Introducción a Regresión Múltiple

Métodos cuantitativos aplicados a estudios urbanos II - MEU
UTDT

Ricardo Pasquini - rpasquini@utdt.edu

June 26, 2025

Plan

- Inclusión de variables categóricas como explicativas
- Múltiples categorías.
- Interpretando coeficientes en regresión múltiple
- Omisión de variable relevante
- Inclusión de variable irrelevante y colinealidad

Inclusión de variables categóricas como explicativas

- Las variables categóricas son variables que tienen un número limitado de valores posibles.
- Ejemplos: género, país, ciudad, etc.
- En algunos casos una condición se cumple o no. Ejemplo: *propiedad posee cochera* o *no la posee*.
- En otros casos pueden existir múltiples categorías. Ejemplo: *tipo de propiedad* (casa, departamento, etc.).
- Ejm: *barrio de origen* (Barrio Almagro, Barrio Belgrano, Barrio Caballito, etc.).

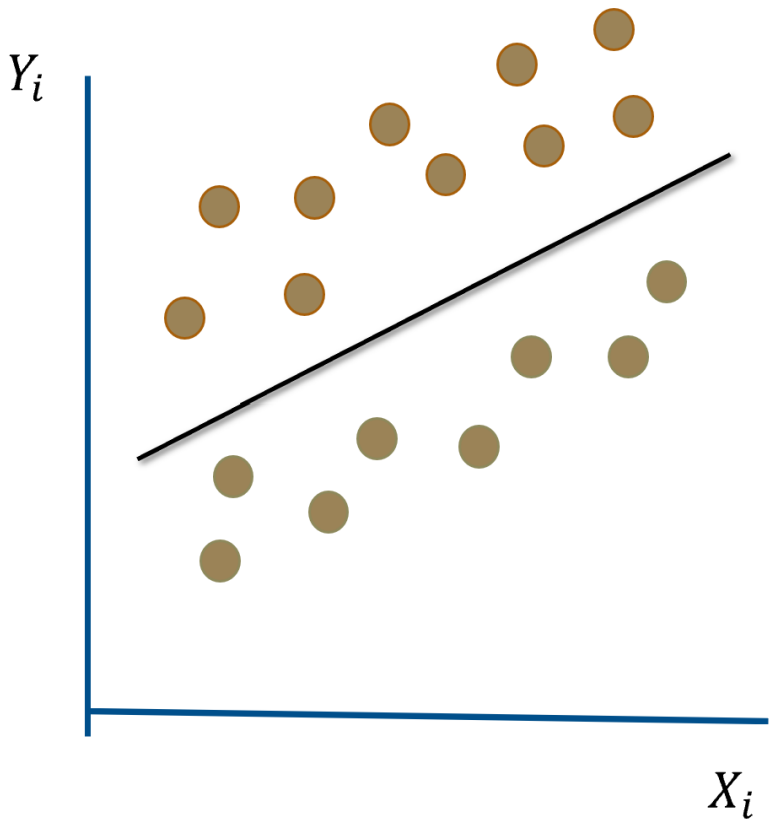
Inclusión de variables categóricas como explicativas

- En general, las variables categóricas se pueden incluir en el modelo como variables binarias, tambien llamadas *dummy variables*.
- Una variable binaria es una variable que solo puede tomar dos valores: 0 o 1.
- 0 representa la ausencia de la característica y 1 representa la presencia de la característica.
- Ejemplo: *propiedad posee cochera* $= 1$ si la propiedad posee cochera y 0 si no la posee.

