Ayudantes:
Axel Canales
Matilde Cerda

EAE310: Microeconometría Aplicada

Prof. Miriam Artiles

Tarea 1

2025-03-15

Esta tarea se deberá realizar de forma **individual**. Además de un informe en PDF, deberán entregarse todas las bases de datos y los correspondientes archivos `.do` y `.log` para poder replicar los resultados. No se evaluará la entrega si falta alguno de estos archivos. Tenga en cuenta que, en caso de usar comandos de Stata que deben ser instalados por el usuario, el archivo `.do` deberá incluir el código necesario para instalarlos (`ssc install`). La entrega se realizará a través de la plataforma del curso en **CANVAS**, en el buzón habilitado para ello a más tardar el 6 de abril, 23:59 horas. En caso de error con CANVAS, la entrega se realizará mediante correo (`miriam.artiles@uc.cl`) con copia a los ayudantes. Bajo ninguna circunstancia se aceptarán entregas después de la hora/fecha límite establecida en el calendario (sin excepciones). Bajo ninguna circunstancia se admitirán casos de copia/plagio. Sea profesional a la hora de redactar el informe PDF:

- Cuando la pregunta se refiera a la ejecución de un comando de Stata, puede presentar una captura de pantalla del comando ejecutado. Redacte la secuencia de comandos usados y sea claro en sus interpretaciones cuando la pregunta lo especifique.
- Cuando la pregunta le pida reportar resultados, presente estos resultados usando gráficos y tablas profesionales (similares a los papers estudiados en clase/ayudantía). Para esto puede emplear LaTeX [opcional].
- Otra opción interesante es presentar los resultados usando un Jupyter Notebook o Markstat *[Avanzado, opcional]*. Algunos recursos para esto:
  - https://www.stata.com/features/overview/jupyter-notebooks/
  - https://kylebarron.dev/stata_kernel/
  - https://grodri.github.io/markstat/

Tenga en cuenta:

- Si el código tiene errores o los archivos proporcionados no permiten replicar los resultados, se restarán 10 puntos de la nota del ejercicio correspondiente (sobre un total de 50 puntos por cada ejercicio).
- Cualquier código que llegue a la misma solución se dará por válido (hay varias formas de llegar a los mismos resultados).
- Se debe adjuntar un informe en **PDF** por separado, fuera del archivo comprimido.
- El archivo comprimido debe contener el `.do` file, `.log` y las bases de datos necesarias para ejecutar el código correctamente.
- Si el buzón en **Canvas** no funciona, la tarea debe enviarse por correo a la profesora con copia a los ayudantes, siempre antes de la fecha y hora límite.
- **No se aceptarán tareas enviadas después del plazo, incluso por correo.**

1. Encuesta Nacional de Empleo (ENE) y Encuesta Suplementaria de Ingreso (ESI)

1. En el sitio Estadísticas de Ocupación y Desocupación del INE, descargue la base de datos de la ENE para el trimestre octubre-noviembre-diciembre (OND) de 2023 (`ene-2023-11-ond.dta`) y el documento de metadatos más reciente disponible (e.g., Encuesta Nacional de Empleo - Año de referencia 2023, o 2020-2021 si no hay actualización).
2. **(2 pts.)** Con base en el documento de metadatos descargado, explique brevemente en qué consiste la Encuesta Nacional de Empleo (ENE), respondiendo: ¿qué tipo de información contiene la encuesta? ¿cuál es su unidad de análisis? ¿qué cobertura tiene? ¿cuál es su universo? ¿cuántas variables contiene la base de datos `ene-2023-11-ond.dta`?
3. **(2 pts.)** Basándose en el documento de metadatos, describa brevemente el procedimiento de muestreo de la ENE. Además, explique ¿qué es un factor de expansión y para qué se utiliza? ¿qué variable representa el factor de expansión en la base de datos de la ENE?

4. **(1 pt.)** Seleccione únicamente las siguientes variables de la base de datos `ene-2023-11-ond`: `est_conyugal, nivel, sexo, tramo_edad, edad, region, tipo, activ, cine, obe, tpi, ftp, habituales, efectivas, fact_cal, asocia, sector, orig1, id_identificacion, idrph,` y guarde el resultado en un archivo llamado `ene-2023-ond_delim.dta`.

5. **(1 pt.)** Genere la variable `ponderador` que redondee los valores de la variable que represente el factor de expansión al entero más cercano.

6. **(2 pts.)** Construya la variable dummy `ocupados` que tome el valor 1 si el individuo está ocupado y 0 en caso contrario. Luego, usando el factor de expansión, calcule las variables que representen el número total de ocupados por sexo, estado conyugal y tipo de estrato, respectivamente.

7. **(2 pts.)** Genere la variable dummy `desocupados` que tome el valor 1 si el individuo está desocupado y 0 en caso contrario. Luego, calcule las variables que representen el número total de desocupados por sexo, estado conyugal y tipo de estrato, ajustados por el factor de expansión.

