TAREA 1

INDICACIONES:

- Dentro de su repositorio en Github, crear un folder llamado hmw1. 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 hmw1_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 hmw1_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 hmw1_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 hmw1_4. Para escribir las respuestas a estas preguntas usen MarkDown.
- Usted deberá realiar el PUSH a la carpeta github ECOP2037_NN, antes de la fecha indicada en esta tarea 1.

1. Análisis de diferencias en log-salarios por género y raza

La Encuesta de Población Actual (CPS) es una encuesta mensual de alrededor de 57.000 hogares estadounidenses realizada por la Oficina del Censo de Estadísticas Laborales. La CPS es la principal fuente de información sobre las características de la fuerza laboral de la población estadounidense. La encuesta cubre empleo, ingresos, nivel educativo, pobreza, cobertura de seguro médico, experiencia laboral, votación y registro, uso de computadoras, condición de veterano y otras variables. Los detalles se pueden encontrar en www.census.gov/cps. De la encuesta de marzo de 2009, trabajaremos solo con 50.742 individuos que estaban empleados a tiempo completo (definidos como aquellos que haían trabajado al menos 36 horas por semana durante al menos 48 semanas el año pasado). Estamos interesados en estudiar el siguiente grupo de variables:

- `logwage` representa el logaritmo del salario por horario,
- `gender` es una variable categórica con valores "male" y "female",
- `race` es una variable categórica con valores "white", "Black" y "other".
- `education`: años de educación formal

Ver el archivo cps09mar.xlsx.

1. Calcule la **media del log-salario** (`logwage`) para cada una de las seis combinaciones posibles de `gender` y `race`. Presente sus resultados en una tabla con el siguiente formato, e incluya una columna adicional con la diferencia entre hombres y mujeres (`men - women`) dentro de cada grupo racial:

	Men	Women	Difference
White			
Black			
Other			

Interprete brevemente las principales desigualdades observadas.

2. Grafique las funciones de **densidad empírica** del `logwage` para los seis grupos combinando género y raza. Utilice líneas diferenciadas por color o estilo para cada grupo y agregue una leyenda que identifique claramente cada uno.

3. Comente brevemente las diferencias observadas en las funciones de densidad. ¿Qué grupo tiene mayor tendencia central? ¿Hay diferencias claras en la dispersión o forma de las distribuciones?

4. Explique por qué se cumple la siguiente identidad utilizando el teorema de la esperanza iterada:

$$\begin{aligned}
&\mathbb{E}[\log(\text{wage}) \mid \text{gender} = \text{man}] = \\
&\mathbb{E}[\log(\text{wage}) \mid \text{gender} = \text{man}, \text{race} = \text{white}] \cdot \mathbb{P}[\text{race} = \text{white} \mid \text{gender} = \text{man}] \\
&+ \mathbb{E}[\log(\text{wage}) \mid \text{gender} = \text{man}, \text{race} = \text{Black}] \cdot \mathbb{P}[\text{race} = \text{Black} \mid \text{gender} = \text{man}] \\
&+ \mathbb{E}[\log(\text{wage}) \mid \text{gender} = \text{man}, \text{race} = \text{other}] \cdot \mathbb{P}[\text{race} = \text{other} \mid \text{gender} = \text{man}]
\end{aligned}$$

5. Calcule empíricamente cada uno de los términos de la derecha utilizando su base de datos. Verifique si su suma coincide con el promedio observado de log(wage) para los hombres.

6. Restrinja la muestra a individuos blancos (`race == ‘‘white’’, y ‘‘Black’’`). Para estos subgrupos:

   - Calcule la **media condicional** de `logwage` para cada año de educación, separadamente para hombres y mujeres, por raza.
   - Presente los resultados en una tabla que muestre: años de educación, media logwage para hombres blancos y negro, y media logwage para mujeres blancas y negras.