Inclusión de variables categóricas como explicativas

Motivación

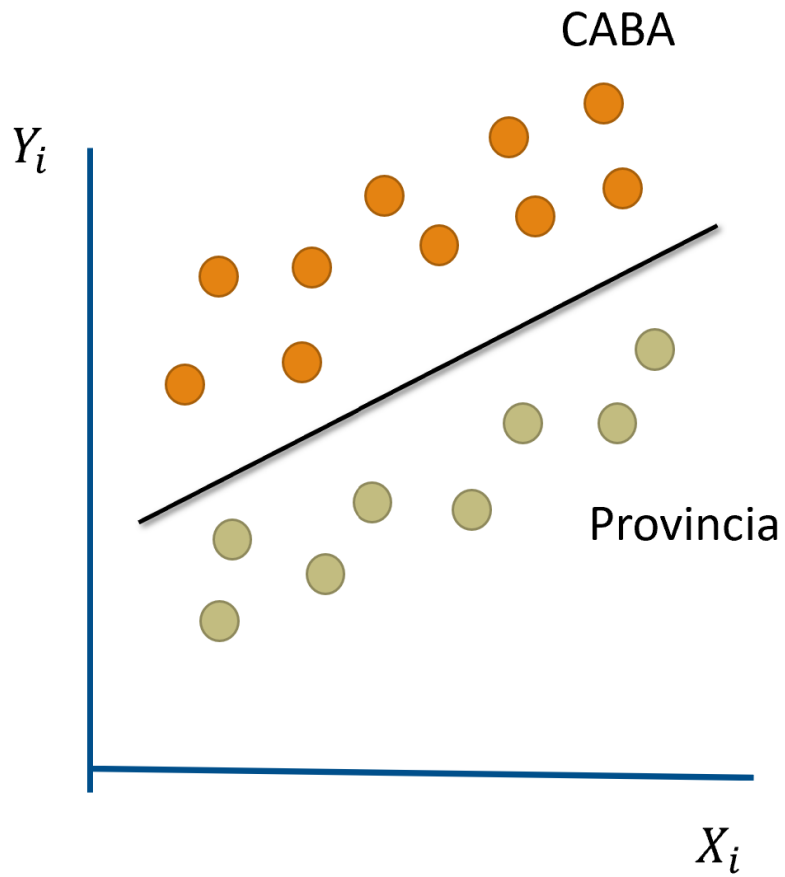
$$Y_i = \beta_0 + \beta_1 X_{1,i} + \varepsilon_i$$



Inclusión de variables categóricas como explicativas

Motivación

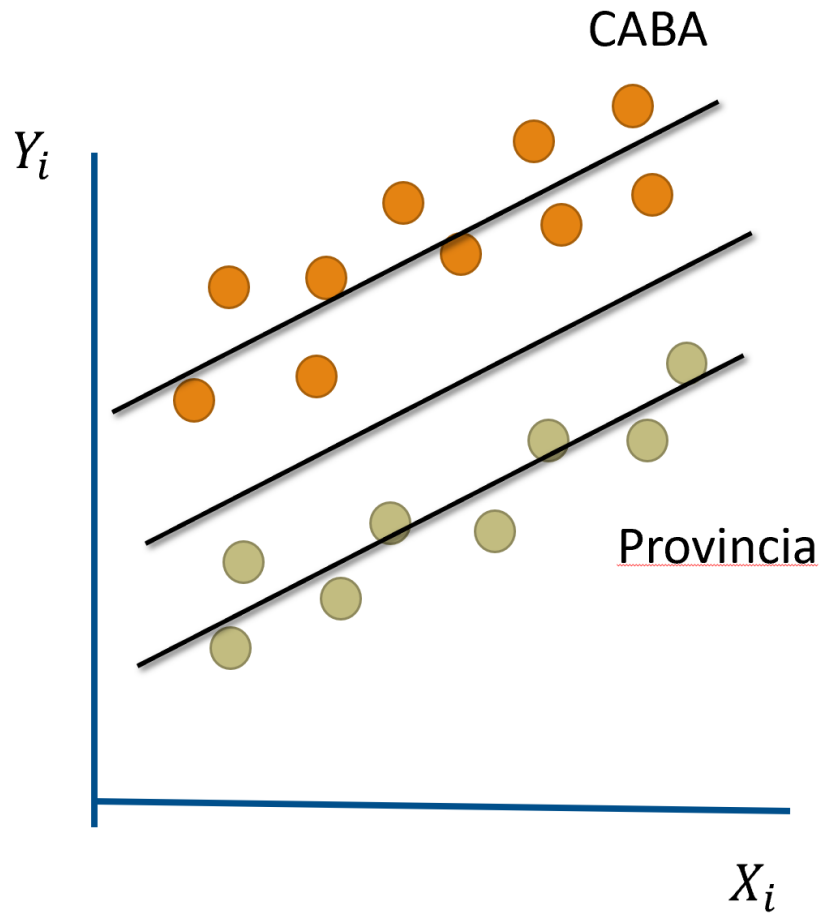
$$Y_i = \beta_0 + \beta_1 X_{1,i} + \varepsilon_i$$



