Tarea 2

IN3242 - Estadística
Departamento de Ingeniería Civil Industrial, Universidad de Chile
Primavera 2023

Profesor Raimundo Undurraga
Auxiliares: Matías Reyes, Leonardo Meneses, Bastián Urzúa, Brandon Galarza, Antonia Villegas, Camila Jáuregui, Francisca Monetta

Puntaje Total: xxxx puntos
Fecha límite de entrega: Jueves 26 de Octubre hasta las 23:59 hrs.

I. Ejercicios Teóricos

1. Se desea estudiar el efecto de la educación en los salarios. Se cree que el género de la persona también puede afectar a los salarios. Originalmente, se propone entonces el siguiente modelo lineal:

$$Salario_i = \beta_0 + \beta_1 Educ_i + \beta_2 Genero_i + \varepsilon_i$$

El Salario está en pesos chilenos. La educación está medida en años de escolaridad. La variable $Genero_i$ toma el valor 1 si es mujer, y 0 si no. Se asume que los errores ε_i son independientes e idénticamente distribuidos, y que $\mathbb{E}(\varepsilon_i | Educ_i, Genero_i) = 0$. Un investigador obtuvo una muestra $\{Salario_i, Educ_i, Genero_i\}_{i=1,\dots,n}$ y observó que: (i) las mujeres en promedio tienen mayores niveles de escolaridad que los hombres, sin embargo (ii) las mujeres en promedio tienen menores niveles de salarios que los hombres.

(i) Derive una expresión para β_1, y otra para β_2. **[5 puntos]**

R: Usamos la formula para medir el efecto parcial de cada regresor. $\hat{\beta}_1 = \frac{Cov(Salario, residuos_1)}{Var(residuos_1)}$, con $residuos_1$ los residuos de regresionar $Educ_i$ contra una constante y $Genero_i$.
El resultado es análogo para $\hat{\beta}_2$, es decir, $\hat{\beta}_2 = \frac{Cov(Salario, residuos_2)}{Var(residuos_2)}$, con $residuos_2$ los residuos de regresionar $Genero_i$ contra una constante y $Educ_i$.

(ii) Un analista le sugiere omitir la variable género del modelo y estimar un modelo univariado de salarios sobre educación. Derive una expresión para β_1. Lo anterior implicaría que el estimador β_1 sobreestima el verdadero impacto de la educación sobre los salarios. Comente. **[5 puntos]**

R: El supuesto del modelo indica que $\mathbb{E}(\varepsilon_i | Educ_i, Genero_i) = 0$, es decir, el error es ortogonal a educación y genero cuando ambas variables entran en la regresión. En efecto, si omitimos la variable genero, va a presentarse un problema de variable omitida, i.e., el error ya no es ortogonal a educación.

$$\hat{\beta}_1 = \frac{Cov(Educ, Salario)}{Var(Educ)} = \beta_1 + \hat{\beta}_2 \cdot \frac{Cov(Educ, Genero)}{Var(Educ)}$$

Notar que (i) las mujeres en promedio tienen mayores niveles de escolaridad que los hombres, i.e., $Cov(Educ, Genero) > 0$. Sin embargo, (ii) las mujeres en promedio tienen menores niveles de salarios que los hombres, i.e., el impacto de ser mujer sobre los salarios es negativo ($\beta_2 < 0$). Luego, $\hat{\beta}_1 < \beta_1$, es decir, $\hat{\beta}_1$ subestima el verdadero impacto de la educación sobre los salarios.

(iii) Otro analista le sugiere conservar la variable genero, pero además agregar la variable Edad al modelo original. Derive una expresión para β_1, y otra para β_2. Cómo cambiaría su respuesta con respecto a (i)? **[5 puntos]**

R: Respuesta no cambia con respecto a pregunta (i). Notar que $\mathbb{E}(\varepsilon_i|Educ_i, Genero_i) = 0$, es decir, el error es ortogonal a educación y genero cuando ambas variables entran en la regresión. En otras palabras, no hay nada en el error que correlacione con educación y/o genero, incluida la edad, por lo tanto la covarianza de educación y genero con edad será 0, obteniéndose así la misma expresión que en (i).

