestas, tablas, y graficos solicitados, y un código que replique los ejercicios empíricos, usando el programa estadístico de su preferencia (Stata, R, Python, etc.).

Defina dos paths dentro de su computador: uno que dirija hacia la carpeta donde está alojada la base de datos (llámelo data), y otro en una carpeta en donde guardará los gráficos y tablas que obtenga (llámelo output). En adelante, asegúrese de exportar todos los procedimientos, en sus formas de gráficos y tablas que se pidan, a esta carpeta.

Abra la base de datos casestudy_dropout.dta en Stata o el programa de su preferencia. La base de datos contempla una muestra de estudiantes que se desvincularon de su escuela en algún momento entre abril y julio del año 2022, definido por la variable DropoutDate. Estar desvinculado de la escuela implica estar fuera del sistema escolar, i.e., se encuentra fuera de la nómina de cualquier escuela pública del país. Algunos/as estudiantes logran re-vincularse a una escuela (sea la misma u otra), otros/as no.

El 28 de Junio de 2022, el MINEDUC envía a cada escuela un reporte sobre la situación de estudiantes desvinculados/as de esa escuela que, al 30 de Mayo de 2022, no han logrado revincularse a ninguna escuela (sea la misma u otra). La variable ListedInDropoutReport toma el valor 1 si el/la estudiante está listado en el reporte, y 0 si no. El objetivo del reporte es que las escuelas activen sus redes locales para lograr la revinculación de los/las estudiantes listados en el reporte, sea en la misma escuela que abandonaron u otra.

Al 31 de Agosto de 2022, algunos/as estudiantes habían logrado revincularse al sistema educacional, i.e., reingresaron a una escuela (sea la misma u otra), lo cual puede medirse por la variable EnrolledByAug312022. Algunos no sólo lograron revincularse, sino que además lograron graduarse a fin de año, medido por la variable GraduatedIn2022.

Su objetivo es ayudar a evaluar el impacto causal de la política de reportes del MINEDUC en la probabilidad de revinculación y graduación de estudiantes desvinculados del sistema escolar. Su aporte puede ser relevante para diseñar políticas de revinculación escolar que cambien la trayectoria educativa de jóvenes en situación de vulnerabilidad social y educacional.

1 Regresión Discontinua

La variable SchoolDaysFromMay302022 cuenta el numero de días escolares entre la fecha de retiro de la escuela (DropoutDate) y la fecha de corte establecida por el MINEDUC para construir el reporte, 30 de Mayo de 2022. Para efectos de un ejercicio de regresión discontinua, usaremos SchoolDaysFromMay302022 como nuestra *running variable*.

1. Construya un gráfico de frecuencias de la *running variable*, y marque con una línea vertical la fecha de corte de asignación (*cutoff*). Superponga en el gráfico una proyección lineal específica a cada lado del cutoff, permitiendo discontinuidad en el cutoff. Hay evidencia de manipulación de la *running variable* en torno al punto de corte de asignación? Compruebe empíricamente utilizando el comando rddensity en Stata (si no usa Stata, construya el test a mano en su programa de preferencia). [15 puntos]

2. Construya un 1st stage graph que mapee la variable ListedInDropoutReport (eje Y) contra la *running variable* (eje X), y dibuje una línea vertical en el cutoff. Superponga en el gráfico una proyección lineal específica a cada lado del cutoff, permitiendo discontinuidad en el cutoff.

   (a) Cuál es la utilidad de este gráfico? [5 puntos]

   (b) Testee empíricamente si la probabilidad de estar listado en el reporte es estadísticamente distinta de cero en torno al cutoff. Describa el resultado (Hint: regresione ListedInDropoutReport contra DroppedOutMay30Before, SchoolDaysFromMay302022, DroppedOutMay30Before × SchoolDaysFromMay302022, controlando por efectos fijos a nivel de escuela.) Interprete. [10 puntos]

   (c) Un colega le sugiere que incluya, además, efectos fijos por día de la semana en que el/la estudiante se retiró de la escuela. Justifique dicha decisión a través de evidencia de un grafico de frecuencia y re-estime siguiendo el consejo del colega creando dummies para la variable DayOfWeekDroppedOut. Describa el resultado [5 puntos]