Inclusión de variables categóricas como explicativas

Estimo 2 modelos

$$Y_i = \beta_0 + \beta_1 X_{1,i} + \varepsilon_i \quad \text{if } i \in \text{CABA}$$
$$Y_i = \beta_2 + \beta_3 X_{1,i} + \varepsilon_i \quad \text{if } i \in \text{Provincia}$$



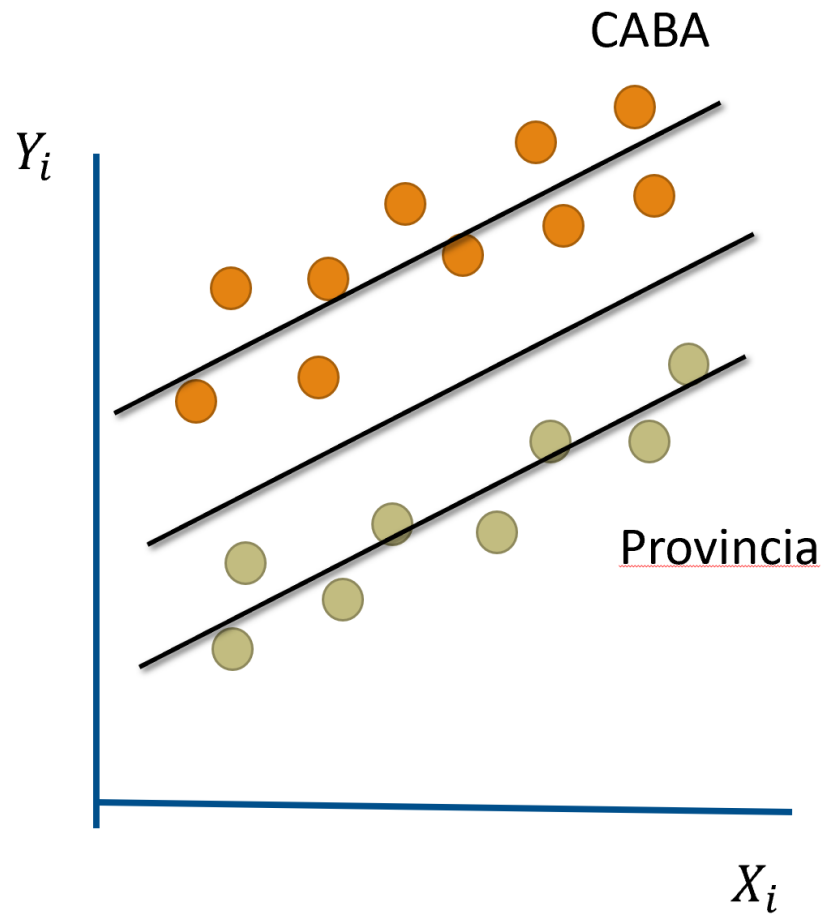
Inclusión de variables categóricas como explicativas

Inclusión de variable binaria

$$Y_i = \beta_0 + \beta_1 X_{1,i} + \beta_2 D_i + \varepsilon_i$$

where

$$D_i = 1 \quad \text{if } i \in \text{CABA}$$
$$D_i = 0 \quad \text{if } i \in \text{Provincia}$$



Inclusión de variables categóricas como explicativas

Múltiples categorías

- Qué pasaría si tengo múltiples (k)barrios?
- Ejemplo: *barrio de origen* (Barrio Almagro, Barrio Belgrano, Barrio Caballito, etc.).
- En este caso, es posible incluir una variable binaria para cada barrio menos uno ($k-1$). La k-ésima categoria queda capturada en β_0 como referencia.
- Ejemplo:

$$Y_i = \beta_0 + \beta_1 X_{1,i} + \beta_2 D_{1,i} + \cdots + \beta_{k-1} D_{k-1,i} + \varepsilon_i$$

where

$$\begin{aligned} D_{1,i} &= 1 \quad \text{if } i \in \text{Barrio Almagro and 0 de otro modo} \\ D_{2,i} &= 1 \quad \text{if } i \in \text{Barrio Belgrano and 0 de otro modo} \\ &\vdots \\ D_{k-1,i} &= 1 \quad \text{if } i \in \text{Villa Ortuzar and 0 de otro modo} \end{aligned}$$

