TAREA 2

INDICACIONES:

- Dentro de su repositorio en Github, crear un folder llamado hmw2. Todos los archivos necesarios para realizar esta tarea deben estar dentro de este folder: código, base de datos, etc.
- Para resolver el problema 1, crear un archivo en `Jupyter Notebook`, y llámelo hmw2_1. Este problema debe desarrollarse con `python`. Este archivo debe llamar a la base de datos necesaria para correr el ejercicio.
- Para resolver el problema 2, crear un archivo en `Jupyter Notebook`, y llámelo hmw2_2. Este problema debe desarrollarse con `python`, específicamente use la libreria numpy. Las respuestas a preguntas teóricas deben estar en el mismo archivo. Para ello use MarkDown.
- Para resolver el problema 3, crear un archivo en `Jupyter Notebook`, y llámelo hmw2_3. Este problema debe desarrollarse con `stata`. Las respuestas a preguntas teóricas deben estar en el mismo archivo.
- Para resolver el problema 4, crear un archivo en `Jupyter Notebook`, y llámelo hmw2_4. Este problema debe desarrollarse con `stata`. Las respuestas a preguntas teóricas deben estar en el mismo archivo.
- Usted deberá realiar el PUSH a la carpeta github ECOP2037_NN, antes de la fecha indicada en esta tarea 2.

1. Simulación de Monte Carlo: Evaluación del Estimador MCO y del Estadístico t

El análisis de Monte Carlo simula una gran cantidad de muestras del modelo para estudiar la distribución en muestras finitas de los estimadores. En este ejercicio, utilizamos esta técnica para confirmar dos resultados en muestras finitas presentados en clases: la insesgadez del estimador de coeficientes MCO y la distribución del cociente t. El modelo es el siguiente modelo de regresión simple que satisface las suposiciones estudiadas en clase con $n = 32$. La ecuación de regresión es:

$$y_i = \beta_1 + \beta_2 x_i + \varepsilon_i \quad (i = 1, 2, \ldots, n)$$

donde $\mathbf{X} = (\mathbf{1} : \mathbf{x})$ y $\boldsymbol{\beta} = (\beta_1, \beta_2)'$. Los parámetros del modelo son $(\beta_1, \beta_2, \sigma^2)$. Elegimos una distribución particular al especificar el modelo de regresión como sigue: fijamos $\beta_1 = 1$, $\beta_2 = 0{,}5$ y $\sigma^2 = 1$. La distribución de $\mathbf{x} = (x_1, x_2, \ldots, x_n)'$ se especifica mediante el siguiente proceso autorregresivo de orden 1 AR(1):

$$x_i = c + \phi x_{i-1} + \eta_i \quad (i = 1, 2, \ldots, n)$$

donde $\{\eta_i\}$ son variables aleatorias i.i.d. $N(0, 1)$ y

$$x_0 \sim N\left(\frac{c}{1-\phi}, \frac{1}{1-\phi^2}\right), \quad c = 2, \quad \phi = 0{,}6$$

Esto fija la distribución conjunta de $(\mathbf{y}, \mathbf{X})$. A partir de esta distribución se generarán muchas muestras. Para programar la simulación, la siguiente expresión de $\mathbf{x}$ será útil. Se resuelve la ecuación en diferencias de primer orden anterior:

$$\begin{aligned} x_i = \phi^i x_0 + \left(1 + \phi + \phi^2 + \cdots + \phi^{i-1}\right) c \\ + \left(\eta_i + \phi\eta_{i-1} + \phi^2\eta_{i-2} + \cdots + \phi^{i-1}\eta_1\right) \end{aligned}$$

o en notación matricial:

$$\underset{(n\times 1)}{\mathbf{x}} = \underset{(n\times 1)}{\mathbf{r}} \cdot x_0 + \underset{(n\times 1)}{\mathbf{d}} + \underset{(n\times n)}{\mathbf{A}} \underset{(n\times 1)}{\boldsymbol{\eta}}$$

donde:

$$\mathbf{d} = (d_1, d_2, \ldots, d_n)', \quad d_i = \left(1 + \phi + \cdots + \phi^{i-1}\right) c$$

$$\mathbf{r} = \begin{bmatrix} \phi \\ \phi^2 \\ \vdots \\ \phi^n \end{bmatrix}, \quad \mathbf{A} = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ \phi & 1 & \cdots & 0 \\ \phi^2 & \phi & 1 & \cdots \\ \vdots & \vdots & \ddots & \vdots \\ \phi^{n-1} & \phi^{n-2} & \cdots & 1 \end{bmatrix}, \quad \boldsymbol{\eta} = \begin{bmatrix} \eta_1 \\ \eta_2 \\ \vdots \\ \eta_n \end{bmatrix}$$

1. Inicializar un contador.
2. Repetir por muchas iteraciones:
   a) Generar una nueva $\mathbf{x}$ (nuevo x_0 y nueva $\boldsymbol{\eta}$).
   b) Generar $\boldsymbol{\varepsilon} \sim N(0, I)$ y calcular $\mathbf{y}$.
   c) Estimar el modelo y calcular el estadístico t para $H_0 : \beta_2 = 0{,}5$.
   d) Si $|t| > t_{0{,}025}(n-2)$, incrementar el contador.
3. Dividir el contador entre el número de repeticiones (frecuencia de rechazo incondicional).