7. Grafique la media condicional de `logwage` como función de los años de educación, separadamente para hombres y mujeres, por tipo de raza. Asegúrese de:

   - Usar líneas diferenciadas por color o estilo para cada grupo.
   - Incluir una leyenda, título y etiquetas de los ejes.

8. Comente brevemente los siguientes puntos observados en el gráfico:

   - ¿La relación entre educación y log-salario parece lineal?
   - ¿Se aprecia un cambio en la pendiente para niveles de educación mayores o menores a 9 años?
   - ¿Cómo se comporta la brecha entre hombres y mujeres a lo largo del espectro educativo? ¿y por raza?
   - ¿Qué implica una brecha aproximadamente constante en términos del salario en niveles (no log)?

2. Algebra del MCO

Usted desea estimar el siguiente modelo de regresión,

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + e_i \quad \forall i = 1, ..., 5 \tag{1}$$

Es decir usted cuenta con 5 observaciones para realizar dicha estimación. A continuación se presenta los datos para estas 5 observaciones,

$$y = \begin{bmatrix} 3 \\ 1 \\ 8 \\ 3 \\ 5 \end{bmatrix} \qquad \boldsymbol{X} = \begin{bmatrix} 1 & 3 & 5 \\ 1 & 1 & 4 \\ 1 & 5 & 6 \\ 1 & 2 & 4 \\ 1 & 4 & 6 \end{bmatrix}$$

De forma matricial la ecuacion a estimar sería,

$$\boldsymbol{y} = \boldsymbol{X}\beta + e \tag{2}$$

1. Obtenga la estimación de mínimos cuadrados ordinarios: $\hat{\beta} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'y$. Indique específicamente el valor para $\hat{\beta}_0$, $\hat{\beta}_1$, $\hat{\beta}_2$.
2. Obtenga el vector de residuos $\hat{e} = y - \boldsymbol{X}\hat{\beta}$
3. Obtenga el vector de predicción $\hat{y} = \boldsymbol{X}\hat{\beta}$. ¿Cuál es la predicción del modelo para la observación 3?
4. Obtenga las matrices idempotentes $\boldsymbol{P}$ y $\boldsymbol{M}$. Muestre las propiedades que se discutieron en clases.
5. Obtenga el vector de predicción $\hat{y} = \boldsymbol{P}y$, compruebe si este resultado es igual al resultado obtenido en el literal 3).
6. Obtenga el vector de residuos $\hat{e} = \boldsymbol{M}y$, compruebe si este resultado es igual al del literal 2).

Considere ahora la matriz $\boldsymbol{M}_0 = I - \iota(\iota'\iota)^{-1}\iota'$, donde ι es un vector $n \times 1$ de 1.

7. Obtenga el producto $\boldsymbol{M}_0 y$. Con base en ese resutlado, ¿cuál es la funcion entonces de la matriz $\boldsymbol{M}_0$?
8. Ahora, premultiplique la ecuación 2 por la matrix $\boldsymbol{M}_0$, y obtenga la fórmula para el estimador de MCO de esta ecuación transformada. Ea decir,

   $$\boldsymbol{M}_0 y = \boldsymbol{M}_0 \mathbf{X}\beta + \boldsymbol{M}_0 e$$

   Llame $=\mathbf{X}_2 = \boldsymbol{M}_0\mathbf{X}$, sea β_2 el vector resultante de esta transformación.

   ¿Cuál es la dimensión del vector β? ¿Cuál es el resultado de $\boldsymbol{M}_0 e$?
9. Usando los datos dados previamente, use la formula del literal anterior para obtener las estiamciones de los betas. ¿qué puede decir con respecto a las estimacion obtenidas previamente en el literal 1?
10. ¿Cómo puede obtener la estimación para β_0? Calcule.
11. Calcule $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{X}_1 - \hat{\beta}_2\bar{X}_2$

12. Calcule la suma de cuadrado totales $TSS = y'M_0 y$
13. Calcule la suma de cuadrado explicado $ESS = \hat{\beta}_2{}' \boldsymbol{X}'_2 M_0 \boldsymbol{X}_2 \hat{\beta}_2$
14. Calcule la suma de cuadrado explicado $RSS = \hat{e}'\hat{e}$
15. Calcule el R^2 y el R^2 ajustado

Ahora aplicaremos del teorema Frish-Waugh (TFW).

