Pontificia Universidad Católica de Chile
Departamento de Ingeniería Industrial y de Sistemas
ICS2563 - Econometría Aplicada
Sección 1

Tarea 3

26 de abril de 2022

2022-1 - Profesor Patricio Domínguez

Aspectos generales

- La tarea puede ser desarrollada en forma indivual o en parejas (2 estudiantes).
- La fecha de entrega de la tarea es el **5 de Mayo a las 23:59 hrs** en el portal del curso en Canvas.
- La entrega debe incluir el script de `R` (o el dofile de `STATA`) y un informe de análisis en formato pdf.
- Sobre el **informe**:
  - El informe debe incorporar las respuestas a todas las preguntas, incorporando las figuras y/o tablas que estime conveniente.
- Sobre el **código**:
  - Cada entrega debe incluir el script de `R` (o el dofile de `STATA`) desde donde se corre todo el análisis.
  - Es importante que el código esté adecuadamente comentado para facilitar su corrección. Por ejemplo, indicar qué parte del informe/preguntas se desarrollan en cada sección del código.
  - **Reproducibilidad:** El código debiera estar escrito de manera tal que cualquier persona lo pueda correr, y reproducir los resultados desde su computador. En el caso de crear datos aleatorios recomendamos fijar una semilla que permita reproducir los resultados tal cual queden especificados en el informe de reporte.
- Parte de la tarea es que se vean enfrentados a tomar decisiones. Si deben aplicar criterio, háganlo y justifiquen sus elecciones.

- Ante dudas o preguntas, se recomienda fuertemente la utilización del foro de preguntas del curso. Las respuestas pueden servir a otros compañeros, que se enfrentan a las mismas dudas. No se permite publicar respuestas.
- La tarea tiene un total de **95 puntos**, en donde la presentación del informe posee 5 puntos (El informe debe incorporar las respuestas a todas las preguntas, y las figuras y/o tablas que estime necesarias), y la presentación del script 5 puntos (todos los cálculos realizados).
- **Bonificacion:** Como un incentivo al uso del procesador de texto LATEX, se entregarán 5 puntos adicionales a quienes lo usen para escribir su informe.

Descripción de la tarea

El objetivo de esta tarea es calcular sesgos de variables relevantes omitidas y conocer más sobre experimentos aleatorios y el diseño de estos. Para ello, utilizaremos las mismas **bases de datos de la Tarea 1 y 2**, y además una adicional llamada `prog_educ.dta` que se utilizará a partir de la Pregunta 2.

Deben conectar las siguientes BDD:

1) SIMCE 8 básico 2019: Puntaje promedio SIMCE de todos los establecimientos que rindieron la prueba ese año. Descargar desde este Link.

2) Matrícula por establecimiento: Número de estudiantes por establecimiento reconocidos por el MINEDUC. Descargar desde este Link. y poner atención a variable mat_total.

3) Datos de estudiantes prioritarios por establecimiento: Número de estudiantes prioritarios para recibir la Subvención Escolar Preferencial (SEP) por establecimiento. Descargar desde este Link y poner atención a variable n_prio.

4) Programa educacional 2021 (`prog_educ.dta`): Esta base de datos contiene información sobre la implementación de un programa educacional en establecimientos educacionales de Chile. Poner atención a las variables `rbd` y `dgv_rbd`.

Preliminar

Nuevamente deben construir la variable proporción de estudiantes prioritarios dividiendo la cantidad de estudiantes prioritarios por el total de estudiantes matriculados en cada colegio.

Para esta tarea, consideren que la variable `dvrbd` en la BDD Simce contiene la misma información que la variable `dgv_rbd`.

Preguntas

1. Variable Omitida (25 puntos)

1. Imagine que el proceso generador de datos (DGP) del puntaje estandarizado del SIMCE de matemáticas (`prom_mate8b_rbd`) se explica por dos variables: proporción de estudiantes prioritarios (`prop_prio`) y cantidad de estudiantes del establecimiento (`mat_total`).