(iv) El mismo colega de la pregunta anterior le indica que la gracia de agregar edad al modelo es que reduce la suma de los errores al cuadrado, lo cual mejora el R^2 del modelo. Comente **[5 puntos]**

R: La edad es una aproximación de la experiencia. Gente con mas experiencia tiene mayores salarios. En efecto, agregar edad al modelo hace que la variabilidad de los salarios sea menos explicada por la variabilidad de los errores, mejorando así la capacidad predictiva del modelo (R^2)

(v) Otro analista le sugiere que la educación tiene efectos no lineales sobre el salario, y por tanto le propone modificar la forma funcional del modelo, y usar la siguiente:

$$Salario_i = \exp\left(\beta_0 + \beta_1 Educ_i + \beta_2 Educ_i^2 + \beta_3 Genero_i + \varepsilon_i\right)$$

Puede este modelo ser estimado vía MCO? Justifique. **[5 puntos]**

R: Si puede ser estimado pues es lineal en los parámetros. Al aplicar logaritmo a la expresión está queda igual a:

$$log(Salario_i) = \beta_0 + \beta_1 Educ_i + \beta_2 Educ_i^2 + \beta_3 Genero_i + \varepsilon_i$$

Lo cual claramente es lineal en los parámetros y se puede estimar por MCO.

(vi) Un colega le sugiere invocar el supuesto de normalidad de los errores, i.e., $e_i \sim N(0, \sigma^2)$, para así poder hacer inferencia estadística sobre los parámetros del modelo. Demuestre matemáticamente la veracidad la sugerencia. **[10 puntos]**

R: Se sabe que $\hat{\beta}$ se puede escribir matricialmente como $\hat{\beta} = \beta + (X^T X)^{-1} X^T \epsilon$, es decir, su distribución depende de la distribución del error. Sabemos además que $\mathbb{E}(\hat{\beta}) = \beta$ y $Var(\hat{\beta}) = \sigma^2 \cdot (X^T X)^{-1}$, ya que el error se distribuye normal con media cero y varianza σ^2. Mas aún, dado que el error se distribuye normal, y toda variable aleatoria combinada con una normal se distribuye normal, entonces β se distribuye normal, lo cual nos permite hacer inferencia estadística sobre β.

(vii) Un colega le hace ver que en realidad el supuesto de independencia, $\mathbb{E}(\varepsilon_i|Educ_i, Genero_i) = 0$, podría no ser correcto en este modelo. En particular, le preocupa el parámetro β_1. De dos ejemplos de por qué podría fallar el supuesto de independencia, y cual sería la dirección esperada del sesgo en β_1 en cada caso. **[10 puntos]**

R: Basta con dar ejemplos de variables omitidas, como estas están contenidas en el error, se asume que son independientes con Educación y Genero, pero esto podría no ser así en

la realidad. La expresión que se espera es $\hat{\beta}_1 = \frac{Cov(Salario,Educ)}{Var(Educ)} = \beta_1 + \hat{\beta}_3 \cdot \frac{Cov(Educ,VO)}{Var(Educ)}$ Dirección del sesgo depende del signo de $\hat{\beta}_3$ y Cov(Educ,VO).

Ejemplo 1: Una posible variable omitida es el estado de salud, el cual está muy correlacionado con los años de educación. Las personas con más educación tienden a tener hábitos de vida más saludables y acceden a mejores servicios de atención médica. En este caso la $cov(Salud, salarios) > 0$ y $cov(Salud, Educ) > 0$, como ambos son mayor que 0 se espera que el sesgo también lo sea.

Ejemplo 2: Ser de región distinta a la metropolitana o no (1 si es de región distinta a la metropolitana, 0 si es de la región metropolitana). Chile es un país muy centralizado en Santiago, por lo que las mayores oportunidades de educación están en la capital. Además los mejores trabajos se encuentran en la capital, por lo que los salarios son mayores en Santiago. Luego la $cov(Region, salarios) < 0$ y $cov(Region, Educ) < 0$, por lo tanto se espera que el sesgo sea positivo.

(viii) El mismo colega de la pregunta anterior le indica que quizas sería bueno construir una nueva variable de género (e.g., Genero2), que tome el valor 0 si es mujer y 1 si no, argumentando que de esa forma podría capturar el impacto de ser hombre. En efecto, le propone estimar el siguiente modelo:

$$Salario_i = \beta_0 + \beta_1 Educ_i + \beta_2 Genero_i + \beta_3 Genero2_i + \varepsilon_i$$

Es útil la recomendación del colega? Justifique detalladamente y muestre matricialmente. **[10 puntos]**

R: No es util, pues viola el supuesto de no multicolinealidad, lo cual impide invertir la matriz de varianza-covarianzas, lo cual impide estimar el modelo.

