Econometría I, 2025. Tarea I

Profesor: Esteban Puentes
Ayudantes: Camila Carrasco y Santiago García
Alumnos: Felipe Díaz, Cristóbal Donoso y Vicente Perales

Introducción

En este problema se busca predecir una variable dependiente $Y \in \mathbb{R}^N$ por medio de un conjunto de predictores independientes $X \in \mathbb{R}^{N\times 3}$. El modelo que explica Y en función de X se muestra en la Ecuación 1. Los parámetros del modelo son dados por un vector β que pondera cada una de las características en X. El vector de error $U \in \mathbb{R}^N$ representa aquellas características no observadas que condicionan el valor de Y, el cual asumimos ruido blanco con varianza σ_ϵ^2. Así mismo, b_0 constituye el termino constante o intercepto del modelo de regresión en la Ecuación 1.

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + U_i \tag{1}$$

Para generar el vector Y de observaciones reales, se realiza una simulación de X los predictores, y U el error. Cada vector sigue una distribucion Normal con media 0 y desviación constante como se muestra en 2.

$$\begin{array}{rcl} X_{1i} & \sim & \mathcal{N}(0,1) \\ X_{2i} & \sim & \mathcal{N}(0,2) \\ X_{3i} & \sim & \mathcal{N}(0,1) \\ U_i & \sim & \mathcal{N}(0,\sigma_\varepsilon^2) \end{array} \tag{2}$$

Los parámetros reales utilizados para ponderar cada unos de los vectores en la Ecuación 1 vienen dados por,

$$\beta = \begin{bmatrix} 1 \\ 2 \\ 1 \\ -1 \end{bmatrix}$$

Para predecir Y debemos estimar los parámetros de β (desde ahora $\hat{\beta}$) los cuales se obtienen utilizando la Ecuación 3. Esta forma de $\hat{\beta}$ es la solución típica para el problema de Mínimos Cuadrados Ordinarios (MCO).

$$\hat{\beta} = (X'X)^{-1}(X'Y) \tag{3}$$

En las siguientes secciones exploraremos cómo afecta a $\hat{\beta}$ los parámetros estimados, cambios en los supuestos de la simulación. En particular, modificaremos las distribuciones de los predictores y la cantidad de simulaciones que constituyen la muestra para estimar β[1].

[1] El código solución a esta tarea se puede encontrar en https://github.com/cridonoso/mc_matlab

Pregunta I

En esta sección variamos el valor de $\sigma_e^2 \in \{1, 2, 10\}$ la varianza del error U, y estudiamos el efecto que tiene en la estimación de los parámetros reales de β.

La Figura 1 muestra la distribución de cada estimador $\hat{\beta}_i$ para los distintos valores de σ_e^2. Se aprecia que para todos los valores de σ_e^2 la distribución toma una forma de campana, similar a una distribución normal, donde la media o valor central de la distribución tiende al valor real del parámetro β_i.

En la misma Figura 1, se muestra cómo los distintos valores de σ_e^2 inciden directamente en la varianza del estimador $\hat{\beta}$. Intuitivamente, un mayor error (e.g., $\sigma_e^2 = 10$) dificulta el ajuste de los parámetros en β, ya que aumenta la incertidumbre en las estimaciones. Por el contrario, cuando el error es menor (e.g., $\sigma_e^2 = 1$), la variable Y puede explicarse con mayor precisión a través de las variables observadas en X. Cabe mencionar que el objetivo del modelo es estimar Y mediante la matriz de predictores observados X, por lo que la componente no observada U –cuyo valor condiciona directamente a Y– representa un factor crítico en la calidad de las estimaciones.

