Universidad de los Andes

Escuela de Gobierno
Alberto Lleras Camargo

Métodos A para Públicos I

EGOB 2101

Andrés Ham
a.ham@uniandes.edu.co

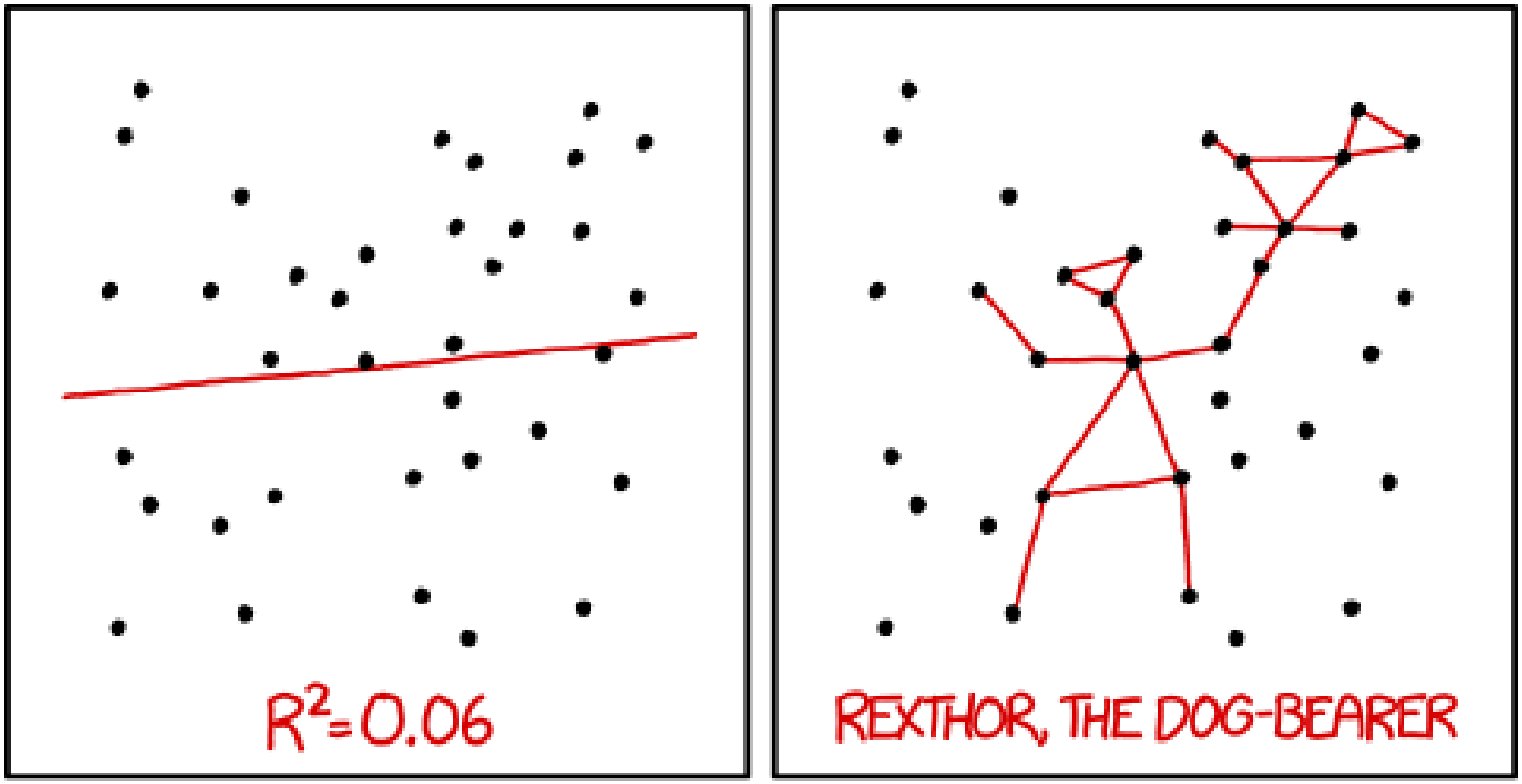
2019-1

20 de febrero del 2018

Agenda de hoy

1. ¿Qué dice y qué no dice una regresión?
2. Enchúlame el Nokia 3310: Regresión multivariada
3. El mensaje del día

Previously...



I DON'T TRUST LINEAR REGRESSIONS WHEN IT'S HARDER TO GUESS THE DIRECTION OF THE CORRELATION FROM THE SCATTER PLOT THAN TO FIND NEW CONSTELLATIONS ON IT.

Previously...

- La regresión líneal es una herramienta que permite estimar la relación entre una variable x y un resultado de interés y.
- Sin embargo, una regresión provee información *parcial*, la cuál es útil hasta cierto punto. Es decir, tiene debilidades.
- Hoy comenzaremos a discutir algunas de estas limitaciones, y por qué estimar una regresión multivariada puede mitigarlas.

La relación entre edad e ICFES

- ¿Cuál era la interpretación de este cuadro?

