Clases 4 y 5: OLS

Denise Laroze

5 de octubre de 2018

CESS - Universidad de Santiago
denise.laroze@usach.cl

Resumen de contenidos

Not so (Ordinary) Least Squares

Supuestos - Recordar

1. Linearidad
2. Full Rank (relevante para regressiones múltiples)
3. Exogeneidad de covariables $E[\epsilon_i|X = 0]$. el valor esperado de las discrepancias o "errores" en la observación i no es una función de las covariables.
4. homoskedasticidad y no auto-correlación de los "errores": Que ϵ_i tiene varianza constante y no está correlacionado con ninguna otra ϵ_j
5. Que los datos son generados exógenamente: el proceso de generación de los datos es externo a los supuestos del modelo, lo que implica que es externo a el/los procesos que generan ϵ. En otras palabras, **las covariables y los disturbios no están correlacionados**.
6. Los disturbios ("errores") se distribuyen normalmente. Más específicamente $\epsilon|X \sim N(0, \sigma^2)$

Importancia de ϵ

$$
\begin{aligned}
b &= (X'X)^{-1}X'Y. \\
&= (X'X)^{-1}X'(X\beta + \epsilon). \\
&= (X'X)^{-1}(X'X)\beta + (X'X)^{-1}X'\epsilon. \\
&= \beta + (X'X)^{-1}X'\epsilon.
\end{aligned}
$$

Take expected values over X.