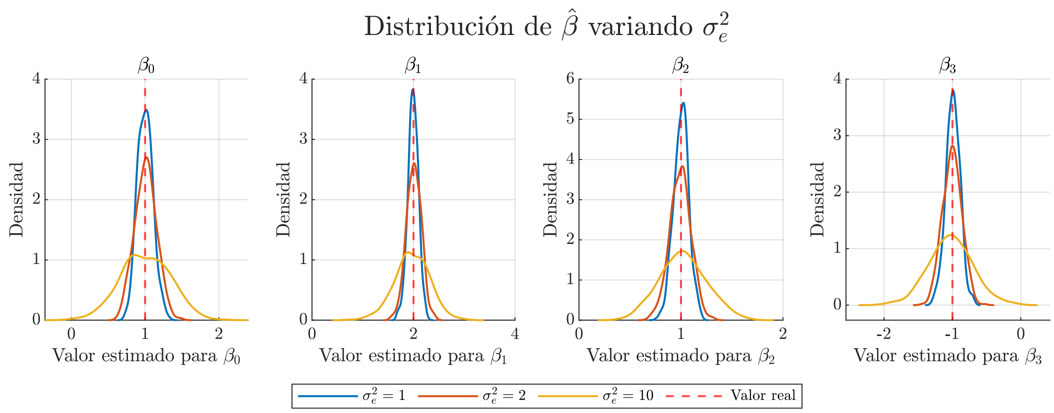

Figura 1: Distribución de los predictores de β. Cada cuadrante contiene tres curvas de densidad, cada una representando el estimador $\hat{\beta}$ cuando el error $U \sim \mathcal{N}(0, \sigma_e^2)$, con $\sigma_e^2 \in \{1, 2, 10\}$.

Los percentiles 5 y 95 se muestran en la Cuadro 1. Estos valores abarcan el 90 % central de los valores estimados para cada $\hat{\beta}$. Una representación visual se muestra en la Figura 1, donde cada estimador tiene asociado tres barras de error (una para cada valor de σ_e^2). Desde los valores obtenidos podemos notar que en todos los casos el modelo es capaz de mantenerse relativamente cercano al valor real de β. Sin embargo, a mayor valor de sigma, ambos percentiles se alejan del valor real de β_i. Así mismo, la Figura 2 muestra que a medida que crece σ_e^2, el intervalo es más amplio, es decir, la incertidumbre del estimador $\hat{\beta}$ aumenta con σ_e^2.

σ_ε^2	Percentil	β_0	β_1	β_2	β_3
1	5 %	0.925	1.93	0.954	-1.06
	95 %	1.07	2.06	1.05	-0.922
2	5 %	0.913	1.91	0.924	-1.10
	95 %	1.10	2.11	1.06	-0.910
10	5 %	0.774	1.76	0.861	-1.23
	95 %	1.24	2.21	1.18	-0.818

Cuadro 1: Percentiles para diferentes valores de σ_ε^2

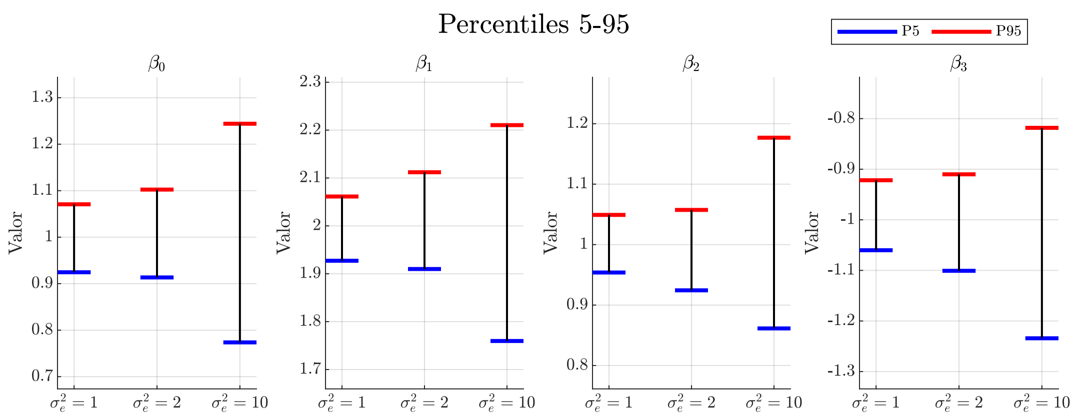

Figura 2: Representación gráfica de los percentiles 5 y 95 para cada valor de σ_e^2. Cada cuadrante representa un estimador distinto segun nuestro modelo en la Ecuación 1.