	Puntaje ICFES
Edad ($\hat{\beta}$)	-0.184
	(0.014)***
Intercepto ($\hat{\alpha}$)	3.227
	(0.255)***
R^2	0.027
Observaciones	6,316

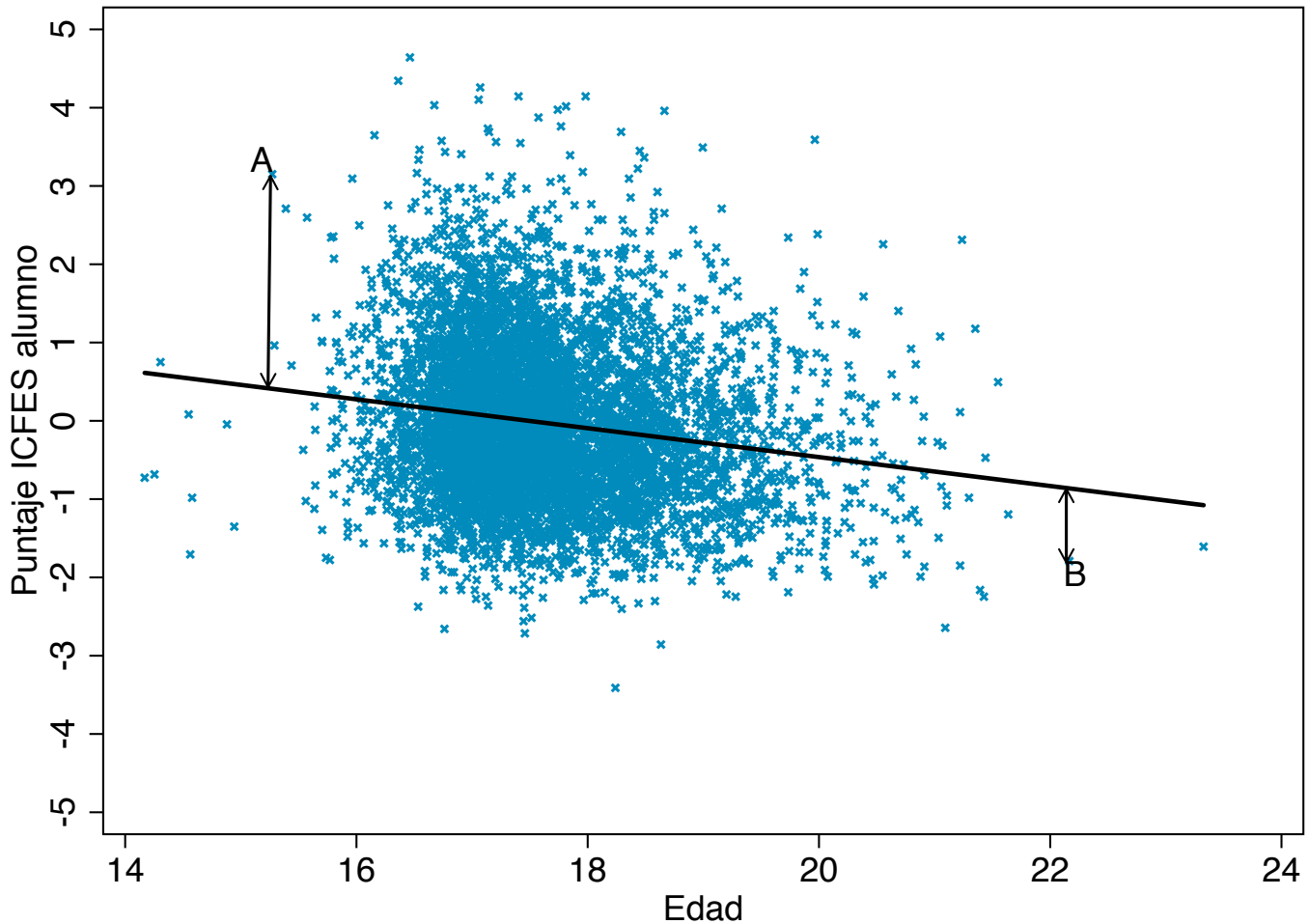
Calculos realizados utilizando datos de Bonilla, Bottan y Ham (2017).
*** Significativo al 1 por ciento, ** 5 por ciento y * 10 por ciento.

Recordemos la mecánica de la regresión

- ¿Qué hace el método gráficamente?

Recordemos la mecánica de la regresión

- ¿Qué hace el método gráficamente?



Deconstruyendo el proceso

- Una regresión minimiza errores en promedio, o sea, asume que todo aquello que no es edad no afecta al puntaje ICFES.

Deconstruyendo el proceso

- Una regresión minimiza errores en promedio, o sea, asume que todo aquello que no es edad no afecta al puntaje ICFES.

$$\underbrace{y}_{\text{Puntaje ICFES}} = \underbrace{\alpha}_{\text{Intercepto}} + \underbrace{\beta}_{\text{Coeficiente}} \times \underbrace{x}_{\text{Edad}} + \underbrace{u}_{\text{Otras cosas}}$$

Deconstruyendo el proceso

- Una regresión minimiza errores en promedio, o sea, asume que todo aquello que no es edad no afecta al puntaje ICFES.

$$\underbrace{y}_{\text{Puntaje ICFES}} = \underbrace{\alpha}_{\text{Intercepto}} + \underbrace{\beta}_{\text{Coeficiente}} \times \underbrace{x}_{\text{Edad}} + \underbrace{u}_{\text{Otras cosas}}$$

- Un coeficiente estimado nos da el efecto causal cuando la relación entre la variable explicada y explicativa es **exógena**.