Preguntas sobre la simulación:

- ¿Cuál es la media de β_1 y β_2?
- ¿Cuál es la frecuencia de rechazar H_0?.

Preguntas teóricas

- ¿El regresor no constante es estrictamente exógeno?
- ¿El término de error es condicionalmente homocedástico?

2. Algebra de MCO: continuación del problema 2 (HMW 1)

2.1. Leverage Values

1. Obtenga el valor h_{ii} para cada una de las observaciones $p_{ii} = X_i' \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} X_i$
2. Muestre que: i) $0 \leq p_{ii} \leq 1$, ii) $p_{ii} \geq 1/n$ si X incluye un intercepto, y iii) $\sum_{i=1}^{n} p_{ii} = k$

3. Para cada una de las observaciones, obtenga el estimador LOO:

$$\hat{\beta}_{(-i)} = \left(\sum_{j \neq i} \mathbf{x}_j \mathbf{x}_j' \right)^{-1} \left(\sum_{j \neq i} \mathbf{x}_j y_j \right)$$

Note que por cada observación que saque, usted obtendrá los $\hat{\beta}_0$, $\hat{\beta}_1$, $\hat{\beta}_2$. En total tendrá 15 estimaciones!

4. Para cada observación obtenga: $\tilde{y}_i = x_i' \hat{\beta}_{(-i)}$. Muestre los valores.
5. Para cada observación obtenga: $\tilde{e}_i = y_i - \tilde{y}_i$. Muestre los valores.
6. Muestre que $\widetilde{e_i} = (1 - p_{ii})^{-1} \widehat{e_i}$. Muestre que los valores. conciden con aquellos del literal 5.

2.2. Matrices de Varianza y Covarianza

7. Obtenga el estimador de la varianza de lo erroes corregida por sesgo: $s^2 = \frac{1}{n-k} \sum_{i=1}^{n} \hat{e}_i^2$. Muestre los valores.
8. Obenga el estimador de la varianza de los errores standarizados: $\bar{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (1 - h_{ii})^{-1} \hat{e}_i^2$. Muestre los valores.
9. Obtenga la matriz de varianza y covarianza bajo homocedasticidad $\tilde{\boldsymbol{V}}_{\beta}^{Q} = (\boldsymbol{X}'\boldsymbol{X})^{-1} s^2$. Muestre los valores.
10. Obtenga el estimador de la matriz de varianza y convarianza consistente para heterocedasticidad HC0: $\widehat{\boldsymbol{V}}_{\beta}^{\mathrm{HC}} = (\boldsymbol{X}'\boldsymbol{X})^{-1} \left(\sum_{i=1}^{n} X_i X_i' \hat{e}_i^2 \right) (\boldsymbol{X}'\boldsymbol{X})^{-1}$. Muestre los valores.
11. Obtenga el estimador de la matriz de varianza y convarianza consistente para heterocedasticidad HC1 $\hat{\boldsymbol{V}}_{\boldsymbol{\beta}}^{\mathrm{HCl}} = \left(\frac{n}{n-k} \right) (\boldsymbol{X}'\boldsymbol{X})^{-1} \left(\sum_{i=1}^{n} X_i X_i' \hat{e}_i^2 \right) (\boldsymbol{X}'\boldsymbol{X})^{-1}$. Muestre los valores.
12. Obtenga el estimador de la matriz de varianza y convarianza consistente para heterocedasticidad HC2 $\hat{\boldsymbol{V}}_{\boldsymbol{\beta}}^{\mathrm{HC2}} = (\boldsymbol{X}'\boldsymbol{X})^{-1} \left(\sum_{i=1}^{n} (1 - h_{ii})^{-1} X_i X_i' \hat{e}_i^2 \right) (\boldsymbol{X}'\boldsymbol{X})^{-1}$. Muestre los valores.
13. Obtenga el estimador de la matriz de varianza y convarianza consistente para heterocedasticidad HC3 $\hat{\boldsymbol{V}}_{\boldsymbol{\beta}}^{\mathrm{HC3}} = (\boldsymbol{X}'\boldsymbol{X})^{-1} \left(\sum_{i=1}^{n} (1 - h_{ii})^{-2} X_i X_i' \hat{e}_i^2 \right) (\boldsymbol{X}'\boldsymbol{X})^{-1}$. Muestre los valores.