(ix) Un colega está preocupado de testear si el modelo en su conjunto tiene capacidad predictiva de los salarios o no. Plantee un test estadístico al respecto, y detalle el paso a paso de su implementación. **[5 puntos]**

R: Se considera un modelo restringido con $\beta_1 = \beta_2 = 0$, luego el test de hipotesis es $H_0 : \beta_1 = \beta_2 = 0; H_1 : \beta_1 o \beta_2 \neq 0$. Con $g = 2$ se calcula el estadístico F. Por ultimo se rechaza la hipótesis nula si $F > c$ (c constante determinada por el nivel de significancia).

(x) Un colega le recuerda que las mujeres en promedio tienen mayores niveles de escolaridad que los hombres, y por tanto $cov(Educ_i, Genero_i) > 0$. En efecto, el colega señala que al estimar el modelo usando MCO, β_1 refleja no solo el impacto de la educación sobre los salarios, sino también indirectamente la influencia de ser mujer en los salarios. Comente y demuestre. **[10 puntos]**

R: Falso. Se mostró en la parte (i) que β_1 no depende del genero, es más, al controlar por genero el impacto se lo lleva β_2 por lo que β_1 no tiene como reflejar un efecto de genero.

(xi) Un colega señala varias apreciaciones respecto al poder estadístico del modelo para hacer inferencia sobre los parámetros. Indique en cada caso si el colega esta en lo correcto o no.

(a) la varianza de los errores reduce los errores estándar de los parámetros, lo cual mejora el poder estadístico **[5 puntos]**

R: Falso. Un aumento de σ^2 aumenta el error estándar de los betas, lo cual reduce la probabilidad de rechazar la hipótesis nula, reduciendo así el poder estadístico del modelo.

(b) Que la educación y el genero esten positivamente correlacionados no ayuda al poder estadístico, no así si estuviesen negativamente correlacionados. **[5 puntos]**

R: Falso. Independiente de la dirección de la correlación entre las variables independientes del modelo, mientras mayor sea la correlación entre ellas, mayor es el error estándar de los betas, lo cual reduce la probabilidad de rechazar la hipótesis nula, reduciendo así el poder estadístico del modelo.

(c) En la base de datos, tanto la educación como el genero muestran una alta variación, lo cual reduce el poder estadístico del modelo. **[5 puntos]**

R: Falso, mejora el poder estadístico.

II. Ejercicios Empíricos

4. Suponga que usted está interesado en modelar el comportamiento de los salarios de los **jefes de hogar** según la última encuesta CASEN 2020 (disponible en https://observatorio.ministeriodesarrollosocial.gob.cl/encuesta-casen-2022, base titulada bajo el rotulo "Base de datos Casen 2022 STATA", y para ello crea el siguiente modelo de regresión:

$$Salario_i = \beta_0 + \beta_1 Sexo_i + \beta_2 RM_i + \beta_3 Escolaridad_i + e_i$$

- $Salario_i$ = Sueldos/Salario Monetario del jefe de hogar ($y0101$)
- $Sexo_i$ = 1 si el jefe de hogar es mujer, 0 si no ($sexo$)
- RM_i = 1 si el jefe de hogar vive en la Región Metropolitana (crear variable)
- $Escolaridad_i$ = años de escolaridad del jefe de hogar (esc)

(a) Estime el modelo de regresión usando R. Según su modelo, vivir en la RM o no, influye en los salarios? Interprete orden de magnitud del parámetro y significancia estadística. Haga lo mismo para el caso del sexo **[5 puntos]**

```
Call:
lm(formula = y0101 ~ sexo + RM + e6a, data = df2)

Residuals:
     Min       1Q   Median       3Q      Max
-1196027  -285115   -80042   117741 23703973

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -460086      11464  -40.13   <2e-16 ***
sexo         -197144       4896  -40.26   <2e-16 ***
RM            163357       5900   27.68   <2e-16 ***
e6a           113768       1086  104.76   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 582300 on 58236 degrees of freedom
Multiple R-squared:  0.1851,    Adjusted R-squared:  0.1851
F-statistic:  4409 on 3 and 58236 DF,  p-value: < 2.2e-16
```