   Calcule el sesgo en que usted incurriría al estimar la relación entre la proporción de estudiantes prioritarios y la variable dependiente, utilizando un modelo de regresión lineal simple que no considere la cantidad de estudiantes del establecimiento como variable independiente. (5 puntos)

2. Derive algebraicamente el sesgo en que incurriría al omitir la cantidad de estudiantes del establecimiento en su estimación anterior. Muestre el valor de cada uno de los (tres) componentes del sesgo estimado en 1.1 y discuta su contribución al sesgo en la estimación. ¿Qué puede decir de la magnitud y signo de cada uno de ellos? (10 puntos)

3. Compare los coeficientes de la regresión *corta* y *larga* utilizando el siguiente procedimiento. (i) Seleccione aleatoriamente una muestra de 1500 colegios y estime ambas regresiones (*corta* y *larga*); (ii) Repita 1.000 veces el procedimiento anterior; (iii) Por medio de un histograma que muestre la distribución de los 1.000 coeficientes obtenidos para cada modelo estimado, compare el sesgo en su estimación de la regresión *corta* vs. *larga*. *Recomendación:* Para una visualización más eficiente de la situación, puede utilizar un *kernel density*, que corresponde a la transformación continua de un histograma. Para más información de cómo implementarlo en `R` puede revisar el siguiente Link. (10 puntos)

2. Número de profesores y Puntaje SIMCE (25 puntos)

Para esta pregunta analizaremos la relación entre la tasa de profesores por estudiante y el puntaje estandarizado en SIMCE matemáticas. Para esto deberá agregar a su base de datos `prog_educ.dta` que incorpora la tasa de profesores por estudiante (`n_prof`) y utilice las variables disponibles adecuadamente para cada pregunta.

1. Considerando que Ud. cuenta con una serie de variables que podrían afectar el puntaje SIMCE de un establecimiento. Estime por medio de diferentes modelos propuestos la relación entre la tasa de profesores por estudiante y el puntaje estandarizado en SIMCE matemáticas. Construya una tabla que resuma sus resultados y discuta la relación entre ambas variables a partir de los resultados resumidos en la tabla. Utilice **al menos cuatro** modelos distintos para su estimación. (10 puntos)

2. Escoja un modelo preferido dentro de los estimados anteriormente y justifique su elección por algún criterio que estime conveniente. ¿Cómo interpretaría el valor del coeficiente de la tasa de profesores por estudiante? (5 puntos)

3. Discuta las razones de por qué el coeficiente de la tasa de profesores por estudiante podría o no estar sesgado, respecto de su efecto causal sobre el puntaje SIMCE. Comente **al menos dos** fuentes de sesgo y sea específico respecto de las posibles amenazas que esto traería. Intente dar ejemplos de cómo afectaría (o no) las estimaciones que ha realizado. (10 puntos)

3. Evaluación de un programa (35 puntos)

Para esta pregunta usted debe evaluar el impacto de un programa (ficticio) que se implementó y respecto del cual nos interesa conocer su efecto. Los establecimientos beneficiados por el programa están identificados por la variable `treat`. Como el programa fue implementado en el año 2021, Ud debiera evaluar su impacto utilizando los datos del SIMCE 2021 (`simce_mat21`), también disponible en la base de datos `prog_educ.dta`.

1. Haga una tabla de balance que compare algunos atributos de aquellos que recibieron el programa (`treat` $== 1$) y quienes no (`treat` $== 0$). Muestre estadística descriptiva de ambos grupos.

   ¿Qué podría concluir respecto de la comparación entre ambos grupos? Además, determine si las diferencias entre las variables observables de estos son estadísticamente significativas. (15 puntos)

   *** Variables a incluir en el balance**: tipo de establecimiento (particular pagado, municipal o particular subvencionado), matrículas totales (`mat_total`) y separadas por totales de hombres y mujeres, área geográfica en que se ubica el establecimiento (`rural_rbd`), región del establecimeinto, vigencia del convenio SEP (`convenio_sep`), grupo socioeconómico (`cod_grupo`), proporción de estudiantes prioritarios para el año 2019 (`prop_prio`), puntaje simce matemáticas y lenguaje. Adapte las variables en forma **conveniente** para su análisis y detalle en el código la adaptación realizada.

2. Estime **al menos cuatro** modelos de regresión lineal distintos para evaluar el impacto del programa en relación a su efecto en puntaje SIMCE matemáticas y contruya una tabla similar a la que construyó en 2.1. Discuta la estabilidad del coeficiente de interés ¿Bajo qué supuesto podría interpretar el coeficiente de esa regresión como el impacto causal del programa? Analice el valor del coeficiente y su nivel de significancia. (10 puntos)

3. Suponga que el programa fue asignado en forma aleatoria a los establecimientos y que en la práctica aumentó el número de horas laborales docentes equivalentes a aumentar en 0.04 la cantidad de profesores por estudiante. ¿Cómo utilizaría esa información para describir el efecto causal de la tasa de profesores por estudiante y el puntaje SIMCE? (10 puntos)