Los errores se eliminan en promedio

- ¿Qué exactamente quiere decir esto?

Los errores se eliminan en promedio

- ¿Qué exactamente quiere decir esto?
- Basicamente, dice que la “esperanza” es que al controlar por edad, el resto de cosas que no incluimos en la regresión no importan.

Los errores se eliminan en promedio

- ¿Qué exactamente quiere decir esto?
- Basicamente, dice que la "esperanza" es que al controlar por edad, el resto de cosas que no incluimos en la regresión no importan.

$$\mathbb{E}[u|x] = 0$$

Los errores se eliminan en promedio

- ¿Qué exactamente quiere decir esto?
- Basicamente, dice que la "esperanza" es que al controlar por edad, el resto de cosas que no incluimos en la regresión no importan.

$$\mathbb{E}[u|x] = 0$$

- ¿Y esto implica que el efecto de la edad sobre el puntaje es?

Los errores se eliminan en promedio

- ¿Qué exactamente quiere decir esto?
- Basicamente, dice que la "esperanza" es que al controlar por edad, el resto de cosas que no incluimos en la regresión no importan.

$$\mathbb{E}[u|x] = 0$$

- ¿Y esto implica que el efecto de la edad sobre el puntaje es?

$$\mathbb{E}[y|x] = \beta$$

Exogeneidad en la estadística

- ¿Todos han escuchado hablar de variables **endógenas** y **exógenas**?

Exogeneidad en la estadística

- ¿Todos han escuchado hablar de variables **endógenas** y **exógenas**?
- En materia estadística, estos conceptos se definen distinto:

Exogeneidad en la estadística

- ¿Todos han escuchado hablar de variables **endógenas** y **exógenas**?
- En materia estadística, estos conceptos se definen distinto:
  - *Variable endógena*: una variable que **SÍ** está correlacionada con el término de error.
  - *Variable exógena*: una variable que **NO** está correlacionada con el término de error.

Exogeneidad en la estadística

- Por ejemplo, que puede explicar que un alumno(a) tome el ICFES a mayor edad?

Exogeneidad en la estadística

- Por ejemplo, que puede explicar que un alumno(a) tome el ICFES a mayor edad? Repitencia.

Exogeneidad en la estadística

- Por ejemplo, que puede explicar que un alumno(a) tome el ICFES a mayor edad? Repitencia.

- Entonces, el verdadero modelo es:

Exogeneidad en la estadística

- Por ejemplo, que puede explicar que un alumno(a) tome el ICFES a mayor edad? Repitencia.

- Entonces, el verdadero modelo es:

$$\underbrace{y}_{\text{Puntaje ICFES}} = \alpha + \beta_1 \times \underbrace{x_1}_{\text{Edad}} + \beta_2 \times \underbrace{x_2}_{\text{Repitencia}} + \underbrace{u}_{\text{Otras cosas}}$$

Exogeneidad en la estadística

- Por ejemplo, que puede explicar que un alumno(a) tome el ICFES a mayor edad? Repitencia.
- Entonces, el verdadero modelo es:

$$\underbrace{y}_{\text{Puntaje ICFES}} = \alpha + \beta_1 \times \underbrace{x_1}_{\text{Edad}} + \beta_2 \times \underbrace{x_2}_{\text{Repitencia}} + \underbrace{u}_{\text{Otras cosas}}$$

- En este caso, la repitencia es una **variable omitida**.

Exogeneidad en la estadística

- Pero, nosotros estimamos esto:

Exogeneidad en la estadística

- Pero, nosotros estimamos esto:

$$\underbrace{y}_{\text{Puntaje ICFES}} = \alpha + \beta_1 \underbrace{x_1}_{\text{Edad}} + \underbrace{e}_{\text{Error}}$$

Exogeneidad en la estadística

- Pero, nosotros estimamos esto:

$$\underbrace{y}_{\text{Puntaje ICFES}} = \alpha + \beta_1 \underbrace{x_1}_{\text{Edad}} + \underbrace{e}_{\text{Error}}$$