Figure 1: Resultados de la regresión

De acuerdo a los resultados entregados por la regresión se interpreta lo siguiente:

La variable "RM", que indica si el individuo reside en la región metropolitana (1 si vive, 0 si no), se destaca con tres asteriscos, indicando significancia con un p-valor menor a 0.001. El coeficiente (beta) asociado a esta variable es positivo, aproximadamente 163 mil. Este resultado sugiere que vivir en la región metropolitana está positivamente relacionado con el salario del jefe de hogar, con un aumento estimado de 163 mil unidades monetarias.

En relación al género ("sexo"), donde 1 representa mujer y 0 hombre, esta variable también muestra significancia, con un p-valor menor a 0.001 y un coeficiente negativo de aproximadamente 197 mil. Esto indica que ser mujer tiene un impacto negativo en el salario, disminuyéndolo en alrededor de 197 mil unidades monetarias.

(b) Cuánto de la variabilidad de los salarios se explican por la variabilidad de las variables del modelo? Construya el indicador utilizando SST, SSR y SSE.[**10 puntos**]

El coeficiente de determinación, comúnmente conocido como R-cuadrado, es un indicador clave en análisis de regresión que nos permite entender cuánta variabilidad en los salarios se explica por las variables incluidas en el modelo de regresión. En este casos los valores obtenidos son:

SST (Suma de Cuadrados Total): $2.423093e^{16}$
SSR (Suma de Cuadrados de la Regresión): $4.485131e^{15}$
SSE (Suma de Cuadrados del Error): $1.97458e^{16}$
R^2: 0.1850994

El R^2 obtenido a través de los cálculos de la suma total de cuadrados (SST) y la suma de cuadrados del error (SSE) es muy similar al valor proporcionado por la regresión. Como se mencionó anteriormente, el R^2 ofrece información acerca de la proporción de la varianza de la variable dependiente que es explicada por el modelo de regresión. En este caso específico, el R^2 indica que aproximadamente el 18.5% de la variabilidad de la variable dependiente es explicada por las variables predictoras incluidas en el modelo

(c) Construya un test F para testear si todos los parámetros del modelo son estadísticamente distintos de cero e interprete.[**10 puntos**]
Dado el anterior el test de hipótesis sería:

$$H_0 : \beta_{sexo} = \beta_{escolaridad} = \beta_{RM} = 0$$

$$H_1 : \text{Al menos un } \beta_i \neq 0$$

Así tendremos un test F de la siguiente forma:

$$F = \frac{(SST - SSE)/g}{SSE/(n - K - 1)}$$

donde g es el número parámetros (en este caso serían 3), n sería el número de observaciones de la muestra y K el número de parámetros a estimar en la regresión (en este caso g = K)

$$F = \frac{(2.423093e^{16} - 1.97458e^{16})/3}{1.97458e^{16}/(58240 - 3 - 1)} = 4409.311$$

Luego el valor de F crítico $F_{2,58236}$ para un $\alpha = 0.05$ es igual a 2.997, por lo que dado que F calculado es mayor al F crítico, entonces rechazamos la hipótesis nula y al menos una de las variables es o son significativas en el modelo.

(d) Un colega le sugiere que el genero y la escolaridad son variables que no debiesen estar dentro del modelo. Usted le replica indicando que para ello tendríamos que testear la validez de un modelo restringido que excluya ambas variables. Estime el modelo restringido y construya un test F que le permite testear la veracidad de la sugerencia de su colega. ¿Qué concluye? **[10 puntos]**

Dado el anterior output del modelo restringido, el test de hipótesis sería:

$$H_0 : \beta_{sexo} = \beta_{escolaridad} = 0$$

$$H_1 : \beta_{sexo} \ \ o \ \ \beta_{esclaridad} \neq 0$$

Repetimos el mismo proceso que para la pregunta anterior:

$$F = \frac{(SSE_r - SSE_{nr})/g}{SSE_{nr}/(n-K-1)}$$

$$F = \frac{(SSE_r - SSE_{nr})/2}{SSE_{nr}/(58240-3-1)}$$

$$F = \frac{(2.370492e^{16} - 1.97458e^{16})/2}{1.97458e^{16}/(58236)} = 4863.197$$

Luego el valor de F crítico $F_{2,58236}$ para un $\alpha = 0.05$ es igual a 2.997, por lo que dado que F calculado es mayor al F crítico, rechazamos la hipótesis nula y las variables son significativas en el modelo.

