**Econometría I, 2025. Tarea II**

**Profesor:** Esteban Puentes
**Ayudantes:** Camila Carrasco y Santiago García
**Alumnos:** Cristóbal Donoso, Gaspar Gonzalez y Francisco Medina [1]

# Introducción

Este estudio explora las propiedades de los estimadores de Mínimos Cuadrados Ordinarios (MCO) en diferentes escenarios de varianza del error, tamaño muestral y heterocedasticidad. Utilizamos un conjunto de datos generados por el siguiente proceso:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + U_i \tag{1}$$

Donde:

- $X_{1i} \sim N(0, 1)$
- $X_{2i} \sim N(0, 2)$
- $X_{3i} \sim N(0, 1)$
- $U_i \sim N(0, \sigma_\epsilon^2)$
- $\beta = [1, 2, 1, -1]'$

El estimador de MCO se define como,

$$\hat{\beta} \quad = \quad (X'X)^{-1}X'Y \tag{2}$$

el cual depende directamente de la matriz de observaciones $X$.

En lo que sigue, necesitaremos estimar la varianza del estimador, ya que esta nos permitirá calcular intervalos de confianza y realizar pruebas de hipótesis sobre los resultados obtenidos. En términos generales, la varianza del estimador $\hat{\beta}$ dado los datos $X$ esta dada por:

$$Var(\hat{\beta}|X) = \sigma^2(XX')^{-1} \tag{3}$$

donde $\sigma^2$ representa la varianza asociada al termino de error. En este caso, es fácil conocer el valor de $\sigma^2 = \sigma_e^2$ porque conocemos el proceso generador de datos (DGP; Ecuación 1). Sin embargo, en un escenario realista, no sabemos la forma del GDP y por ende tampoco los parámetros reales que subyace sobre cada distribución.

Cuando asumimos que el termino de error es homocedástico (i.e, no varia con las observaciones) un buen estimador de la varianza del estimador esta dada por,

$$\widehat{\text{Var}}(\hat{\beta}) = s^2(X'X)^{-1} \tag{4}$$

donde, $s^2$ es el estimador de la varianza del error, comúnmente calculado como $s^2 = \frac{\sum \hat{u}_i^2}{N-K}$, siendo $\hat{u}_i = (Y - X\hat{\beta})$ los residuos estimados, $N$ el número de observaciones y $K$ el número de parámetros. Una adaptación del estimador de varianza para el caso con error heterocedástico se muestra en la Pregunta III.

A continuacion se detalla cada uno de los experimentos y sus resultados.

[1] El codigo con la solucion a esta tarea se puede encontrar en: https://github.com/cridonoso/htest_mco.git

# Pregunta I

En este ejercicio se estima un modelo de regresión lineal en el que el vector de parámetros es $\beta = [1, 2, 1, -1]'$, y se analiza específicamente el comportamiento de los intervalos de confianza para el coeficiente $\beta_3$ (correspondiente a la restricción $R = [0, 0, 0, 1]$). Asumiendo normalidad sobre los errores [2] se deriva que,

$$\hat{\beta}|X \sim N(\beta, \sigma^2(X'X)^{-1})$$

donde los coeficientes estimados $\hat{\beta}$ siguen una distribución normal con media $\beta$ y varianza $\sigma^2(X'X)^{-1}$. Al conocer la distribución de los estimadores es posible aplicar intervalos de confianza basados en el valor $Z$ de la distribución normal estándar. No obstante, dado que la varianza original de los datos se asume no observada, entonces utilizamos el estadístico $t$ de Student para construir los intervalos de confianza.

Formalmente, los errores de predicción de nuestro modelo siguen una distribución $t$-Student con $N - K$ grados de libertad,

$$\frac{R'\hat{\beta} - R'\beta}{\sqrt{s^2 R'(X'X)^{-1}R}} \sim t_{N-K}$$

donde $R$ es el vector de restricción, y $\sqrt{s^2 R'(X'X)^{-1}R}$ el error estándar estimado. Lo anterior nos permite construir la siguiente estructura de intervalos de confianza:

$$\beta_3 \in [\hat{\beta}_3 \pm \sqrt{s^2 R'(X'X)^{-1}R} t_{1-\alpha/2,N-K}]$$

La idea central de este ítem es evaluar si los intervalos de confianza construidos cumplen con la cobertura nominal teórica (por ejemplo, si se usa un intervalo de confianza al 95 %, se espera que, de las 500 simulaciones realizadas, cerca del 95 % de ellas contengan al valor verdadero de $\beta_3$).