Interpretando coeficientes en regresión múltiple

- En regresión múltiple, los coeficientes se interpretan como el cambio en la variable dependiente por unidad de cambio en la variable explicativa, manteniendo las otras variables explicativas constantes.
- Ejemplo Modelo 1:

$$\text{Precio}_i = \beta_0 + \beta_1 \text{Habitaciones}_{1,i} + \varepsilon_i$$

  - β_1 indica el incremento en el precio ($) por cada habitación adicional.
- Modelo 2:

$$\text{Precio}_i = \beta_0 + \beta_1 \text{Habitaciones}_{1,i} + \beta_2 \text{Baños}_{1,i} + \varepsilon_i$$

  - β_1 indica el incremento en el precio ($) por cada habitación adicional **cuando ya se tuvo en cuenta el efecto del número de baños**.

Consecuencias de la mala especificación de un modelo

Enfoque tradicional

Consecuencias de Mala Especificación del Modelo			
		Modelo Verdadero	
		$Y = \beta_1 + \beta_2 X_2 + u$	$Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + u$
Modelo Estimado	$\hat{Y} = b_1 + b_2 X_2$		
	$\hat{Y} = b_1 + b_2 X_2 + b_3 X_3$		

Consecuencias de la mala especificación de un modelo

Enfoque tradicional

Consecuencias de Mala Especificación del Modelo			
		Modelo Verdadero	
		$Y = \beta_1 + \beta_2 X_2 + u$	$Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + u$
Modelo Estimado	$\hat{Y} = b_1 + b_2 X_2$	Especificación correcta	
	$\hat{Y} = b_1 + b_2 X_2 + b_3 X_3$		Especificación correcta

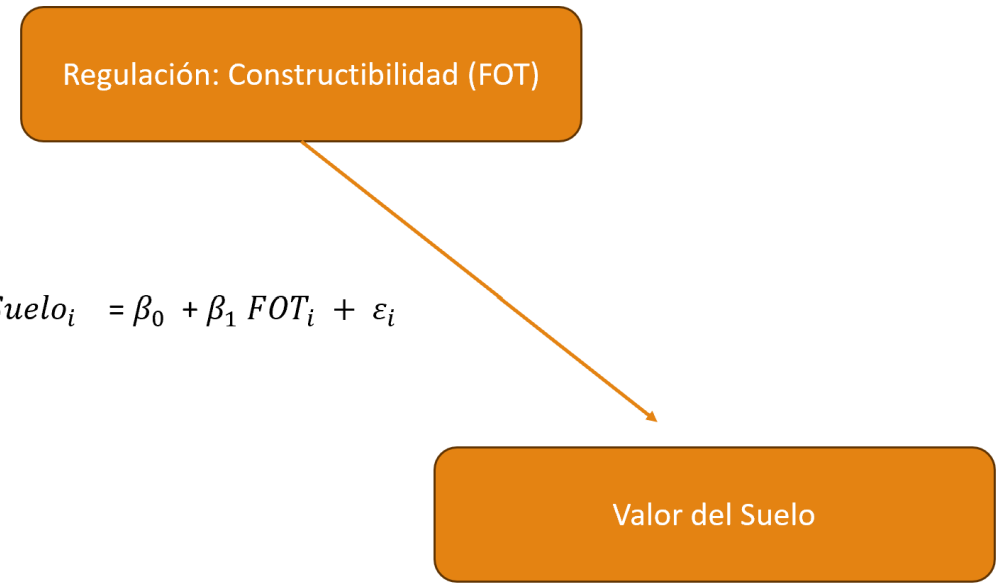
Consecuencias de la mala especificación de un modelo

Enfoque tradicional

Consecuencias de Mala Especificación del Modelo			
		Modelo Verdadero	
		$Y = \beta_1 + \beta_2 X_2 + u$	$Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + u$
Modelo Estimado	$\hat{Y} = b_1 + b_2 X_2$	Especificación correcta	Los coeficientes son sesgados. Los errores estándar invalidos.
	$\hat{Y} = b_1 + b_2 X_2 + b_3 X_3$		Especificación correcta