(e) Un colega se le acerca y le indica que el modelo que usted estimó en realidad sufre de sesgo por variable omitida, y que habrían al menos tres variables omitidas dentro del modelo (y quizás más también), es decir, variables que con seguridad podían afectar el nivel de salarios, y al mismo tiempo estar correlacionadas ya sea con el sexo, lugar donde vive o la escolaridad. De tres ejemplos de posibles variables omitidas dentro del modelo (en cada caso, entregue evidencia empírica de lo sugerido utilizando los datos de la encuesta CASEN) **[20 puntos]**

Pregunta abierta, justificaciones acordes al modelo y supuestos de MCO, se espera que analice las correlaciones entre las variables y el aporte de estas al modelo

(f) Finalmente, otro colega le indica que es posible que los salarios de los hombres estén bastante correlacionados entre ellos, idem para el caso de los salarios de las mujeres. El colega le indica que, de ser así, entonces todo aquello no observable en el modelo también podría estar correlacionado dentro de cada grupo de género, lo cual violaría el supuesto de homocedasticidad. Qué problema se genera si la aseveración de su colega es correcta? Implemente un test que le permita chequear si se viola el supuesto de homocedasticidad o no. Describa en detalle una solución al respecto e implémentela. La solución implementada, confirma que se estaba violando el supuesto? Analice los errores estándar en cada caso y explique. **[20 puntos]**

En caso de que la aseveración sea correcta implicaría que existe heterocedasticidad en el modelo lo que afectan la validez de las inferencias estadísticas y la eficiencia de los estimadores

Un test de hipótesis que sirve para saber si el modelo presenta heterocedasticidad es el llamado Test de Breuch Pagan, cuya formulación es:

$$H_0 : \text{La varianza es constante}$$

$$H_A : \text{La varianza no es constante}$$

```
Breusch Pagan Test for Heteroskedasticity
-----------------------------------------
Ho: the variance is constant
Ha: the variance is not constant

             Data
---------------------------------
Response : y0101
Variables: fitted values of y0101

       Test Summary
-----------------------------
DF            =    1
Chi2          =    38336.8350
Prob > Chi2   =    0.0000
```

Figure 2: Resultados del test de homocedasticidad

El último término $Prob > Chi2$ es el p-valor correspondiente al test, el cual, en este caso es cero. Por lo tanto, a un 99% de significancia, se rechaza que la varianza es constante (Se rechaza la hipótesis nula). O sea, hay heterocedasticidad, y no se cumple el supuesto de homocedasticidad de OLS.
Una solución a este problema es la corrección de White, que ajusta los errores estándar para corregir la heterocedasticidad condicional. La idea principal de la corrección de White es estimar la matriz de varianza-covarianza de los errores y utilizarla para ajustar los errores estándar en las inferencias estadísticas. Al realizar dicha corrección se obtienen errores estándar más robustos.

```
t test of coefficients:

             Estimate Std. Error t value  Pr(>|t|)
(Intercept) -460085.5    11969.6 -38.438 < 2.2e-16 ***
sexo        -197143.8     5055.2 -38.998 < 2.2e-16 ***
RM           163356.6     7909.7  20.653 < 2.2e-16 ***
e6a          113768.3     1330.6  85.504 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 3: Resultados de la aplicación de la correción de White.

Luego de implementada la solución se puede realizar una comparación de los errores donde se observa que en las variables de RM y Educación existe una diferencia del 10%, en el caso de las otras variables la diferencia es menor al 10% pero no es nula, por lo que se puede interpretar que si se está violando el supuesto de homocedasticidad (Ver Figura 4

```
> # Mostrar la tabla
> print(tabla_resultados)
  ErroresConvencionales ErroresRobustos DiferenciaRelativa
1             11464.489         11969.6         0.04405875
2              4896.308          5055.2         0.03245139
3              5900.523          7909.7         0.34050829
4              1085.970          1330.6         0.22526405
```

Figure 4: comparación entre errores del modelo y errores robustos.