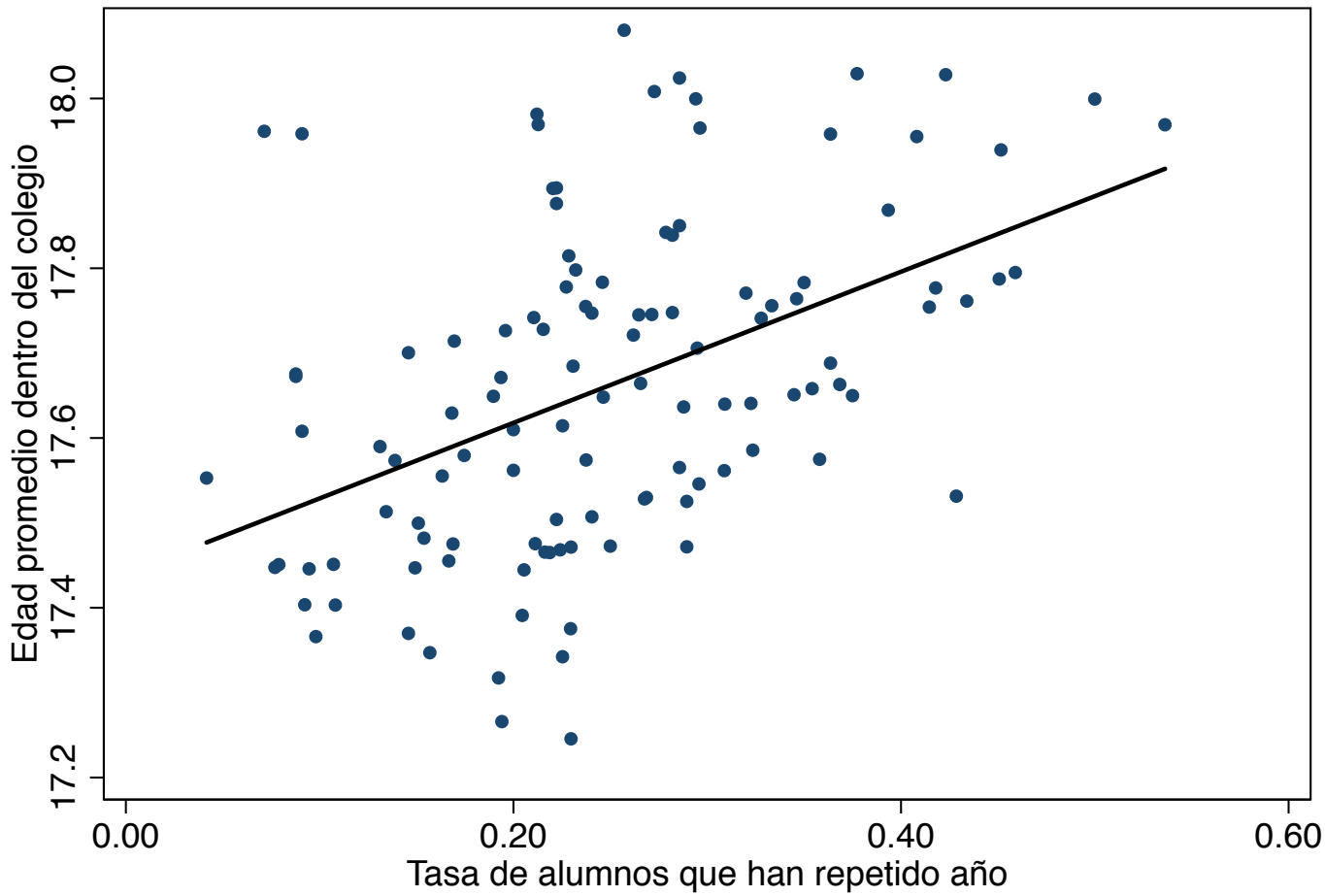
- Donde $e = \beta_2 \underbrace{x_2}_{\text{Repitencia}} + \underbrace{u}_{\text{Otras cosas}}$

¡Tu modelo es endógeno!

- La regresión estaría bien si la edad y la repitencia no están correlacionadas.

¡Tu modelo es endógeno!

- La regresión estaría bien si la edad y la repitencia no están correlacionadas.



Coeficiente de correlación: 0.465

¿Dónde veo esa endogeneidad?

- Si calculamos el efecto de la edad sobre puntaje, este sería:

$$
\begin{aligned}
\mathbb{E}[y|x_1] &= \beta + \mathbb{E}[e|x_1] \\
&= \beta + \mathbb{E}[\beta_2 x_2 + u|x_1] \\
&= \beta + \mathbb{E}[\beta_2 x_2|x_1] + \underbrace{\mathbb{E}[u|x_1]}_{=0} \\
&= \underbrace{\beta}_{\text{Efecto de la edad en ICFES}} + \underbrace{\pi}_{\text{Relación entre repitencia y edad}}
\end{aligned}
$$

¿Dónde veo esa endogeneidad?

- Si calculamos el efecto de la edad sobre puntaje, este sería:

$$
\begin{aligned}
\mathbb{E}[y|x_1] &= \beta + \mathbb{E}[e|x_1] \\
&= \beta + \mathbb{E}[\beta_2 x_2 + u|x_1] \\
&= \beta + \mathbb{E}[\beta_2 x_2|x_1] + \underbrace{\mathbb{E}[u|x_1]}_{=0} \\
&= \underbrace{\beta}_{\text{Efecto de la edad en ICFES}} + \underbrace{\pi}_{\text{Relación entre repitencia y edad}}
\end{aligned}
$$

- Entonces, ¡estaríamos sacando conclusiones erróneas!