$$E[b|X] = \beta + E[(X'X)^{-1}X'\epsilon|X].$$

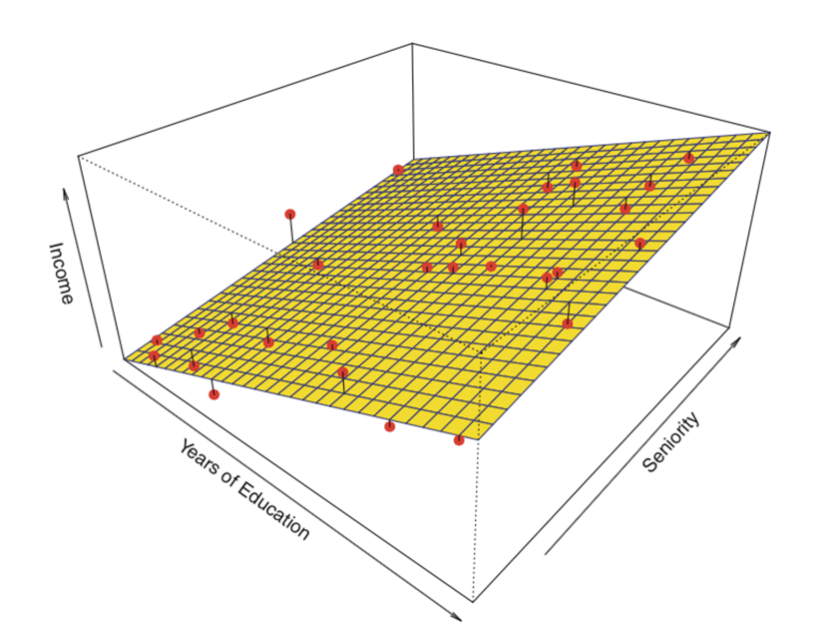
$$E[b|X] = \beta.$$

Regresión Múltiple

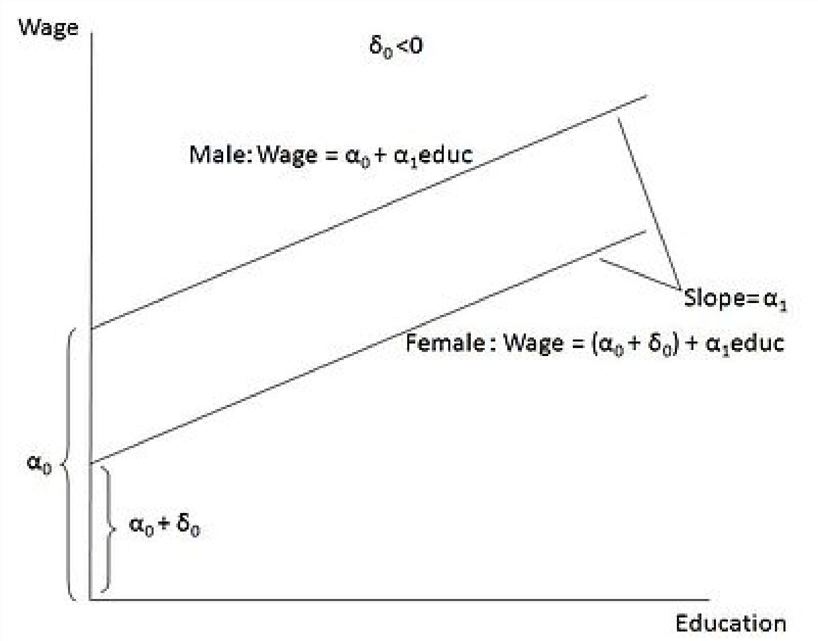
Regresión lineal múltiple

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + ... \beta_n X_n + \epsilon$$

Regresión lineal múltiple



Dummy variables

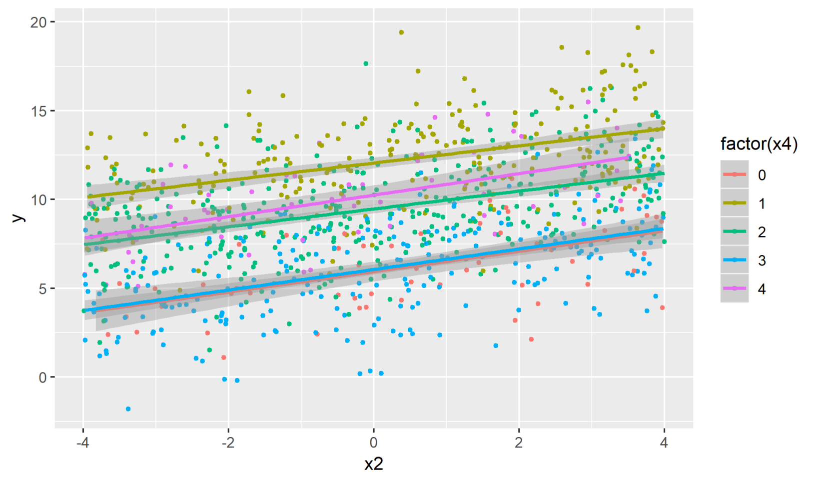


Regresión lineal múltiple

¿Cómo se interpretan los siguientes resultados?

	(1)	(2)
x1	0.112***	0.099***
	(0.006)	(0.006)
x2	0.452***	0.525***
	(0.027)	(0.027)
x3	1.472***	1.540***
	(0.125)	(0.124)
factor(x4)1		6.041***
		(0.280)
factor(x4)2		3.494***
		(0.271)
factor(x4)3		0.165
		(0.279)
factor(x4)4		4.146***
		(0.360)
Constant	4.929***	4.750***
	(0.094)	(0.259)
Observations	1,000	1,000
R^2	0.421	0.690
Adjusted R^2	0.419	0.687
F Statistic	241.250*** (df = 3; 996)	314.843*** (df = 7; 992)

Variables Categóricas



Selección del modelo

¿Cómo evaluar si una variable aporta a la explicación de la varianza de Y?

Selección del modelo

¿Cómo evaluar si una variable aporta a la explicación de la varianza de Y?

t-test

Selección del modelo

¿Cómo evaluar si una variable aporta a la explicación de la varianza de Y?

t-test

¿Cómo saber si una variable categórica o conjunto de variables son un aporte al modelo de estimación o no?

Selección del modelo

¿Cómo evaluar si una variable aporta a la explicación de la varianza de Y?

t-test

¿Cómo saber si una variable categórica o conjunto de variables son un aporte al modelo de estimación o no?