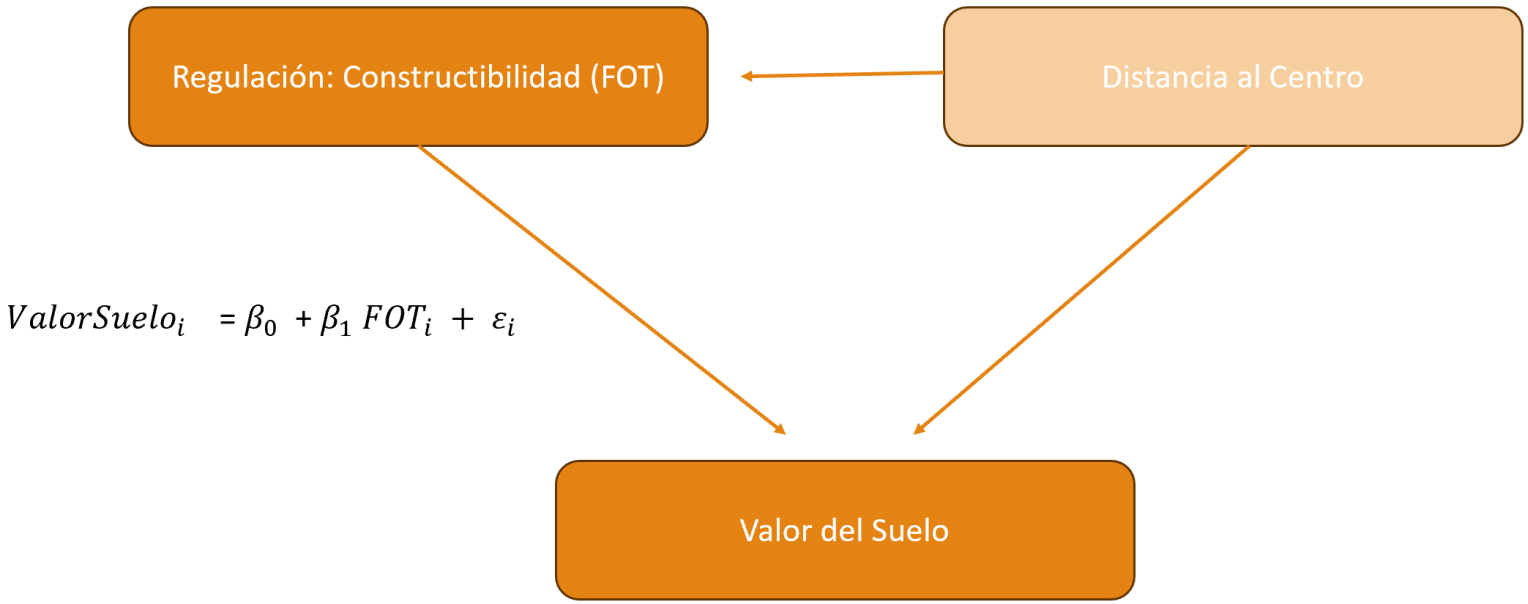
Consecuencias de la mala especificación de un modelo

Ejemplo efecto constructibilidad



Consecuencias de la mala especificación de un modelo

Ejemplo efecto constructibilidad



- $\hat{\beta}_1$ resulta sesgado.
- Tendriamos que estimar el modelo:

$$\text{Valor Suelo}_i = \beta_0 + \beta_1 \text{FOT}_{1,i} + \beta_2 \text{Distancia CBD}_{1,i} + \varepsilon_i$$

Consecuencias de la mala especificación de un modelo

Omisión de variable relevante

- Si el modelo verdadero es:

$$Y_i = \beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + \varepsilon_i$$

- Pero estimamos:

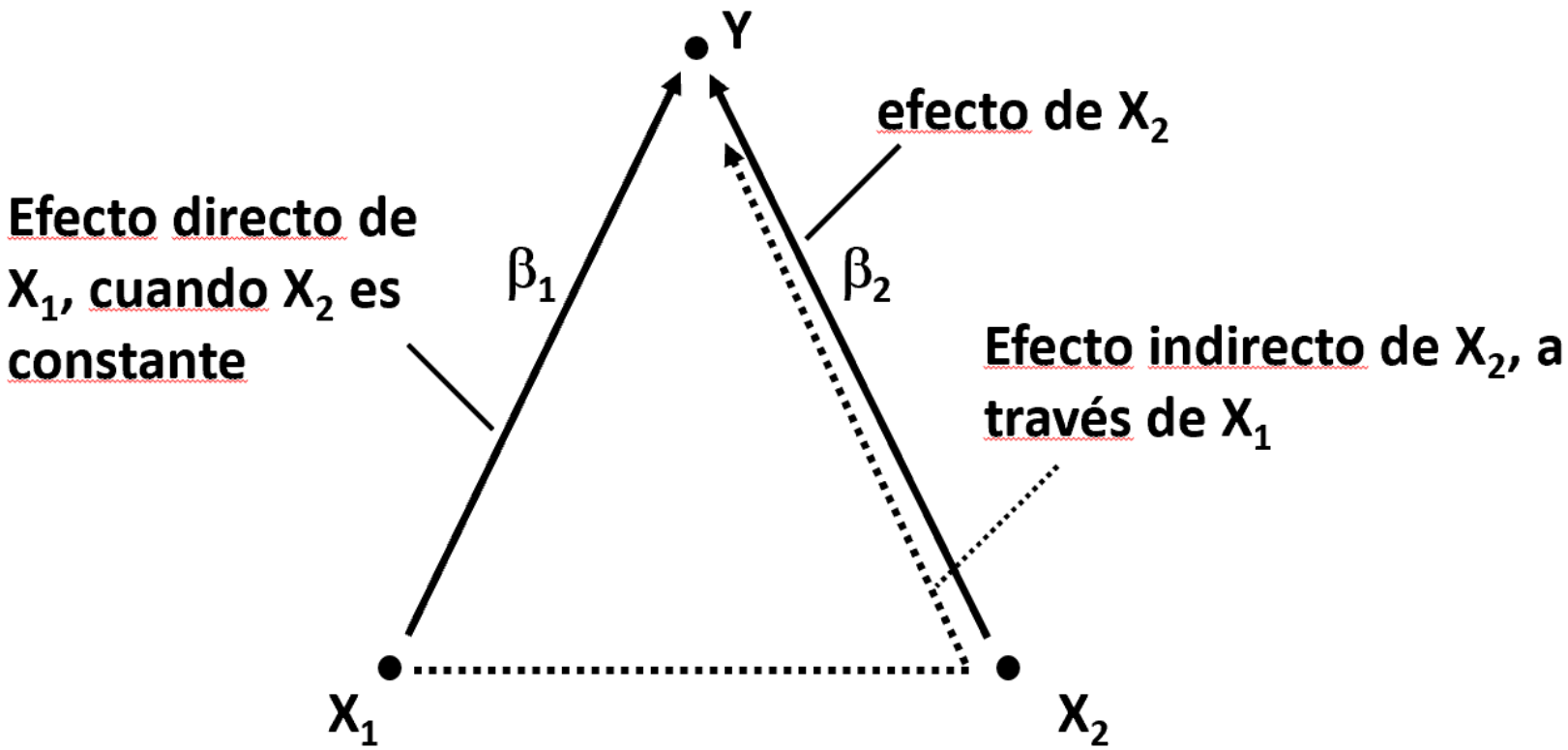
$$Y_i = b_0 + b_1 X_{1,i} + \varepsilon_i$$

- La teoría indica que:

$$E(\hat{b}_1) = \beta_1 + \underbrace{\beta_2 \frac{\text{Cov}(X_1, X_2)}{\text{Var}(X_1)}}_{\text{Sesgo}}$$