Teorema de Frisch-Waugh-Lovell (FWL)
En el modelo $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + e_i$, el estimador MCO de β_2 y los residuos MCO
1. Regresar $\boldsymbol{Y}$ sobre $\boldsymbol{X}_1$ y obtener los residuos $\widetilde{\boldsymbol{e}}_1$;
2. Regresar $\boldsymbol{X}_2$ sobre $\boldsymbol{X}_1$ y obtener los residuos $\widetilde{\boldsymbol{X}}_2$;
3. Regrese $\widetilde{\boldsymbol{e}}_1$ sobre $\widetilde{\boldsymbol{X}}_2$, obtenga estimaciones MCO $\widehat{\beta}_2$ y residuos $\widehat{\boldsymbol{e}}$.

16. Siguiendo el algoritmo propuesto en el teorema FWL, desarrolle un codigo donde se evidencia cada una de los tres paso, y finalmente se muestre la estimacion de β_2.

3. Efects de la asistencia en el examen final

La base de datos `ATTEND` contiene información sobre 680 estudiantes de un curso de microeconomía. A continuación se describen las variables disponibles:

- `attend`: número de clases asistidas (de un total de 32).
- `priGPA`: GPA acumulado previo al curso.
- `ACT`: puntaje de admisión estandarizado ACT.
- `atndrte`: porcentaje de clases asistidas (entre 0 y 100).
- `frosh`: indicador (=1) si el estudiante es de primer a no (freshman).
- `soph`: indicador (=1) si el estudiante es de segundo a no (sophomore).
- `skipped`: número de clases que el estudiante se perdi'o.
- `stndfnl`: nota estandarizada del examen final .

Un modelo para explicar el resultado estandarizado de un examen final (`stndfnl`) en términos del porcentaje de clases asistidas (`atndrte`), el promedio de calificaciones de la universidad previa y la puntuación del `ACT` es:

$$\begin{aligned}\texttt{stndfnl} =& \beta_0 + \beta_1 \texttt{atndrte} + \beta_2 \texttt{priGPA} + \beta_3 \texttt{ACT} + \beta_4 \texttt{priGPA}^2 \\ &+ \beta_5 \texttt{ACT}^2 + \beta_6 \texttt{priGPA} \cdot \texttt{atndrte} + e\end{aligned}$$

1. ¿Por qué usamos el resultado estadarizado en lugar de la nota original?
2. ¿Por qué inlcuimos $\texttt{priGPA}^2$, $\texttt{ACT}^2$, y $\texttt{priGPA} \cdot \texttt{atndrte}$?
3. Estime el modelo propuesto. Reporte los coeficientes estimados y sus errores estándar.
4. ¿Por qué no se debe interpretar directamente el coeficiente de `atndrte` en presencia de la interacción con `priGPA`? ¿A qué valor de `priGPA` corresponde esta interpretación y por qué es poco relevante en la muestra?

5. Calcule el efecto marginal de `atndrte` sobre `stndfnl` cuando `priGPA` $= 2.59$ (la media). Interprete el resultado en términos de desviaciones estandarizadas.