F-test

$$F = \frac{(R_a^2 - R_0^2)/(k_a - k_0)}{(1 - R_a^2)/[n - (K_a + 1)]}$$

$$F = \frac{(R_{cambio}^2)/df_{cambio}}{(1 - R_a^2)/[n - (K_a + 1)]}$$

Donde R_a^2 corresponde al del modelo completo y R_0^2 al del modelo restringido.

Buen use del F-test

La prueba de F sólo se puede hacer si el modelo restringido es un subconjunto del modelo completo.

$$1) y = \alpha + \beta_1 x1 + \beta_2 x2 + \beta_3 x3 + \beta_4 x4 + \epsilon$$

$$2) y = \alpha + \beta_1 x1 + \beta_2 x2 + \epsilon$$

$$3) y = \alpha + \beta_1 x1 + \beta_2 x2 + \beta_3 x3 + \beta_5 x5 + \epsilon$$

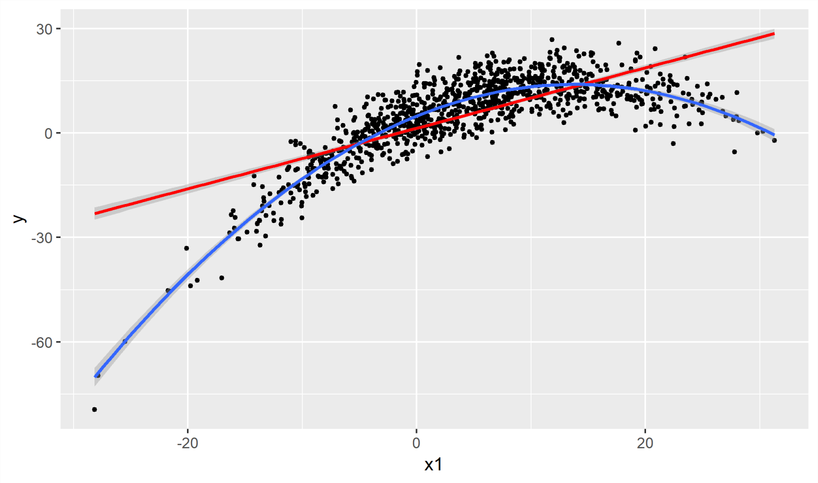
Comparar modelo 1) y 2) **Sí**, modelo 1) y 3) **NO**

Functional Form

Efectos no lineales

Función real: $y < -5 + 1{,}3 * x1 - 0{,}05 * x1^2 + e$

	(1)	(2)
x1	0.869^{***}	1.318^{***}
	(0.026)	(0.020)
I(x1^2)		-0.048^{***}
		(0.001)
Constant	1.373^{***}	4.849^{***}
	(0.278)	(0.198)
Observations	1,000	1,000
R^2	0.530	0.811
Adjusted R^2	0.529	0.810
F Statistic	$1{,}124.753^{***}$ (df = 1; 998)	$2{,}136.135^{***}$ (df = 2; 997)
Note:		*p<0.1; **p<0.05; ***p<0.01