Consecuencias de la mala especificación de un modelo

Omisión de Variable Relevante



Consecuencias de la mala especificación de un modelo

Omisión de Variable Relevante

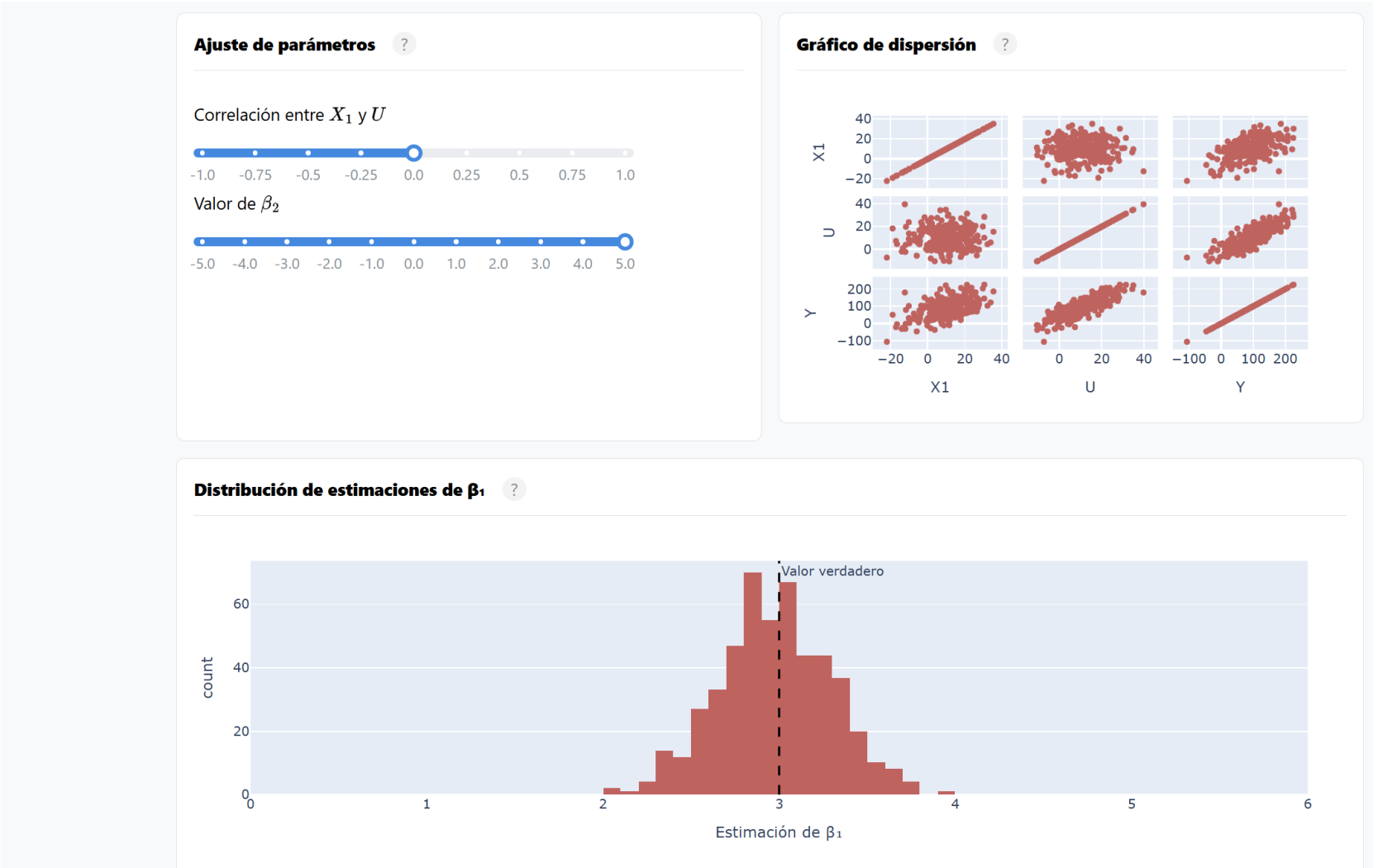


Figure: Simulación url: `https://simuecon.com/es/ch2_regresion_multiple/2_omitted_variable_bias.html`