8. **(2 pts.)** La fuerza de trabajo se define como la suma de ocupados y desocupados. Construya la variable `fuerza_trabajo` y luego tres distintas variables que muestren el total de la fuerza de trabajo por sexo, estado conyugal y tipo de estrato, ajustando por el factor de expansión.

9. **(3 pts.)** La población en edad de trabajar se define como los individuos cuya edad es mayor o igual a 15 años. Cree una variable dummy que sea igual a 1 si los individuos tienen la edad de trabajar y 0 en caso contrario. Luego, usando el factor de expansión, construya una variable que calcule el total de la población en edad de trabajar llamada `mayoresde15`.

10. La tasa de participación laboral se define como: $Tasa\ de\ partipacion = \left(\frac{fuerza\ de\ trabajo}{mayores\ de\ 15}\right) * 100$. Calcule las tasas de participación laboral por sexo, estrato y estado conyugal usando la fuerza de trabajo y el calculo del total de la población en edad de trabajar generados en incisos anteriores. Use el formato de 2 decimales para porcentajes

11. **(4pts.)** Reporte los estadísticos de de tasa de participación en una tabla y en tres gráficos de barras uno por cada categoría (sexo, estado conyugal y tipo de estrato) ***Hint:*** *Recuerde filtrar los valores faltantes o respuestas como* ***No sabe*** *o* ***No responde*** *si existen.*

12. **(2 pts.)** La tasa de desempleo se calcula como $Tasa\ de\ Desempleo = \left(\frac{Desocupados}{Fuerza\ de\ trabajo}\right) * 100$. Reporte en una tabla y muestre un gráfico de barra el top 5 de regiones que tienen mayor desempleo.

    La Encuesta Suplementaria de Ingresos (ESI), es un módulo complementario que se aplica dentro de la Encuesta Nacional de Empleo (ENE). Se levanta una vez al año durante el trimestre octubre-diciembre en todas las regiones de Chile, tanto en zonas urbanas como rurales. Su objetivo es caracterizar los ingresos laborales de las personas que son clasificadas como ocupadas en la ENE, así como los ingresos de otra(s) ocupación(es) distinta(s) de la ocupación principal, tanto a nivel nacional como regional.

13. **(1pt.)** Descargue desde la página web del INE la Encuesta Suplementaria de Ingresos del 2023 y combínela con `ene-2023-ond_delim.dta` en una sola base de datos, reportando el resultado del merge (cantidad de coincidencias en ambas bases, etc.) De la ESI conserve únicamente las siguientes variables: `ing_t_t ing_ot ing_t_d ing_t_p`, y guarde el resultado en un archivo llamado `tarea1_parte1_nombre_apellido.dta`.

14. **(2pt.)** Elimine de la base las observaciones para las que los últimos dos dígitos de `idrph` es un número múltiplo de su cumpleaños más 20. (por ejemplo, si nació el 18 de julio, elimine las observaciones con `idrph` terminado en 38).

15. **(4 pts.)** Para efectos de esta tarea tome como factor de expansión de ahora en adelante el de la ESI_2022. Calcule y presente en un gráfico de barras el ingreso medio de las personas por sector (formal o informal).

16. **(3 pts.)** Genere un histograma que incluya la densidad kernel para la variable de ingreso total del trabajo (ing_t_t). A continuación, realiza un análisis descriptivo de la distribución observada, identificando posibles anomalías. Finalmente, proponga una estrategia específica para mitigar el potencial problema.

17. **(3 pts.)** Cree una variable *dummy* que sea igual 1 si el individuo posee educación universitaria aprobada y cero en caso contrario. Estime mediante MCO una regresión del logaritmo del ingreso total del trabajo sobre la dummy anterior, ajustando por el factor de expansión. Interprete el resultado.

18. **(3 pts.)** Añada las siguientes variables de control de una en una (i.e. comience añadiendo solo la primera, después las dos primeras, etc.):

a) Edad
b) Edad al cuadrado
c) Sexo
d) Región
e) Estado conyugal
f) *Dummy* de perteneciente a un pueblo originario.
g) Tipo de estrato (ciudad/rural).
h) Horas habituales semanales de trabajo.
i) Persona ocupada según sector (formal/informal).
j) Participante en al menos una organización de trabajadores

19. **(4 pts.)** Describa los resultados. ¿Cómo cambia el coeficiente estimado de educación universitaria al añadir las variables de control? ¿Qué controles parecen más importantes desde el punto de vista estadístico? ¿Cómo de robusta le parece la relación entre educación universitaria e ingreso? Según su criterio ¿se cumple el supuesto de exogeneidad? ¿Por qué si o no?