Cuadro 1: Parámetro en los intervalos de confianza para $\beta_3$ variando la varianza del error

| $\sigma_e^2$ | $\alpha = 1\,\%$ | $\alpha = 5\,\%$ | $\alpha = 10\,\%$ |
|---|---|---|---|
| **1** | 492 (98.4 %) | 472 (94.4 %) | 453 (90.6 %) |
| **2** | 497 (99.4 %) | 482 (96.4 %) | 458 (91.6 %) |
| **10** | 494 (98.8 %) | 478 (95.6 %) | 451 (90.2 %) |

Nota: La tabla muestra la cantidad veces en que el parámetro $\beta_3$ se encuentra en el intervalo de confianza en 500 simulaciones. A través de las columnas varía la insignificancia del intervalo. En las filas se indican la varianza del error de la esperanza condicional.

Los resultados en el Cuadro 1 muestran que la proporción de iteraciones que contienen el valor verdadero de $\beta_3$ dentro del intervalo de confianza respecto del total, confirma la cobertura nominal teórica.

[2] En verdad sabemos que el termino de error se distribuye normal porque conocemos el proceso generador de datos. En general esto esto es desconocido en la practica, por lo tanto lo asumimos.

En la Tabla 1, se muestra ademas el efecto de variar los valores de $\sigma_e^2$, lo cual afecta la dispersión de la distribución de los errores. Al aumentar la magnitud del error entonces el proceso de optimización se torna más ruidoso, lo cual afecta $E[\hat{\beta}|X]$ el valor esperado de nuestro coeficiente estimado. No obstante, dado que nuestra muestra $N = 100$ es suficientemente grande, la distribución $t$ nos sigue proporcionando resultados que confirman los valores teóricos. También es relevante mencionar que a niveles de $\sigma_e^2$ más altos, los intervalos de confianza son más amplios y por lo tanto las estimaciones son menos precisas. En ese sentido, los intervalos de confianza internalizan la mayor varianza, por lo que no debería variar sistemáticamente el número de veces que el parámetro esta contenido.

## Pregunta II

En esta pregunta exploraremos el efecto que tiene variar $N$ el tamaño de la muestra sobre la inferencia estadística, incluyendo los intervalos de confianza y el rechazo de la hipótesis nula en un test-t bilateral. Como se mencionó en la sección pasada (Pregunta I) el estimador $t$ se define como,

$$t = \frac{R'\hat{\beta} - c}{\sqrt{s^2 R'(X'X)^{-1}R}}, \tag{5}$$

donde $R$ un vector de restricción que permite trabajar sobre los parámetros de interés, y $c = R'\beta$ es el valor propuesto para la hipótesis nula. Rechazaremos la hipótesis nula a un nivel de significancia de $\alpha$ si

$$|t| > t_{1-\alpha/2} \tag{6}$$

donde $t_{1-\alpha/2}$ representa el valor critico de la distribución $t$ de Student que corresponde al cuantil $1 - \alpha/2$.

Es importante mencionar que existe una correspondencia directa entre los intervalos de confianza y el rechazo de la hipótesis nula. Un valor fuera del intervalo de confianza para un nivel $1 - \alpha$ implica un rechazo de la hipotesis nula a un nivel de signiciancia $\alpha$ y viceversa.

### Parte 2.1

Al igual que la Pregunta I, el primer experimento consistió en cuantificar el número de veces que, el valor real de $\beta_3$ cayó dentro de los intervalos de confianza cuando variamos el tamaño de la muestra $N$. Los resultados se muestran en el Cuadro 2.

Cuadro 2: Parámetro en los intervalos de confianza para $\beta_3$ variando el tamaño muestral

| **N** | $\boldsymbol{\alpha = 1\,\%}$ | $\boldsymbol{\alpha = 5\,\%}$ | $\boldsymbol{\alpha = 10\,\%}$ |
|---|---|---|---|
| **50** | 493 (98.6 %) | 475 (95.0 %) | 449 (89.8 %) |
| **100** | 497 (99.4 %) | 474 (94.8 %) | 447 (89.4 %) |
| **500** | 495 (99.0 %) | 479 (95.8 %) | 452 (90.4 %) |

Nota: La tabla muestra la cantidad veces en que el parámetro $\beta_3$ se encuentra en el intervalo de confianza en 500 simulaciones. A través de las columnas varía la insignificancia del intervalo. En las filas se indican el tamaño muestral de cada simulación.

Hay dos aspectos interesantes a analizar sobre el número de veces en que el parámetro esta en el intervalo de confianza, el primero es como cambia cuando varia el nivel de significacia y el segundo es como cambia cuando varia el tamaño de las muestras. Para estos análisis es importante primero recordar que significa intervalo de confianza al x % de confianza. Este se define como que la probabilidad (ex-ante) que un **intervalo aleatorio** del parámetro contenga el valor verdadero es de (1-x) %.