Pregunta II

En esta sección se explora el efecto que tiene el tamaño de la muestra N en la precisión del estimador $\hat{\beta}$. Para generar nuestra variable dependiente Y utilizamos el mismo modelo descrito en la Sección . Así mismo, las distribuciones para cada vector estimador en X se mantiene igual que en 2 y fijamos $\sigma_e^2 = 2$.

Para ver el efecto que tiene el volumen de la muestra, realizamos varias estimaciones para β con $N \in \{50, 100, 500\}$. Las distribuciones de los estimadores para cada tamaño de muestra se encuentran en la Figura 3.

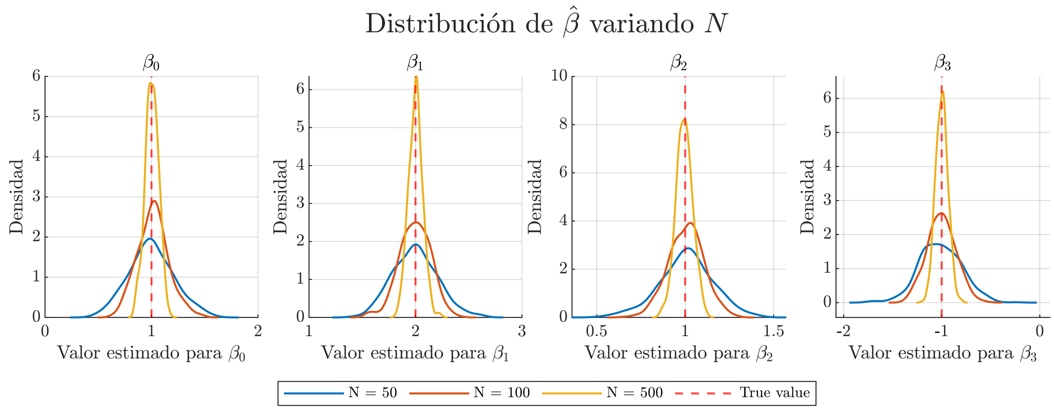

Figura 3: Distribución de los predictores de β. Cada cuadrante contiene tres curvas de densidad, cada una representando el estimador $\hat{\beta}$ cuando el tamaño de la muestra $N \in \{50, 100, 500\}$.

La Figura 3 muestra que el tamaño de la muestra reduce la varianza de la distribución, llevando a una distribución con una campana más estrecha alrededor del valor real del parámetro β en cuestión. Esto es consistente con la varianza del estimador,

$$V(\hat{\beta} \mid X) = \sigma^2 (X^\top X)^{-1}, \tag{4}$$

la cual disminuye a medida que aumenta la cantidad de datos presente en $X^\top X \in \mathbb{R}^{3\times 3}$. Dicho de otra forma, si tenemos mas datos entonces los valores de $X^\top X$ aumentaran en magnitud, reduciendo la varianza del modelo. Esto, junto a la forma normal de la distribución, es consistente además con el Teorema Central del Limite (TCL).

N	Percentil	β_0	β_1	β_2	β_3
50	5 %	0.857	1.85	0.917	-1.16
	95 %	1.14	2.14	1.10	-0.867
100	5 %	0.918	1.89	0.941	-1.10
	95 %	1.10	2.10	1.07	-0.908
500	5 %	0.962	1.96	0.964	-1.04
	95 %	1.05	2.04	1.02	-0.960

Cuadro 2: Percentiles para diferentes valores de N

Los percentiles de los estimadores según el tamaño de la muestra (Cuadro 2 y Figura 4) muestran desde otra óptica lo mencionado anteriormente; a medida que la cantidad de la muestra N aumenta, los estimadores (el 90 % de la distribución) disminuye su desviación, concentrándose con mayor precisión al rededor del valor real de β.