3. Modelo de regresión simple

Sea $\varepsilon_1, ..., \varepsilon_n$ be $iidN \sim [0, \sigma^2]$. Considere que cada ε_i puede ser escrito como una fumción de datos observados para x_i y y_i. Es decir, $\varepsilon_i = y_i - \alpha - \beta x_i$.

1. Escriba la funcion log-likelihood function.
2. Encuentre el estimador de maxima verosimilitud para α, β and σ^2.
3. Encuentre el limite inferior Cramer-Rao.
4. Escriba el Likelihood Ratio test para $H_0 : \beta = 0$
5. Escriba un programa en Stata que le permita estimar por ML la siguiente expresion:

$$\texttt{wage} = \beta_1 \texttt{education} + \beta_2 \texttt{experience} + \beta_3 \texttt{exp2} + \beta_4 + e$$

Para ello, use la base de datos de la Encuesta de Población Actual (CPS), provista en HMW 1. A continuación entrego una versión incompleta del codigo que debe realizar.

```
capture program drop myLikelihood
program myLikelihood
args logl b1 b2 b3 b4 sig
tempvar xb stde p

quietly gen double ‘xb’=‘b1’*education + ‘b2’*experience + ‘b3’*exp2 + ‘b4’
quietly gen double ‘stde’ = (wage-‘xb’)/‘sig’
quietly gen double ‘p’=normalden(‘stde’)/‘sig’
quietly replace ‘logl’=ln(‘p’)
end
```

Recuerde que su programa debe llamar la base de datos, y realizar la estimación.

4. Efecto del Tracking Escolar y Inferencia con Muestras Agrupadas

En contextos donde las unidades de observación están naturalmente agrupadas (como estudiantes dentro de escuelas), la independencia entre observaciones puede no cumplirse. Esto afecta la inferencia estadística, ya que los errores estándar convencionales pueden subestimar la varianza real del estimador. En este ejercicio trabajaremos con datos de un experimento aleatorizado en escuelas de Kenia (Duflo, Dupas y Kremer, 2011), que evaluó el efecto de dividir a los estudiantes en clases según su nivel de rendimiento inicial (“tracking”).

Objetivo

Replicar la estimación del efecto promedio del tracking sobre el rendimiento escolar con inferencia adecuada para datos agrupados.

Base de datos

Utilice el archivo `DDK2011.dta`. La variable dependiente es el puntaje estandarizado `TestScore`, y la variable de tratamiento es `Tracking`, que indica si la escuela aplicó tracking.

Parte I: Estimación básica

1. Estime el siguiente modelo por Mínimos Cuadrados Ordinarios (OLS):

$$\texttt{TestScore}_{ig} = \alpha + \gamma \cdot \texttt{Tracking}_{g} + e_{ig}$$

2. Reporte el valor de $\hat{\gamma}$ y su error estándar usando:
   (a) Errores estándar convencionales (homocedásticos).
   (b) Errores estándar robustos agrupados por escuela.
3. Comente las diferencias. ¿Qué implicancias tiene para la significancia estadística del efecto estimado?

Parte II: Desafíos adicionales

Challenge 1: Robustez con controles individuales

Agregue las siguientes variables de control al modelo anterior:

- Edad del estudiante.
- Género.
- Puntaje inicial.

¿Cambia la magnitud o significancia de $\hat{\gamma}$? ¿Por qué?

Challenge 2: Heterogeneidad del efecto del tracking

Cree una variable `BottomHalf` que sea 1 si el estudiante estaba en la mitad inferior del puntaje inicial en su escuela.

Estime:

$$\texttt{TestScore}_{ig} = \alpha + \gamma_1 \cdot \texttt{Tracking}_g + \gamma_2 \cdot \texttt{BottomHalf}_{ig} + \gamma_3 \cdot (\texttt{Tracking}_g \times \texttt{BottomHalf}_{ig}) + e_{ig}$$

Interprete el coeficiente $\hat{\gamma}_3$. ¿Es diferente el efecto del tracking para estudiantes de menor rendimiento inicial?

Challenge 3: Inferencia errónea por no agrupar

Compare los errores estándar de $\hat{\gamma}$ usando:

- Errores convencionales.
- Errores robustos de White (HC0)
- Errores robustos agrupados por escuela.

¿En cuál caso cambia la significancia estadística del efecto de tracking?