Se aprecia claramente en la tabla que al aumentar el nivel de significacia se reduce la cantidad de veces que el parámetro esta dentro del intervalo. Esto tiene sentido, ya que por lo explicado en el párrafo anterior, si aumenta la signaficancia, entonces hay menos probabilidad que, al realizar una simulación, el intervalo de confianza generado contenga al parámetro real. Dicho de otra forma, hay menos simulaciones cuyo intervalo de confianza efectivamente contiene el parámetro $\hat{\beta}_3$ para una significacia de 10 %, que para una de 5 %, que para una de 1 %. Se puede apreciar también, que los porcentajes de aciertos del intervalo de confianza son muy similares a la probabilidad "teórica".

Dado que el estimador $t$-Student es robusto sobre muestras pequenas, no hay una diferencia significativa en el porcentaje de veces en las que el parámetro efectivamente esta en el intervalo. La razón es que para construir el intervalo aleatorio se ocupa la información del tamaño de la muestra. Por ende, independiente del tamaño de la muestra, la probabilidad ex-ante es la misma. Por lo que es esperable que no varie mucho. Efectivamente esto es lo que se observa en el Cuadro 2. Para cualquier tamaño de la muestra $N$, el porcentaje efectivo es similar a la probabilidad teórica; y ningún tamaño muestra es mejor sistemáticamente. El único detalle importante es que se esta usando la varianza estimada, sin embargo, todos los experimentos utilizan el mismo estimador insesgado, por lo que no hay desventaja relativa.

### Parte 2.2

En esta Sección estudiamos el efecto que tiene el tamaño de la muestra sobre la cantidad de rechazos sobre la hipótesis nula $H_0 : \beta_2 \in \{0.5, 1\}$.

Cuadro 3: Rechazos Test de Hipótesis para el Parámetro $\beta_2$

| **N** | $\boldsymbol{H_0 : \beta_2 = 0.5}$ | $\boldsymbol{H_0 : \beta_2 = 1}$ |
|---|---|---|
| **50** | 452 (90.4 %) | 37 (7.4 %) |
| **100** | 499 (99.8 %) | 20 (4.0 %) |
| **500** | 500 (100.0 %) | 31 (6.2 %) |

Nota: La tabla muestra la cantidad de rechazos del test de hipótesis para el parámetro $\beta_2$ en 500 simulaciones. La primera columna se tiene como hipótesis nula que que $\beta_2 = 0,5$, lo que es falso; La segunda columna se tiene como hipótesis nula que que $\beta_2 = 1$, lo que es verdadero. En las filas se indican el tamaño muestral de cada simulación.

Los resultados se muestran en el Cuadro 3. Note que, el segundo test tiene como hipótesis nula $\beta_2 = 1$. Esta hipótesis es igual al valor verdadero de la población. Por ende, lo esperable es que se rechace pocas veces, lo que efectivamente se observa que ocurre. El test se construye de forma tal que la probabilidad ex-ante que se rechace es 5 %. Es destacable que en varias simulaciones se rechaza la nula, pese a ser verdadera realmente, esto es error

tipo I, es verdadero que en la población (o PGD) $\beta_2 = 1$, pero a veces la evidencia lo rechaza.

Se puede observar que en para los diferentes tamaños muéstrales el porcentaje de rechazos es similar al 5 % teórico; si aumentarán la cantidad de simulaciones esperaríamos que converja a 5 %, ya que así esta hecho por construcción. Es esperable que los rechazos no varíen según el tamaño muestral, dado que los intervalos de confianza se construyen utilizando $N$ el tamano de la muestra. Por lo tanto, independiente del $N$, se espera un 5 % de rechazo.

El primer test tiene como hipótesis nula $\beta_2 = 0.5$. Esta hipótesis es distinta al valor verdadero del proceso generador de datos. Por ende, lo esperable es que se rechace muchas veces, lo que efectivamente se observa que ocurre. Se destaca que con 50 observaciones hay veces en que no se rechaza la nula, siendo que esta es falsa para la población (o PGD), ya que $\beta_2 = 0.5$, esto es error tipo II, pese a ser falso, a veces la evidencia no es suficiente para rechazarlo.

Teóricamente es esperable que, mientras mayor es el tamaño de la muestra, más simulaciones deberían rechazar la hipótesis nula.