6. Centre la variable `priGPA` alrededor de su media y reemplace `priGPA`$\cdot$`atndrte` por (`priGPA` $-$ 2,59) $\cdot$ `atndrte`. Estime este modelo nuevamente y reporte el nuevo coeficiente de `atndrte`. Especifacmente el modelo a estimar es

$$\begin{aligned}\texttt{stndfnl} =&\beta_0 + \beta_1\texttt{atndrte} + \beta_2\texttt{priGPA} + \beta_3\texttt{ACT} + \beta_4\texttt{priGPA}^2 \\ &+ \beta_5\texttt{ACT}^2 + \beta_6(\texttt{priGPA} - 2{,}59)\cdot\texttt{atndrte} + e\end{aligned}$$

   Compare con el obtenido previamente.

   Este resultado se conoce coomo el **efecto parcial promedio (APE)**. El literal previo y este, muestran dos maneras de calcular APE, aunque es preferible el propuesto en este literal por ser un cáculo directo.

7. Derive el efecto marginal de `priGPA` sobre `stndfnl`. Evalúelo cuando `priGPA` $= 2.59$ y `atndrte` $= 82$. Interprete este resultado.

8. Centre simultáneamente `priGPA` y `atndrte` en sus medias. Especifacmente el modelo a estimar es

$$\begin{aligned}\texttt{stndfnl} =&\beta_0 + \beta_1\texttt{atndrte} + \beta_2\texttt{priGPA} + \beta_3\texttt{ACT} + \beta_4(\texttt{priGPA} - 2{,}59)^2 \\ &+ \beta_5\texttt{ACT}^2 + \beta_6\texttt{priGPA}\cdot(\texttt{atndrte} - 82) + e\end{aligned}$$

   ¿Cómo se interpreta ahora el coeficiente de `priGPA`? Este resultado se conoce como el **efecto parcial promedio (APE)** de `priGPA` sobre `stndfnl`.

9. Explique por qu'e los efectos marginales en este modelo dependen de los valores de las variables explicativas. ¿Por qu'e los APEs son una forma *resumida* de interpretar estos efectos?

10. Una visión teŕica: Suponga el modelo

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + e$$

    Reescríbalo como

$$y = \alpha_0 + \delta_1 x_1 + \delta_2 x_2 + \beta_3(x_1 - \mu_1)(x_2 - \mu_2) + e$$

    .

    Muestre que $\delta_2 = \beta_2 + \beta_3\mu_1$ y explique por qué esta reparametrización es útil para interpretar efectos parciales en el promedio. ¿Cómo interpretaría ahora los coeficientes sobre x_1 y x_2 en este nuevo modelo?¿En vez de μ_1y μ_2 puede usar otros valores de la distribución de x_1 y x_2?

4. Supuestos sobre el término de error y sus implicaciones

Considere el modelo simple lineal:

$$y = \beta x + e$$

donde $x \in \mathbb{R}$ es un regresor aleatorio con $\mathbb{E}[x^2] < \infty$, y e es el término de error.

Suponga que se sabe lo siguiente:

- $\mathbb{E}[e \mid x] = 0$
- $\mathbb{E}[e^2 \mid x] = \sigma^2(x)$, donde $\sigma^2(x)$ es una función no constante.
- x tiene una distribución simétrica alrededor de su media.
- Se sospecha que e no es independiente de x.

1. Muestre que $\mathbb{E}[x^k e] = 0$ para todo $k \in \mathbb{N}$, si x y e cumplen que $\mathbb{E}[e \mid x] = 0$ y x es no estocásticamente relacionado con e.
2. Verifique si la afirmación $\mathbb{E}[xe] = 0$ implica que $\mathbb{E}[x^2 e] = 0$. Justifique con una breve demostración o contraejemplo.
3. ¿La condición $\mathbb{E}[e \mid x] = 0$ implica que e y x son independientes? Explique.
4. ¿La condición $\mathbb{E}[xe] = 0$ implica que $\mathbb{E}[e \mid x] = 0$? Justifique su respuesta.
5. ¿La combinación $\mathbb{E}[e \mid x] = 0$ y $\mathbb{E}[e^2 \mid X] = \sigma^2$ (constante) implica independencia entre e y x? Justifique.