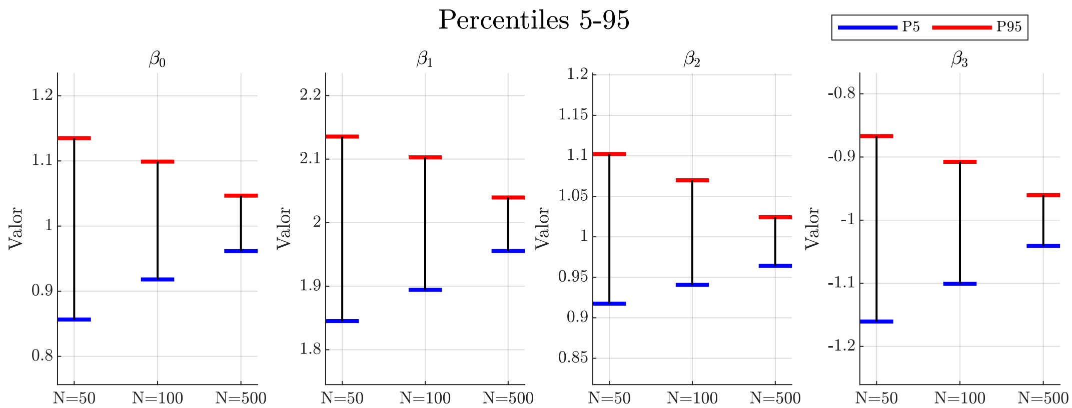

Figura 4: Representación gráfica de los percentiles 5 y 95 para cada valor de N el tamaño de la muestra. Cada cuadrante representa un estimador distinto según nuestro modelo en la Ecuación 1.

Pregunta III

En este caso, se explora como afecta que uno de los predictores en X rompa el supuesto de indepdencia y correlacione con el termino de error U. Formalmente, se tiene,

$$\begin{aligned} U_i &\sim \mathcal{N}(0, \sigma_\epsilon^2) \\ X_{1i} &\sim \mathcal{N}(0, 1) \\ X_{2i} &\sim \mathcal{N}(0, 2) \\ Z_i &\sim \mathcal{N}(0, 1), \quad X_{3i} = Z_i + \lambda U_i \end{aligned}$$

donde X_{3i} el tercer predictor correlaciona con el ruido U_i en una fracción λ. En lo que respecta al modelo real de Y, este se mantiene igual al definido en la Ecuación 1.

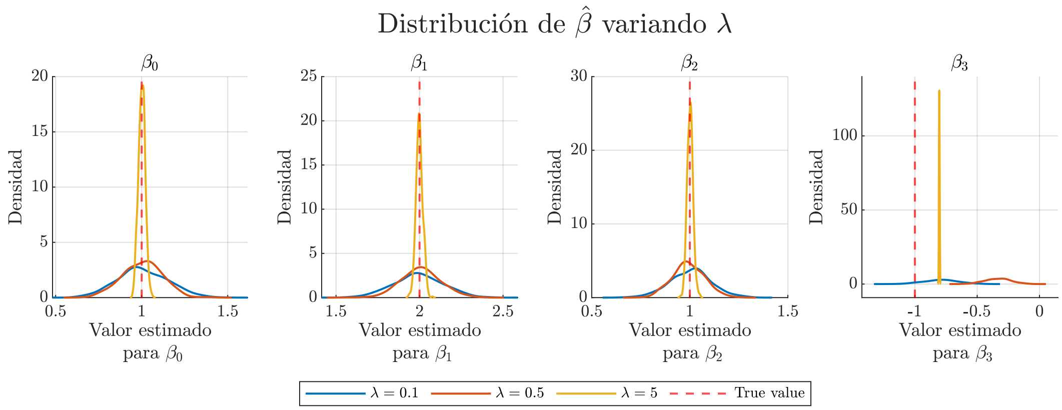

Figura 5: Distribución de los predictores de β. Cada cuadrante contiene tres curvas de densidad, cada una representando el estimador $\hat{\beta}$ cuando $\lambda \in \{0{,}1, 0{,}5, 5\}$, donde λ es un factor de peso en $X_3 = Z + \lambda U$.

