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

Problem Set 1

Instrucciones Generales:

- La tarea es de carácter estrictamente individual
- Fecha de entrega: Viernes 30 de Agosto, 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 Estadística Descriptiva

1. Describa cuál es la unidad de análisis, y cuál es el tamaño de la muestra. ¿A qué se refieren las columnas IDstudent y IDschool? [5 puntos]

2. Genere una tabla de estadística descriptiva en la que muestre la distribución univariada de las siguientes variables: Age, Female, AttendanceIn2021, GPAin2021, MunicipalSchool. Compute la media, la mediana, la desviación estándar, el percentil 25 y el percentil 75 de cada una de estas variables. Incluya, además, el tamaño muestral. ¿Es este último igual para todas las variables? [5 puntos]

3. Replique la misma tabla de estadística descriptiva, pero esta vez separando para estudiantes chilenos y migrantes. ¿Se observan diferencias entre ambos grupos? [5 puntos]

4. Para las mismas 5 variables analizadas, realice un test de medias entre características de estudiantes chilenos y migrantes, utilizando la distribución t-student. Asuma que las varianzas de los grupos son distintas. ¿Qué diferencias son significativas? Distinga en su análisis entre variables dicotómicas y continuas. [10 puntos]

5. Replique el análisis de la pregunta anterior, pero esta vez sólo considerando estudiantes de educación básica (puede obtener esa información a partir de la variable SchoolGrade2022). [10 puntos]

6. Realice un histograma de frecuencias que muestre la distribución de la fecha en la que el/la alumno/a se desvincula de la escuela, que puede obtener utilizando la variable DropoutDate. Marque con una línea vertical el día 30 de Mayo. ¿Observa algún tipo de estacionalidad? Confirme esto haciendo uso de la variable DayOfWeekDroppedOut. En particular, verifique si la desvinculación es mas probable en algunos días de la semana que en otros. [15 puntos]

7. *Colapse* la base de datos *por fecha en la que el/la alumno/a se desvincula de la escuela* (i.e., cada observación es una fecha), y detalle la fracción de alumnos que entraron al reporte en cada fecha (dado por la variable ListedInDropoutReport). Realice un scatterplot que muestre en el eje y la fracción de alumnos que fueron listados en el reporte, y en el eje x la fecha en que se desvincularon de la escuela. Incluya una línea vertical en el día 30 de mayo. Observa alguna discontinuidad? Interprete [20 puntos]

2 Causalidad y Sesgo de Selección

8. Considere el siguiente modelo de regresión lineal simple:

$$y_i = \beta_0 + \beta_1 T_i + \varepsilon_i \tag{1}$$

Donde y_i es una variable dicotómica que toma el valor 1 si el/la estudiante i se ha revinculado al sistema escolar al 31 de Agosto de 2022, y 0 si no. T_i es una variable dicotómica que indica si el/la estudiante i está listada en el reporte, y 0 si no. ε_i es un término de error, que contiene características no observadas del el/la estudiante i asociadas a la

probabilidad de revinculación escolar. Describa, utilizando la **notación e intuición** del modelo de Rubin, el significado de las siguientes cuatro expresiones, señalando cuáles son observables y cuáles no:

(a) $E[y_{1i} \mid T_i = 1]$ [5 puntos]

(b) $E[y_{0i} \mid T_i = 1]$ [5 puntos]

(c) $E[y_{1i} \mid T_i = 0]$ [5 puntos]

(d) $E[y_{0i} \mid T_i = 0]$ [5 puntos]

9. Muestre formalmente que, al estimar β_1 por Mínimos Cuadrados Ordinarios - esto es, hacer una diferencia de medias entre estudiantes listados y no listados-, obtenemos una expresión que contiene el efecto promedio sobre los listados (ATET), y otro componente que es una potencial fuente de sesgo de selección. [10 puntos]

10. Utilizando las cuatro expresiones (a)-(d) de la pregunta 8, describa cuál es el supuesto que permite identificar, a través de Mínimos Cuadrados Ordinarios, el verdadero efecto causal del reporte. Describa cuál es el rol de ε. [10 puntos]

11. Considerando la ecuación (1), señale al menos tres variables observables que, al omitirse, podrían estar generando sesgo por variable omitida. [10 puntos]

12. Ahora entremos directamente en el caso de estudio. Tome T_i como la variable ListedInDropoutReport en su base de datos, y y_i como la variable EnrolledByAug312022, y corra la regresión. Para la inferencia, considere **errores estándar clusterizados a nivel de colegio** (a través de la variable IDSchool). Reporte sus resultados en una tabla y confirme que lo que obtiene es análogo a realizar un test-t de medias en la variable de salida, con los grupos de tratamiento y control definidos por T_i. [10 puntos]

13. Ahora corra las siguientes versiones alternativas del modelo de regresión (1):

    (a) Incluyendo efectos fijos a nivel de colegio.

    (b) Incluyendo como regresores al vector de variables de control conformado por Migrant, Female, Age, PrioritaryStudentIn2021, GPAin2021, además de los efectos fijos.

    Reporte sus resultados en una tabla (cada columna es una regresión), y describa cómo cambia el coeficiente estimado para β_1. Qué nos indica el hecho que las estimaciones de β_1 cambien conforme agregamos variables de control a la regresión? [10 puntos]

    (c) Restringiendo la muestra a estudiantes que se desvincularon en los meses de Mayo y/o Junio, con efectos fijos.

    (d) Restringiendo la muestra a estudiantes que se desvincularon en los meses de Mayo y/o Junio, con efectos fijos y el vector de variables de control.