¿Qué tan grande es el sesgo?

$$\mathbb{E}[y|x_1] = \underbrace{\beta}_{\text{Efecto de la edad en ICFES}} + \underbrace{\pi}_{\text{Relación entre repitencia y edad}}$$

- No sabemos exactamente qué tan grande es π, pero podemos definir en qué dirección se mueve.

¿Qué tan grande es el sesgo?

$$\mathbb{E}[y|x_1] = \underbrace{\beta}_{\text{Efecto de la edad en ICFES}} + \underbrace{\pi}_{\text{Relación entre repitencia y edad}}$$

- No sabemos exactamente qué tan grande es π, pero podemos definir en qué dirección se mueve.

- Dos preguntas claves:

¿Qué tan grande es el sesgo?

$$\mathbb{E}[y|x_1] = \underbrace{\beta}_{\text{Efecto de la edad en ICFES}} + \underbrace{\pi}_{\text{Relación entre repitencia y edad}}$$

- No sabemos exactamente qué tan grande es π, pero podemos definir en qué dirección se mueve.

- Dos preguntas claves:

  1. ¿En qué dirección creemos que afecta la edad al ICFES?

¿Qué tan grande es el sesgo?

$$\mathbb{E}[y|x_1] = \underbrace{\beta}_{\text{Efecto de la edad en ICFES}} + \underbrace{\pi}_{\text{Relación entre repitencia y edad}}$$

- No sabemos exactamente qué tan grande es π, pero podemos definir en qué dirección se mueve.

- Dos preguntas claves:

  1. ¿En qué dirección creemos que afecta la edad al ICFES?
  2. ¿En qué dirección se mueve la relación entre edad y repitencia?

¿Qué tan grande es el sesgo?

$$\mathbb{E}[y|x_1] = \underbrace{\beta}_{\text{Negativo}} + \underbrace{\pi}_{\text{Positivo}}$$

- En este caso, van en direcciones opuestas.

¿Qué tan grande es el sesgo?

$$\mathbb{E}[y|x_1] = \underbrace{\beta}_{\text{Negativo}} + \underbrace{\pi}_{\text{Positivo}}$$

- En este caso, van en direcciones opuestas.
- ¿Qué pasaría con nuestro cálculo de β si ignoramos la repitencia?

Comparemos estimaciones

	Sin repitencia	Con repitencia
Edad ($\hat{\beta_1}$)	-0.184	-0.227
	(0.014)***	(0.017)***
Repitencia ($\hat{\beta_2}$)	—	0.165
		(0.034)***
Intercepto ($\hat{\alpha}$)	3.227	3.935
	(0.255)***	(0.296)***
R^2	0.027	0.030
Observaciones	6,316	6,269

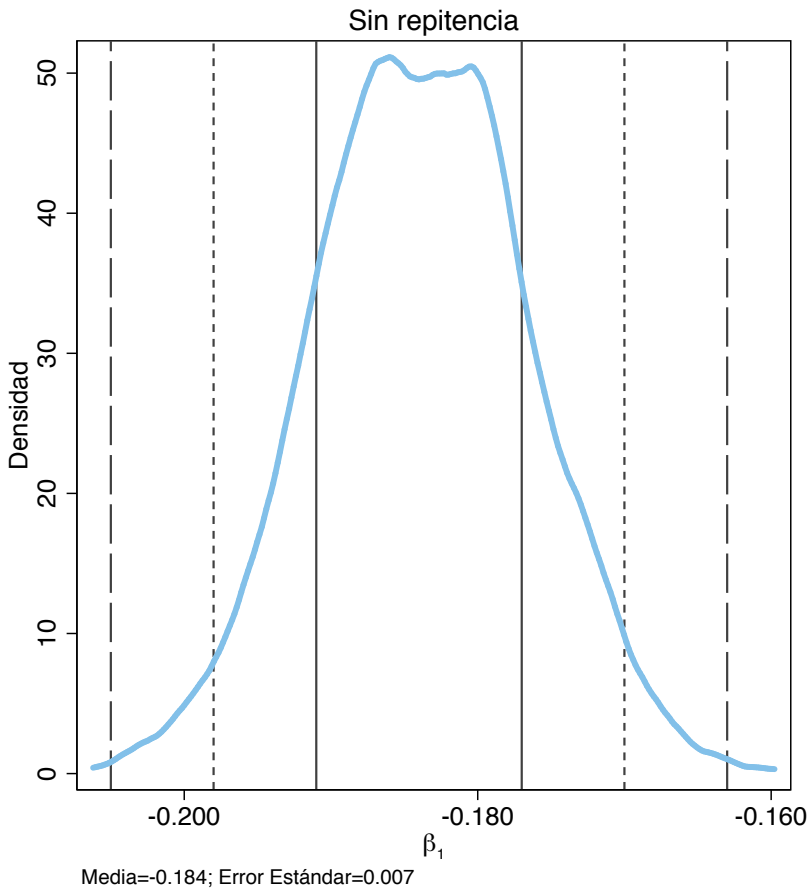
Calculos realizados utilizando datos de Bonilla, Bottan y Ham (2017).
*** Significativo al 1 por ciento, ** 5 por ciento y * 10 por ciento.