La Figura 5 muestra los resultados de la distribución de los estimadores $\hat{\beta}$ en el escenario descrito anteriormente. Se puede apreciar un sesgo que impulsa el valor de $\hat{\beta}_3$ a tomar valores distintos de 0. Esto se ve claramente en el gráfico para el estimador $\hat{\beta}_3$ donde, a mayor λ la distribución se desplaza hacia la derecha del valor real ($\beta = -1$).
El efecto de la endogeneidad de la variable X_3 se mantiene incluso para valores de λ menores y que, por lo tanto, introducen un sesgo menor. También es importante notar que los parámetros β_0, β_1, y β_2 no se ven afectados directamente por X_3 pues son independientes entre si. Esto se puede ver en la Figura 5 donde la media de los estimadores se ajusta con mayor precision al valor real de β.

λ	Percentil	β_0	β_1	β_2	β_3
0.1	5 %	0.911	1.90	0.933	-0.880
	95 %	1.11	2.09	1.07	-0.706
0.5	5 %	0.923	1.93	0.943	-0.404
	95 %	1.08	2.08	1.05	-0.267
5	5 %	0.987	1.99	0.991	-0.806
	95 %	1.01	2.01	1.01	-0.802

Cuadro 3: Percentiles para diferentes valores de λ

Tanto el Cuadro 3 como la Figura 6 muestran los percentiles 5 y 95 para cada uno de los coeficientes según el valor de λ. Lo que destaca es que para β_3, el valor real del parámetro no está en el intervalo, haciendo evidente la confusión del modelo para predecir β_3 debido al sesgo insertado durante la simulación.

Como es de esperar, los otros parámetros se ajustan al valor real de β pues no se ven directamente impactados por el valor de X_3. Sin embargo, podemos notar que, si λ crece, los estimadores para β_0, β_1 y β_2 disminuyen su desviación alrededor del valor real de β. Esto se debe a que valores altos de λ aumentan la endogeneidad de X_3 disminuyendo la interferencia en la estimación de los otros parámetros. Dicho de otra forma, si la correlación entre X_3 y U aumenta, entonces X_3 comienza a comportarse como ruido en el modelo.

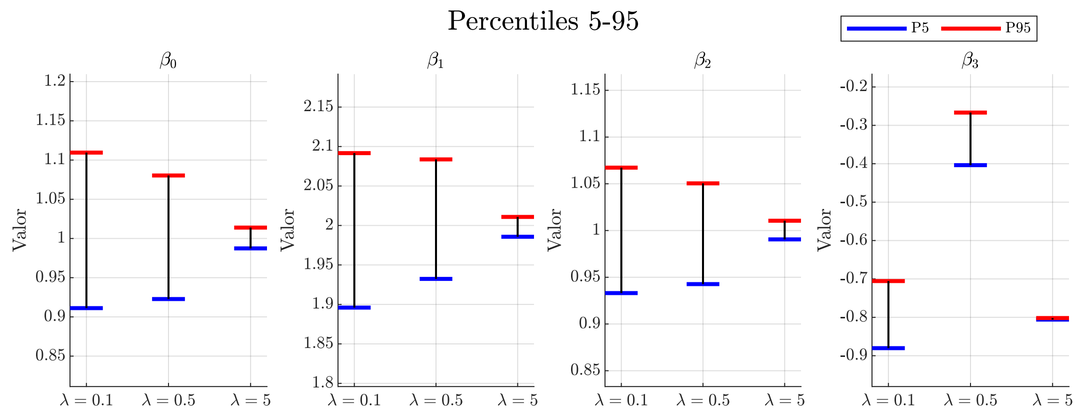

Figura 6: Representación gráfica de los percentiles 5 y 95 para cada valor de N el tamaño de la muestra. Cada cuadrante representa un estimador distinto según nuestro modelo en la Ecuación 1