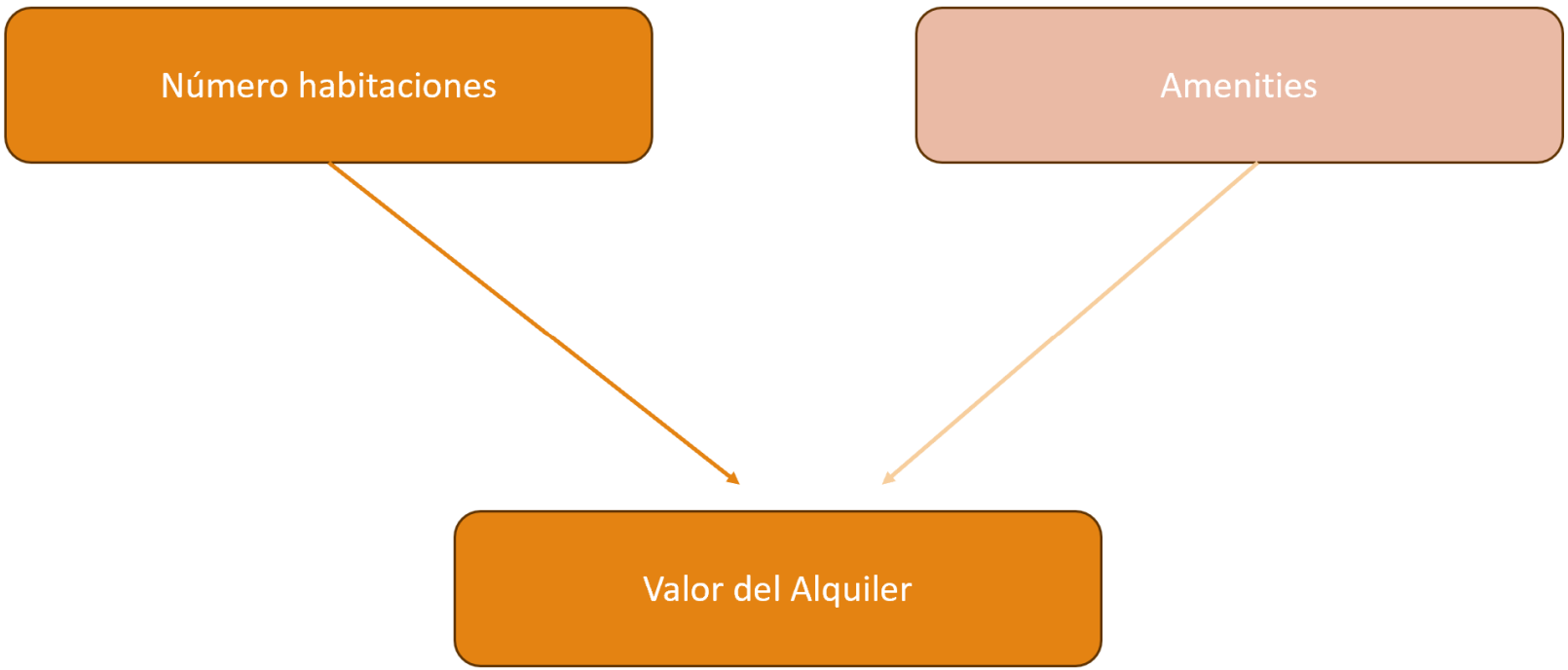
Consecuencias de la mala especificación de un modelo

Inclusión de variable irrelevante

Consecuencias de Mala Especificación del Modelo			
		Modelo Verdadero	
		$Y = \beta_1 + \beta_2 X_2 + u$	$Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + u$
Modelo Estimado	$\hat{Y} = b_1 + b_2 X_2$	Especificación correcta	Los coeficientes son sesgados. Los errores estándar invalidos.
	$\hat{Y} = b_1 + b_2 X_2 + b_3 X_3$	Los coeficientes son insesgados (en general), Pero ineficiente. Los errores estándar son válidos (en general)	Especificación correcta

Consecuencias de la mala especificación de un modelo

EjemploInclusión de variable irrelevante



Inclusión de variable irrelevante

- Si el modelo verdadero es:

$$Y_i = \beta_0 + \beta_1 X_{1,i} + \varepsilon_i$$

- Pero estimamos:

$$Y_i = b_0 + b_1 X_{1,i} + b_2 X_{2,i} + \varepsilon_i$$

- $\hat{b}_1$ es insesgado.
- Esto es asi porque el modelo estimado es el correcto cuando $\beta_2 = 0$!:

$$Y_i = \beta_0 + \beta_1 X_{1,i} + 0 \cdot X_{2,i} + \varepsilon_i$$

Inclusión de variable irrelevante

- Sin embargo, existe una pérdida de eficiencia. Esto puede verse en la **varianza del estimador en regresión múltiple**:

$$\text{Var}(\hat{b}_1) = \frac{\sigma^2}{n \cdot \text{Var}(X_1)} \cdot \frac{1}{1 - \text{Corr}(X_1, X_2)^2}$$

- La correlación entre X_1 y X_2 es la que determina la pérdida de eficiencia.
- Intuitivamente, si X_1 y X_2 tienen alguna correlación, al modelo le cuesta más distinguir el efecto de X_1 sobre Y.