*¿Por qué en este caso importa el tamaño muestral y cuando la nula era verdadera no importaba?*

La razón es que, como se mencionó, cuando la nula es verdadera hay un 5 % de rechazo esperado, determinado por el nivel de significancia $\alpha$ (no por el tamaño muestral $N$. En cambio, cuando la hipótesis nula es falsa, esta propone un valor alejado del parámetro verdadero. La distribución del estimador $\hat{\beta}_2$ está centrada en el valor real, y su varianza disminuye con el tamaño muestral. A mayor varianza (muestras pequeñas), los intervalos de confianza son más amplios, lo que hace más probable que incluyan el valor propuesto en la hipótesis nula ($H_0 : \beta_3 = 0.5$), incluso siendo falso. Por el contrario, con muestras grandes (100 o 500), la varianza se reduce, los intervalos son más estrechos y el valor 0.5 queda fuera. Esto explica por qué, con $n = 50$, algunas simulaciones no rechazan la nula (error Tipo II), mientras que con tamaños mayores el rechazo es consistente.

En síntesis, estos experimentos sirven para entender con mayor profundidad los test de hipótesis. Estos se construyen con una referencia a la significancia. Esto no es perfecto, puede ocurrir que algo que es verdadero se rechace ($\beta_2 = 1$), error tipo I. Pero también puede ocurrir que algo falso no se pueda rechazar $\beta_2 = 0.5$, error tipo II. Se aprecia que la tasa de rechazo no depende del tamaño muestral cuando la hipótesis es verdadera, pero sí afecta cuando es falsa.

## Pregunta III

En el siguiente inciso exploraremos el efecto de la hetorecedasticidad en la estimación de $\hat{\beta}$. En particular, se asume que $U_i$ el ruido está correlacionado con $X_{2i} \in X$ el segundo predictor en nuestra matriz de diseño. Formalmente,

$$U_i = X_{2i} \cdot \epsilon_i, \tag{7}$$

donde $\epsilon \sim N(0,1)$. Si tomamos la varianza condicional en $X$ sobre la Ecuación 7,

$$Var(U_i|X) \quad = \quad E[U_i^2|X] - (E[U_i|X])^2 \tag{8}$$

donde,

$$E[U_i^2|X] \quad = E[(X_{2i} \cdot \epsilon_i)^2|X] = X_{2i}^2 \cdot E[\epsilon_i^2|X] = X_{2i}^2 \sigma_e^2 \tag{9}$$
$$E[U_i|X] \quad = E[X_{2i} \cdot \epsilon_i|X] = X_{2i} \cdot E[\epsilon_i|X] = 0 \tag{10}$$

reemplazando 9 y 10 en la Ecuacion 8, obtenemos,

$$Var(U_i|X) \quad = \quad X_{2i}^2 \sigma_e^2 \tag{11}$$

donde el efecto condicional en $X_{2i}$ actúa como un factor de escala sobre el ruido. Si $X_{2i}^2$ toma valores grandes, la variabilidad del error aumenta proporcionalmente al cuadrado de $X_{2i}$. En contraste si $X_{2i}$ toma valores pequeños, la variabilidad disminuye.

### Estimador Robusto Tipo White

El estimador clásico de varianza en Mínimos Cuadrados Ordinarios, asume homocedasticidad, es decir, que la varianza del termino de error sea constante a lo largo de las observaciones. Sabemos que en este caso el supuesto es falso, ya que en el proceso de generacion de datos tenemos una correlacion entre el predictor $X_{2i}$ y el termino de ruido blanco $\epsilon$.

Para solucionar este problema utilizamos un estimador de varianza-covarianza robusto a la heterocedasticidad propuesto por White [1],

$$\hat{V}(\hat{\beta} \mid X) = (X'X)^{-1} \left( \sum_{i=1}^{N} X_i X_i' \hat{e}_i^2 \right) (X'X)^{-1}, \tag{12}$$

donde el termino que contiene la sumatoria recibe el nombre de *Kernel de White*. Intuitivamente, dado que el error $\hat{e}_i$ no se asume constante[3], utilizamos la $X_i X_i' \in \mathbb{R}^{4\times4}$ una matriz de productos exteriores, para ponderar el error $e_i$. Esto permite que las observaciones con mayor variabilidad en los errores tengan un mayor peso en la estimación de la incertidumbre de los coeficientes, ajustando así el estimador para reflejar la heteroscedasticidad presente en los datos.

### Resultados

Se realizaron 500 simulaciones donde cada una de ellas comprendio una muestra de 100 datos. El estimador $\hat{\beta}$ sigue la misma forma que en la Ecuacion 2.