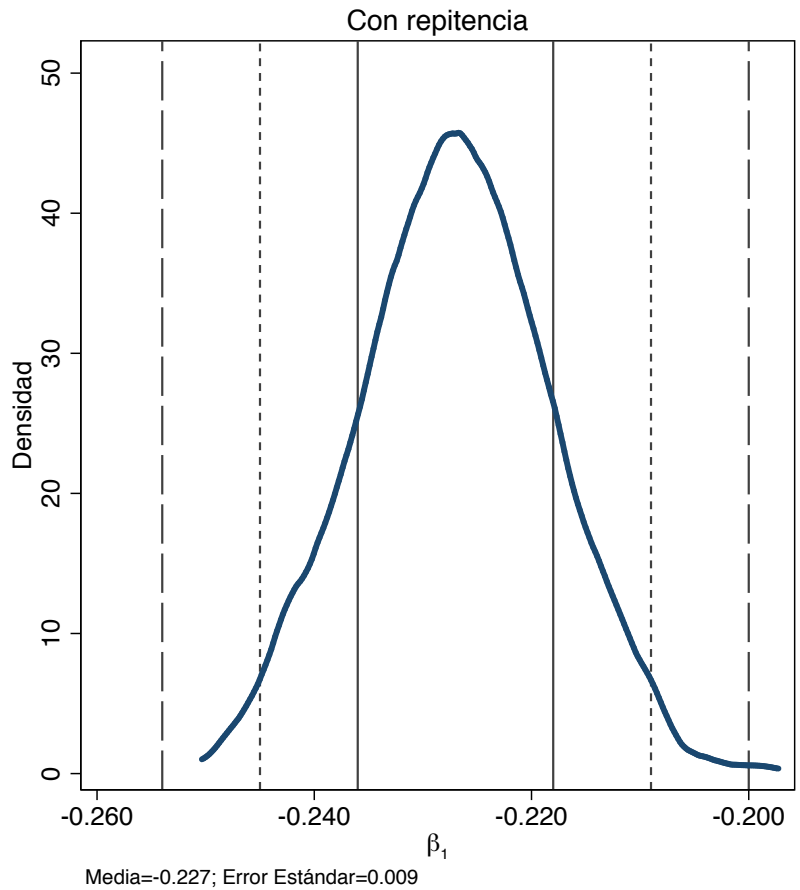
El costo de olvidar variables en el análisis

- ¿Son realmente diferentes ambas versiones de $\hat{\beta}_1$?

El costo de olvidar variables en el análisis

- ¿Son realmente diferentes ambas versiones de $\hat{\beta}_1$?