2. Pruebas de admisión a la educación superior

Descargue desde la pagina web Datos Abiertos del Ministerio de Educacion, las siguientes bases de datos:

- Pruebas de admisión a la educación superior - Prueba de transicion Universitaria - Inscritos Puntajes 2022.
- Pruebas de admisión a la educación superior - Prueba de transicion Universitaria - Datos socioeconomicos 2022.
- Pruebas de admisión a la educación superior - Prueba de transicion Universitaria - Matricula 2022.

Incluye en un `do.file` el código necesario para replicar los siguientes ejercicios:

1. **(3 pts.)** Explique brevemente (150 palabras como máximo) en qué consisten los registros administrativos del Ministerio de Educación (MINEDUC) utilizados en esta tarea. Para ello, responda las siguientes preguntas: ¿Qué tipo de información contienen estos registros? ¿Cuál es su unidad de análisis?
2. **(3 pts.)** Realiza un `merge` entre las bases de datos de Inscritos Puntajes y Datos Socioeconómicos correspondientes al año 2022. Guarda la base de datos resultante con el nombre `puntaje_economico.dta`.
3. **(1 pts.)** Verifica que la base final contenga todas las variables esperadas y muestra el número de observaciones con `summarize`.
4. **(2 pts.)** Identifica y elimina observaciones duplicadas, así como aquellas que contengan valores perdidos.
5. **(2 pts.)** Crea una variable dummy que tome el valor 1 si al menos uno de los padres tiene como mínimo el grado de profesional.
6. **(2 pts.)** Recodifica las siguientes variables numéricas a formato categórico: `DEPENDENCIA`, `TIENE_TRABAJO_REM`, `ECONOMICAMENTE` y `SEXO`.
7. **(3 pts.)** Analiza si existe una diferencia en el Puntaje estándar asignado al promedio general de notas, `ptje_nem` entre estudiantes cuyos padres tienen educación superior y aquellos cuyos padres no la poseen.
8. **(2 pts.)** Determina cuántos estudiantes existen en la base de datos con padres con educación superior y sexo femenino.
9. **(4 pts.)** Calcula el puntaje promedio en la Puntaje estándar asignado al promedio general de notas, `ptje_nem` para cada tipo de colegio (`DEPENDENCIA`), junto con la desviación estándar y el número de observaciones.
10. **(4 pts.)** Crea un boxplot para comparar el puntaje NEM entre distintos tipos de colegios (públicos, subvencionados y privados).

11. **(3 pts.)** Crea una nueva variable `ingreso` que excluya la categoría "Prefiero no responder" y asigne a cada observación la mediana del intervalo de ingreso correspondiente. Por ejemplo, para la categoría 1, el ingreso per cápita se calcula como $(71,093 + 0)/2$. Si alguien esta en la categoria 10 suponga un ingreso per capita de 1,000,000.

12. **(3 pts.)** Genera un gráfico utilizando el comando `binscatter` que muestre la relación incondicional entre el ingreso familiar del estudiante(variable creada en el inciso anterior) y su Puntaje estándar asignado al promedio general de notas (`ptje_nem`).

13. **(2 pts.)** Estima una regresión mediante el método de mínimos cuadrados ordinarios (MCO) entre el puntaje estándar asignado al promedio general de notas (`PTJE_NEM`) y el ingreso familiar. Interpreta los resultados.

14. **(2 pts.)** Reestima la regresión anterior incorporando efectos fijos por establecimiento de graduación (`RBD`). ¿Es el estimador asociado al ingreso familiar del inciso anterior robusto a controlar por efectos fijos?

15. **(3 pts.)** Agrega un conjunto de variables de control relacionadas con características individuales en la regresion anterior y analiza cómo cambia la interpretación de los resultados. ¿Se puede dar una interpretación causal a los resultados obtenidos?

16. **(4 pts.)** Realiza un `merge` entre las bases de datos `matricula` y `puntaje_economico`. Guarde la nueva base de datos en un archivo `.dta` llamado `tarea1_parte2_nombre_apellido.dta`. ***Hint:*** *Tenga en cuenta que un alumno puede estar inscrito en más de una universidad.*

17. **(2 pts.)** Elimina todas las observaciones de estudiantes que estén matriculados en más de una universidad.

18. **(2 pts.)** Crea una variable dummy llamada `elite` que tome el valor de 1 si el alumno se encuentra inscrito en la Universidad Catolica o Universidad de Chile. De lo contrario, que tome el valor de cero si se encuentra matriculado en cualquier otra universidad o no se encuentra matriculado en un instituto de educacion superior. Etiquete dicha variable.

19. **(4 pts.)** En un mismo gráfico, presenta la densidad de la variable `ptje_nem` utilizando `kdensity`, diferenciando entre estudiantes matriculados en una universidad de élite y aquellos que no lo están.

20. **(4 pts.)** Calcula la proporción de estudiantes matriculados en universidades de élite según si al menos uno de sus padres tiene educación profesional o superior. Presenta los resultados en una tabla de frecuencias y en un gráfico de barras apiladas.