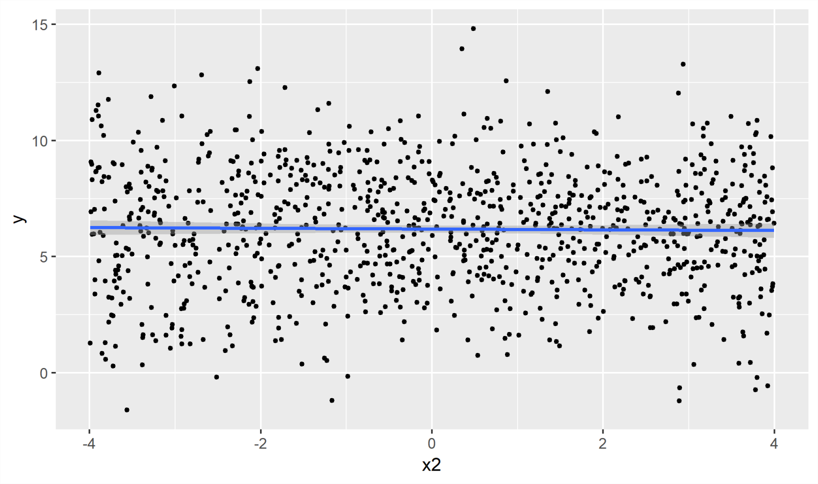

30
0
-30
-60
y
-20
0
20
x1

Interacciones

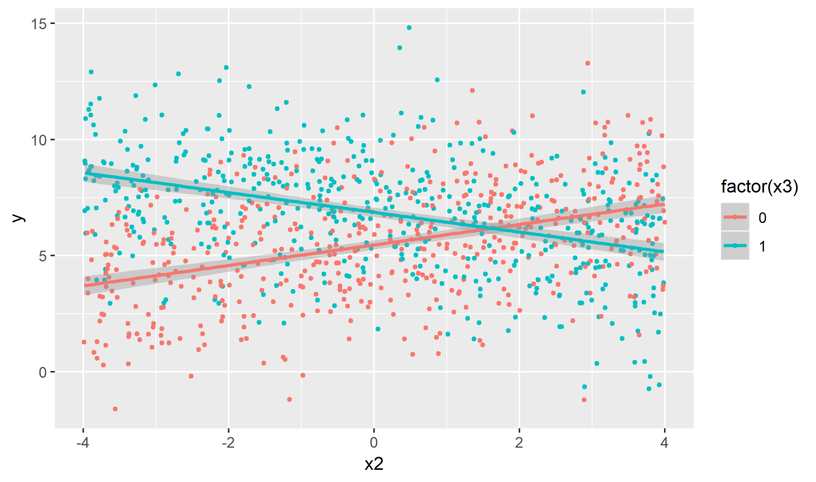
Función real: $y < -5 + 0{,}1 * x1 + 0{,}5 * x2 + 1{,}5 * x3 + -0{,}9 * x2 * x3 + e$

	(1)	(2)
x1	0.109^{***}	0.112^{***}
	(0.007)	(0.006)
x2	-0.019	0.450^{***}
	(0.030)	(0.039)
x3	1.381^{***}	1.471^{***}
	(0.142)	(0.125)
x2*x3		-0.897^{***}
		(0.053)
Constant	5.004^{***}	4.929^{***}
	(0.106)	(0.094)
Observations	1,000	1,000
R^2	0.238	0.406
Adjusted R^2	0.235	0.403
Residual Std. Error	2.240 (df = 996)	1.979 (df = 995)
F Statistic	103.438^{***} (df = 3; 996)	169.846^{***} (df = 4; 995)
Note:		*p<0.1; **p<0.05; ***p<0.01

Sin interacción - cuando existe



Con interacción - cuando existe



Efecto Marginal de X2

$y < -\alpha + \beta_1 * x1 + \beta_2 * x2 + \beta_3 * x3 + \beta_4 * x2 * x3 + e$

$y < -5 + 0,1 * x1 + 0,5 * x2 + 1,5 * x3 - 0,9 * x2 * x3 + e$

$$\frac{dE[y|x]}{dx2} = 0,5 - 0,9 * x3$$

Efecto Marginal de X2

$$y < -\alpha + \beta_1 * x1 + \beta_2 * x2 + \beta_3 * x3 + \beta_4 * x2 * x3 + e$$

$$y < -5 + 0{,}1 * x1 + 0{,}5 * x2 + 1{,}5 * x3 - 0{,}9 * x2 * x3 + e$$

$$\frac{dE[y|x]}{dx2} = 0{,}5 - 0{,}9 * x3$$

¿Pero cuál es la varianza de x2? ¿Por qué importa?