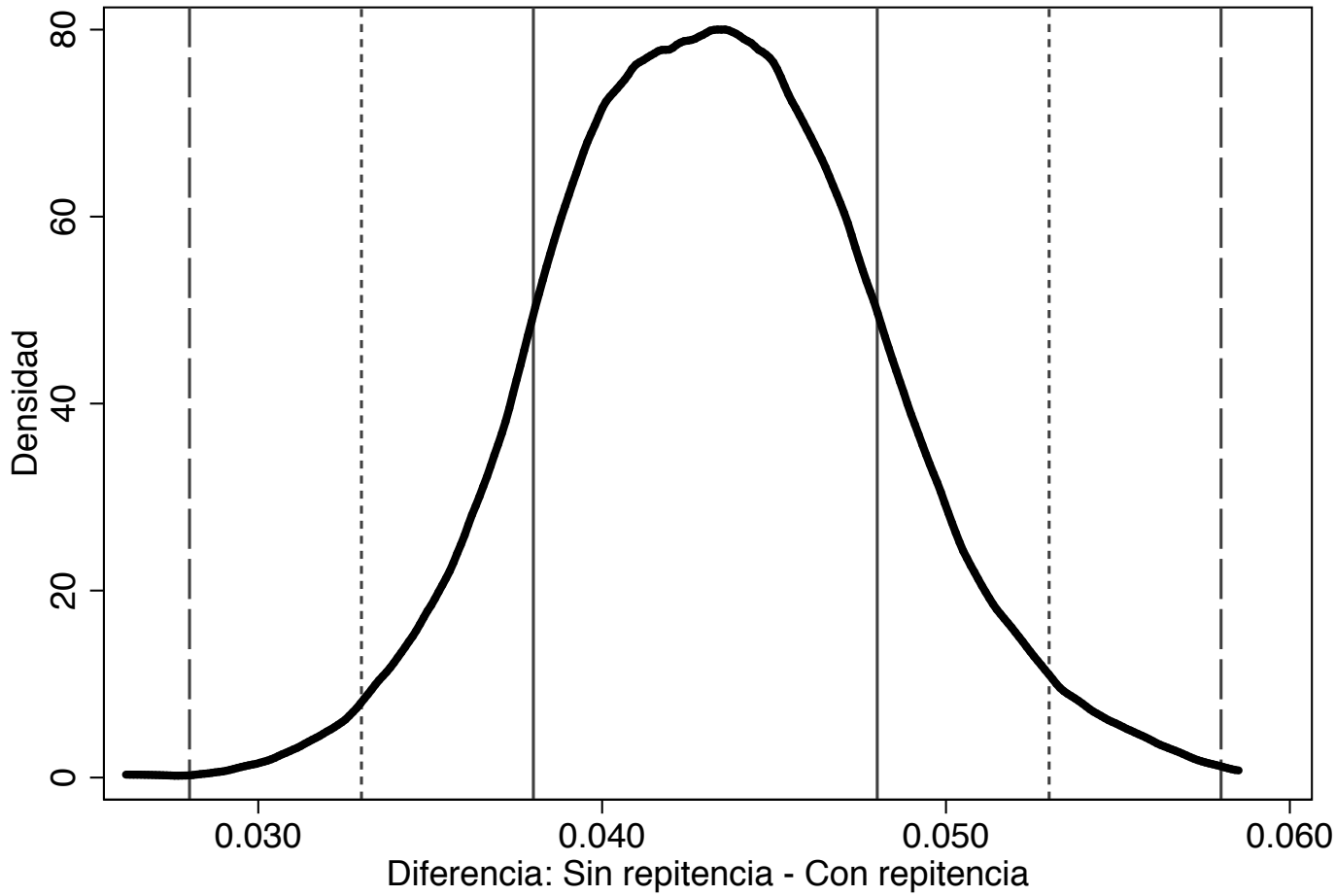


El costo de olvidar variables en el análisis

- ¿Y si tomamos la diferencia, será $\neq 0$?

El costo de olvidar variables en el análisis

- ¿Y si tomamos la diferencia, será $\neq 0$?



¿Qué aprendemos de todo esto?

- Es importante definir bien el modelo, pues la regresión hace su trabajo según el insumo que recibe.

¿Qué aprendemos de todo esto?

- Es importante definir bien el modelo, pues la regresión hace su trabajo según el insumo que recibe.
- Dependiendo qué cosas olvidamos o no podemos incluir, el coeficiente de interés puede ser más grande o pequeño.

¿Qué aprendemos de todo esto?

- Es importante definir bien el modelo, pues la regresión hace su trabajo según el insumo que recibe.
- Dependiendo qué cosas olvidamos o no podemos incluir, el coeficiente de interés puede ser más grande o pequeño.
- Entonces, ¿qué podemos hacer?

Incluyamos más variables explicativas

- El rendimiento escolar promedio se puede explicar por muchos factores diferentes.

Incluyamos más variables explicativas

- El rendimiento escolar promedio se puede explicar por muchos factores diferentes.
- Entonces, con buenos datos podemos intentar reducir el sesgo incluyendo más variables explicativas.

Incluyamos más variables explicativas

- El rendimiento escolar promedio se puede explicar por muchos factores diferentes.
- Entonces, con buenos datos podemos intentar reducir el sesgo incluyendo más variables explicativas.
- La idea es que si incluimos más variables explicativas, quedan menos cosas en el término de error para jodernos la vida.

$$y = \alpha + \beta_1 x_1 + ... + \beta_k x_k + u \qquad \text{Dónde } k = 1, ..., K$$

Por ejemplo

Edad	-0.171
	(0.016)***
Hombres	0.349
	(0.024)***
Repitencia	0.022
	(0.033)
Padres con educación primaria	0.082
	(0.037)**
Padres con educación secundaria	0.162
	(0.038)***
Padres con educación superior	0.374
	(0.046)***
Ingreso alto	0.148
	(0.033)***
Ingreso medio	0.243
	(0.038)***
Ingreso alto	0.065
	(0.025)***
Puntaje ICFES del colegio	0.825
	(0.050)***
Número de alumnos	-0.001
	(0.000)**
Intercepto	2.420
	(0.295)***
R^2	0.141
Observaciones	6,255

Calculos realizados utilizando datos de Bonilla, Bottan y Ham (2017).
*** Significativo al 1 por ciento, ** 5 por ciento y * 10 por ciento.

Interpretación: *Ceteris Paribus*

- Cuándo ven una regresión multivariada, los coeficientes se interpretan de una manera particular.

Interpretación: *Ceteris Paribus*

- Cuándo ven una regresión multivariada, los coeficientes se interpretan de una manera particular.

- Cada coeficiente dice cuál es el efecto de una variable sobre el resultado **manteniendo el resto de las variables constantes**.

Fortalezas

1. Permite controlar por otros factores relevantes (y disponibles), reduciendo la posibilidad de sesgo por variables omitidas.

Fortalezas

1. Permite controlar por otros factores relevantes (y disponibles), reduciendo la posibilidad de sesgo por variables omitidas.

2. Puede aproximarse al verdadero efecto causal si argumentamos exitosamente que la relación estimada captura la verdad.

Fortalezas

1. Permite controlar por otros factores relevantes (y disponibles), reduciendo la posibilidad de sesgo por variables omitidas.

2. Puede aproximarse al verdadero efecto causal si argumentamos exitosamente que la relación estimada captura la verdad.

3. El método funciona para lo que fue hecho (minimizar errores), fijénse que los problemas surgen por especificar mal el modelo.