    Reporte sus resultados en una tabla (cada columna es una regresión). Qué nos indica el hecho que las estimaciones de β_1 cambien conforme restringimos la muestra a

estudiantes que se desvincularon en los meses de Mayo y/o Junio? [10 puntos]

14. Usando el modelo de Rubin, muestre formalmente que aún cuando se controle en la regresión por variables relevantes, la estimación de $\hat{\beta}_{1,MCO}$ aún puede contener sesgo. [10 puntos]

15. Una colega le sugiere usar una fuente de variación exógena para resolver el problema de endogeneidad que trae la variable ListedInDropoutReport. Para ello, recomienda usar la variable DroppedOutMay30Before como un predictor exógeno de ListedInDropoutReport. La variable DroppedOutMay30Before toma el valor 1 si el estudiante se desvinculó de la escuela antes del 30 de Mayo y 0 si no. La colega argumenta que estudiantes que se desvincularon antes o después del 30 de Mayo debiesen ser, en promedio, similares, i.e., haberse desvinculado antes o después de esa fecha constituye un hecho fortuito ortogonal a las características de los estudiantes. Sin embargo, estudiantes que se desvincularon antes del 30 de Mayo tienen mas chances de estar listados en el reporte que aquellas/os que lo hicieron después de esa fecha.

    Cómo podríamos probar que DroppedOutMay30Before es una variable exógena al tratamiento? Realice un test de medias que compare estudiantes que se desvincularon antes y después del 30 de Mayo de 2022 para cada una de las siguientes variables: Migrant, Female, Age, PrioritaryStudentIn2021, GPAin2021. Son estos dos grupos estadísticamente similares en este set de variables observables? [10 puntos]

16. Ahora considere el siguiente modelo de regresión lineal simple:

$$y_i = \gamma_0 + \gamma_1 D_i + \varepsilon_i \tag{2}$$

    Donde y_i es una variable dicotómica que toma el valor 1 si el/la estudiante i se ha revinculado al sistema escolar al 31 de Agosto de 2022, y 0 si no. D_i es una variable dicotómica que indica si el/la estudiante i se desvinculó de la escuela antes o después del 30 de Mayo de 2022, y ε_i es un término de error, que contiene características no observadas de el/la estudiante i asociadas a la probabilidad de revinculación escolar. Corra la regresión del modelo 2, y además las siguientes alternativas del modelo de regresión:

    (a) Incluyendo efectos fijos a nivel de colegio.

    (b) Incluyendo como regresores al vector de variables de control conformado por Migrant, Female, Age, PrioritaryStudentIn2021, GPAin2021, además de los efectos fijos.

    Reporte sus resultados en una tabla (cada columna es una regresión) y describa cómo cambia el coeficiente estimado para γ_1. Qué nos indica el hecho que las estimaciones de γ_1 se mantengan relativamente robustas conforme agregamos variables de control a la regresión? [20 puntos]

17. Describa por qué las distintas estimaciones anteriores deben interpretarse bajo el concepto de *Intention to Treat* (y no bajo el concepto *Treatment Effect on the Treated*). [10 puntos]

18. Considere la versión de la regresión dada por (2), para revisar efectos heterogéneos entre hombres y mujeres. En primer lugar, corra la regresión separadamente para hombres y para mujeres. Describa qué observa y compare las estimaciones de γ_1 en cada subpoblación. [10 puntos]

19. Considere una versión extendida del modelo (2) dada por:

$$y_i = \beta_0 + \beta_1 D_i + \beta_2 \text{Mujer}_i + \beta_3 \text{Mujer}_i \times D_i + \varepsilon_i \tag{3}$$

Realice la estimación, interprete los coeficientes y presente sus conclusiones respecto de posibles efectos heterógeneos para hombres y mujeres. Cómo se interpreta β_0, β_1, β_2, y β_3? [10 puntos]

20. Considere ahora como outcome (esto es, como variable y_i) la variable dada por graduación en 2022 (GraduatedIn2022), y corra la regresión dada por (2). Adicionalmente, vuelva a correr la especificación, pero esta vez agregando como variable de control la variable EnrolledByAug312022. Explique, utilizando la materia vista en clases, por qué la segunda regresión presentaría problemas de sesgo de selección, y describa formalmente -matemáticamente- por qué la estimación de γ_1 cambia, y por qué cambia en la dirección que cambia.

21. Realice una prueba de Rerandomización de Fisher. Para esto, siga los siguientes pasos:

    (a) Corra la regresión inicial, dada por (2). Obtenga el estimador para γ_1, al que llamaremos gamma_original.

    (b) Cree artificialmente 1000 variables nuevas de re-randomización, en donde para cada una de las 1000 variables D_1 a D_{1000}, aleatoriamente le asigna a cada estudiante el valor 0 o 1 (tratado o no). [1]

    (c) Vuelva a correr 1000 veces la regresión dada por la ecuación (2), en donde cada una de las 1000 regresiones usa el D_k (con $k \in (1, 1000)$) correspondiente a las variables nuevas *rerandomizadas* que usted creó. Obtenga el estimador de γ_1 para cada una de estas regresiones, y declárelo como gamma_k en donde k es el valor del número de iteración.

    (d) Grafique la distribución de los 1000 gamma_k, y dibuje tres barras verticales: una barra vertical en el valor correspondiente al percentil 2.5 de la distribucion, otra barra vertical en el valor correspondiente al percentil 97.5 de la distribución, y finalmente una barra vertical en el valor del coeficiente gamma_original. ¿En qué percentil de la distribución de estos 1000 gamma_k se encuentra gamma_original? Que concluye? Interprete. [20 puntos]

[1] Esto puede hacerse de distintas maneras. Una posible estrategia es crear una variable auxiliar que siga una distribución uniforme entre 0 y 1, y asignar 0 si el valor es menor o igual a 0.5, y 1 en caso contrario.