Efecto Marginal de X2

$y < -\alpha + \beta_1 * x1 + \beta_2 * x2 + \beta_3 * x3 + \beta_4 * x2 * x3 + e$

$y < -5 + 0{,}1 * x1 + 0{,}5 * x2 + 1{,}5 * x3 - 0{,}9 * x2 * x3 + e$

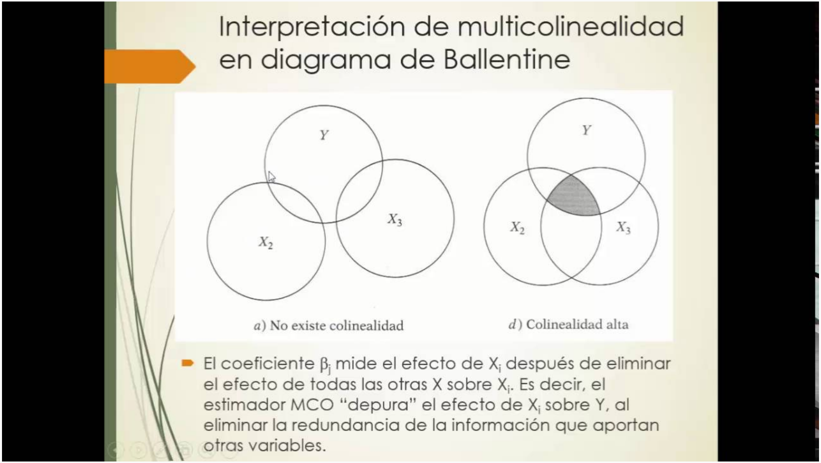
$$\frac{dE[y|x]}{dx2} = 0{,}5 - 0{,}9 * x3$$

¿Pero cuál es la varianza de x2? ¿Por qué importa?

Delta method

$Var(\beta_2 - \beta_4 * x3) = 1^2 + Var(\beta_2) + x3^2 + Var(\beta_4) + (2)(1)(x3)Cov(\beta_2, \beta_4)$

Multicolinealidad



Violaciones de supuestos - endogeneidad

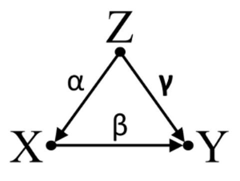
Temas a considerar

Causalidad: $X - - > Y$

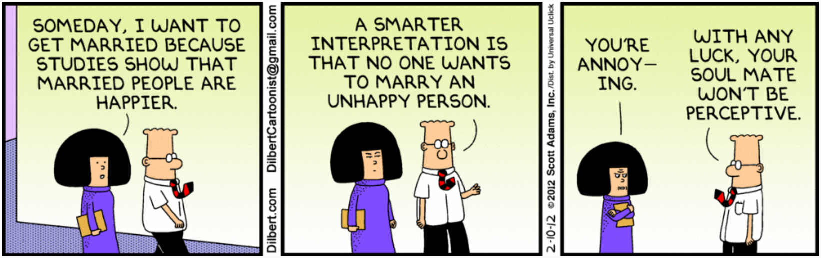
Causalidad inversa: $X < - - Y$

Simultaneidad: $X < -- > Y$

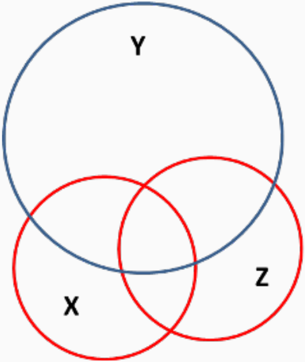
Variables Omitidas:



Causalidad inversa

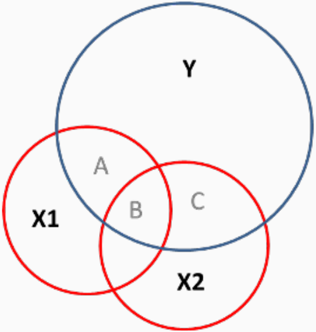


Omitted Variable Bias



$$Y = b_0 + b_1 X + e$$
$$Y = \beta_0 + \beta_1 X + \beta_2 Z + \varepsilon$$
$$b_1 \neq \beta_1$$

Omitted Variable Bias



$$b_1 \approx A + B$$

$$\beta_1 \approx A$$

$$bias \approx B$$

$$\beta_1 = b_1 - bias \approx (A + B) - B$$