Debilidades

1. Dadas las limitaciones de datos, no podemos controlar por todos los factores relevantes aunque quisieramos.

Debilidades

1. Dadas las limitaciones de datos, no podemos controlar por todos los factores relevantes aunque quisieramos.

2. Para convencer a las personas que realmente estimamos una relación causal, toca ser muuuuuuuuy convincentes.

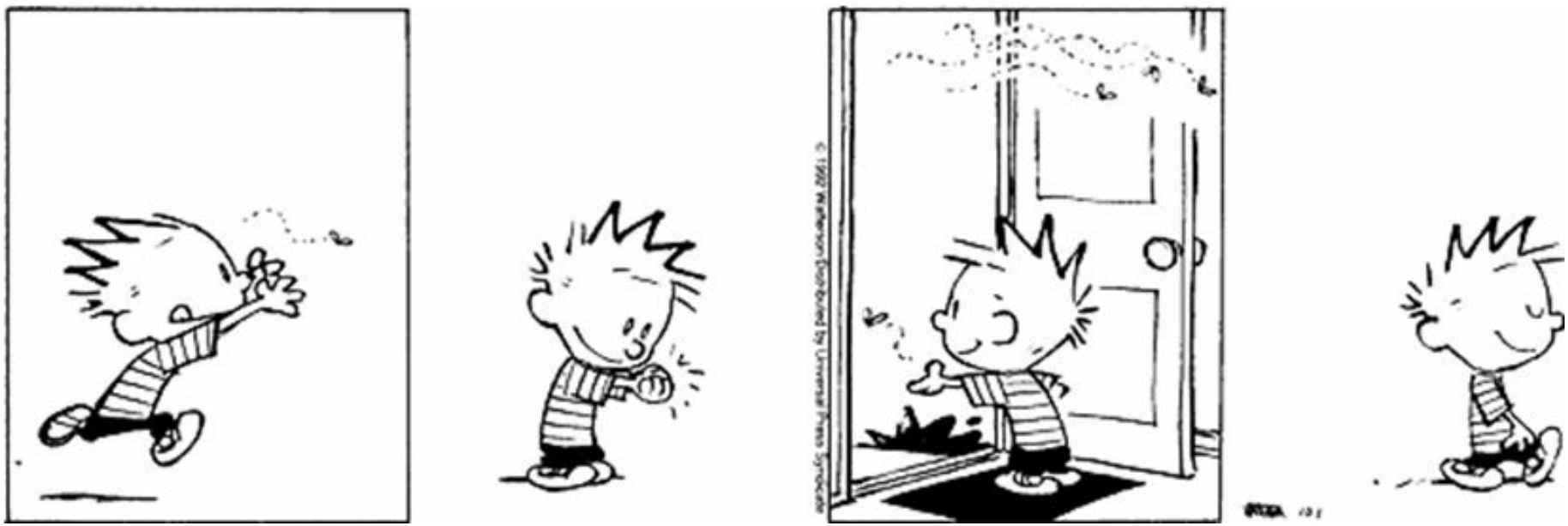
Debilidades

1. Dadas las limitaciones de datos, no podemos controlar por todos los factores relevantes aunque quisieramos.

2. Para convencer a las personas que realmente estimamos una relación causal, toca ser muuuuuuuuy convincentes.

3. Incluir más variables no necesariamente es mejor (esto lo vamos a discutir más adelante).

Regression:
"when you fix one bug, you introduce several newer bugs."



Un mapa: ¿Cuándo hay problemas de endogeneidad?

1. **Variables omitidas**: ¿Hay otras cosas además de la edad que afectan el puntaje del ICFES y no incluimos en nuestra regresión?

Un mapa: ¿Cuándo hay problemas de endogeneidad?

1. **Variables omitidas**: ¿Hay otras cosas además de la edad que afectan el puntaje del ICFES y no incluimos en nuestra regresión?

2. **Error de medición en** x: ¿Qué pasa si la variable edad está medida con error sistemático?

Un mapa: ¿Cuándo hay problemas de endogeneidad?

1. **Variables omitidas**: ¿Hay otras cosas además de la edad que afectan el puntaje del ICFES y no incluimos en nuestra regresión?

2. **Error de medición en** x: ¿Qué pasa si la variable edad está medida con error sistemático?

3. **Simultaneidad**: ¿En qué dirección realmente opera la relación causal? ¿Edad $\leftrightarrows$ ICFES?

El mensaje del día

- La regresión univariada (una sola x) suele sufrir de problemas de endogeneidad debido al **sesgo de variables omitidas**.
- Es casi imposible cuantificar el tamaño del sesgo, pero sí es factible determinar su signo (positivo o negativo).
- Esto es un problema porque podemos llegar a conclusiones erróneas a la hora de tomar decisiones de política pública.
- Una potencial solución es incluir más variables, pero esto tiene sus fortalezas y debilidades. Conózcanlas bien.

En el próximo capítulo...

- (T)errores comúnes en el análisis de regresión.

- Lecturas:
  - Capítulo 12 de Wheelan: "*Naked Statistics*".