3. Utilice un modelo de regresión discontinua lineal para estimar el impacto de haberse retirado de la escuela antes del cutoff sobre la probabilidad de estar revinculado en la escuela al 31 de agosto (Hint: regresione EnrolledByAug312022 contra DroppedOutMay30Before, SchoolDaysFromMay302022, SchoolDaysFromMay302022 × DroppedOutMay30Before, controlando por efectos fijos a nivel de escuela y efectos fijos a nivel de día de la semana en que sucedieo el retiro.). Replique usando como outcome la variable GraduatedIn2022. Describa los resultados [10 puntos]

4. Usando el comando rdplot en Stata, construya un gráfico que mapee EnrolledByAug312022 (eje Y) contra la *running variable* (eje X). Dibuje una línea vertical en el cutoff, y la proyección lineal de Y sobre X separadamente a cada lado del cutoff. Replique usando GraduatedIn2022 en el eje Y. Postee ambos graficos side-by-side. Qué concluye? Interprete. [10 puntos]

5. Un colega le sugiere chequear que el efecto no sea sensible a la forma funcional utilizada. Hasta ahora hemos asumido una relación lineal entre el outcome y la *running variable*.

En efecto, el colega le sugiere utilizar una forma cuadrática. Verifique según le aconseja su colega (Hint: corra la misma regresión lineal del punto 3, pero incluya además SchoolDaysFromMay302022^2 y SchoolDaysFromMay302022^2 $\times$ DroppedOutMay30Before en la regresión). Interprete [10 puntos]

6. Verifique el resultado anterior haciendo uso del comando rdplot en Stata, i.e., construya un gráfico que mapee EnrolledByAug312022 (eje Y) contra la *running variable* (eje X). Dibuje una línea vertical en el cutoff, y la proyección cuadrática de Y sobre X separadamente a cada lado del cutoff. Interprete. Cambia el resultado si usa una forma funcional cúbica? [10 puntos]

7. Usando el comando rdplot en Stata, construya un gráfico que mapee GraduatedIn2021 (eje Y) contra la *running variable* (eje X). Dibuje una línea vertical en el cutoff, y la proyección lineal de Y sobre X separadamente para cada lado del cutoff. Por qué es importante realizar este ejercicio? Qué concluye? Interprete. [10 puntos]

8. Replique el ejercicio anterior usando pre-treatment covariates adicionales, incluyendo GPAin2021, AttendanceIn2021, Female, y Migrant. Qué concluye? [10 puntos]

9. El 1st stage graph de la pregunta 2 sugiere que el cambio en la probabilidad de ser listado en el reporte no es *sharp*, sino *fuzzy*. Lo anterior indica que para estimar el impacto de ser listado en el reporte (ListedInDropoutReport) en revinculación escolar (EnrolledByAug312022) debiésemos usar un RD Fuzzy, que es, en el fondo, un estimador por variables instrumentales.

   (a) Escriba matemáticamente el estimador RD Fuzzy. Qué es el numerador? Qué es el denominador? [5 puntos]

   (b) Describa los supuestos de identificación del método [5 puntos]

   (c) Describa operacionalmente como se implementa el método [5 puntos]

10. Haga uso del estimador RD Fuzzy para estimar el impacto de ser listado en el reporte (ListedInDropoutReport) sobre revinculación escolar (EnrolledByAug312022). RD Fuzzy es un estimador en variables instrumentales, lo cual requiere estimar en dos etapas. Para ello:

    (a) Vaya a buscar el coeficiente de impacto de la pregunta 3, cuando el outcome es EnrolledByAug312022, y llamele "Intention-to-treat estimate". Luego vaya a buscar el coeficiente asociado al 1st stage de la pregunta 2, y llamele "1st stage estimate". El RD Fuzzy es el cuociente de Intention-to-treat estimate sobre 1st stage estimate. Interprete. [5 puntos]

    (b) Ahora estimemos el RD Fuzzy usando el método de variables instrumentales. Luego de correr la regresión del 1st stage, genere una predicción de la probabilidad de ser incluido en el reporte usando el comando predict en Stata. Llame a esa predicción

ListedInDropoutReport_Predicted. Luego corra la regresión de la variable EnrolledByAug312022 contra ListedInDropoutReport_Predicted, SchoolDaysFromMay302022, SchoolDaysFromMay302022 $\times$ DroppedOutMay30Before, controlando por efectos fijos a nivel de escuela y efectos fijos a nivel de día de la semana en que sucedieo el retiro. Interprete. [10 puntos]