En la Figura 1 se muestran las matrices de covarianzas asumiendo homocedasticidad (ver Ecuacion 4) y utilizando el Estimador de White. Es claro ver el efecto de la heterocedasticidad sobre la covarianza del predictor $X_2$ -cuando se ocupa el estimador de White la magnitud

[3]cuando asumíamos homocedasticidad asumíamos que los residuos eran similar a la varianza (i.e., $\sigma^2 = s^2 = \frac{\hat{e}}{n-k}$), la cual era constante y no dependía de las observaciones. En el caso heterocedástico, los residuos dependen de cada observación por lo tanto no es constate.

de la covarianza aumenta a más del doble. Esto es consistente con el DGP propuesto en la Ecuación 7, donde el error es proporcional a $X_2$. Esto aumenta la inceridumbre sobre $\hat{\beta}_2$, la cual es correctamente capturada por el estimador robusto de covarianza.

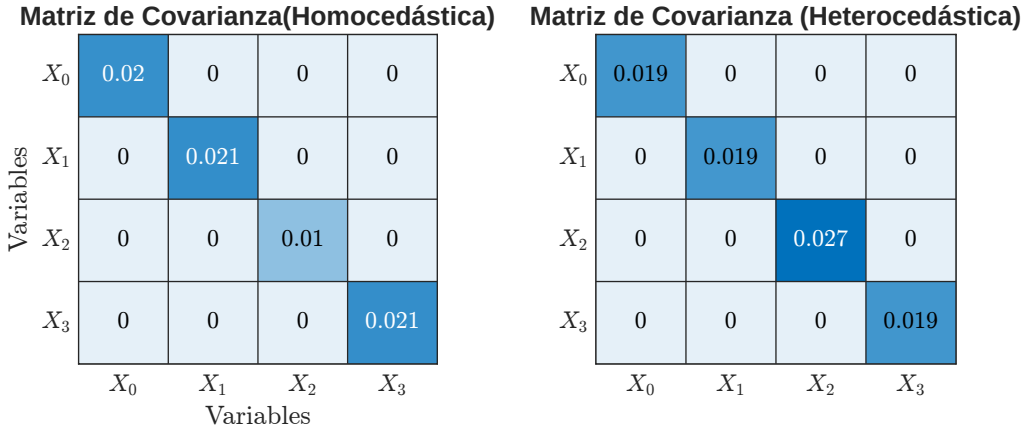


Figura 1: Matrices de Covarianzas estimadas asumiendo Homocedasticidad (Izquierda) y Heterocedasticidad utilizando el estimador de White (Derecha). $X_0$ es la columna en la matriz $X$ asociada al intercepto $b_0$

La matriz de covarianzas tipo White es efectiva en capturar la variabilidad del error en un escenario Heterocedastico. Sin embargo, en la Figura 1 tambien podemos ver que subestima el valor de las covarianzas asociadas a los otros parámetros ($\{X_0, X_1, X_3\}$).

Como se ve en la Ecuación 12, la ponderación del error depende directamente de la cantidad de datos. Mientras mayor sea $N$ mejor será nuestra estimación [4]. Esto convierte al estimador de White sensible a la cantidad de datos.

Al igual que en la Pregunta II, realizamos un test $t$ bilateral para la hipótesis $H_0 : \beta_2 = 1$. Luego registramos cuantas veces se rechaza la hipótesis nula bajo cada estimador (homocedástico/heterocedástico). En la tabla 4 se muestra el número de rechazos para cada caso. Podemos ver que cuando asumimos homocedasticidad hay 14.2 puntos porcentuales más de rechazo sobre hipótesis nula. Esto quiere decir que aumenta la cantidad de falsos positivos.

Cuadro 4: Rechazos de la nula asumiendo homocedasticidad y heterocedasticidad

| | **Homocedástico** | **Estimador de White** |
|---|---|---|
| **Rechazos $H_0$** | 130 (24.6 %) | 42 (10.40 %) |

Nota: Número de rechazos sobre la Hipótesis Nula $H_0 : \hat{\beta} = 1$ utilizando el estimador homocedástico y tipo White (heterocedástico). Resultados obtenidos con 500 simulaciones.

Una explicación sobre el aumento de falsos positivos sobre el test $t$ dice relación con la forma de la varianza. En la Ecuación 5 se puede observar que, si la desviación en el denominador es subestimada (como se evidencia en la Figura 1), entonces el valor del estadístico $t$ se

[4] Esto igual aplica para el caso homocedástico, recordemos que los residuos también dependen de $X$ los datos. Sin embargo, el estimador de White aumenta aún mas esta dependencia.

infla artificialmente, produciendo una mayor cantidad de rechazos. Este efecto se disminiuye considerablemente utilizando el estimador de White.

## Referencias

[1] Halbert White. A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica*, 48(4):817–838, 1980.