Universidad de los Andes
Escuela de Gobierno Alberto Lleras Camargo

Métodos Cuantitativos para Asuntos Públicos I

EGOB 2101

Andrés Ham
a.ham@uniandes.edu.co

2019-1

15 de febrero del 2019

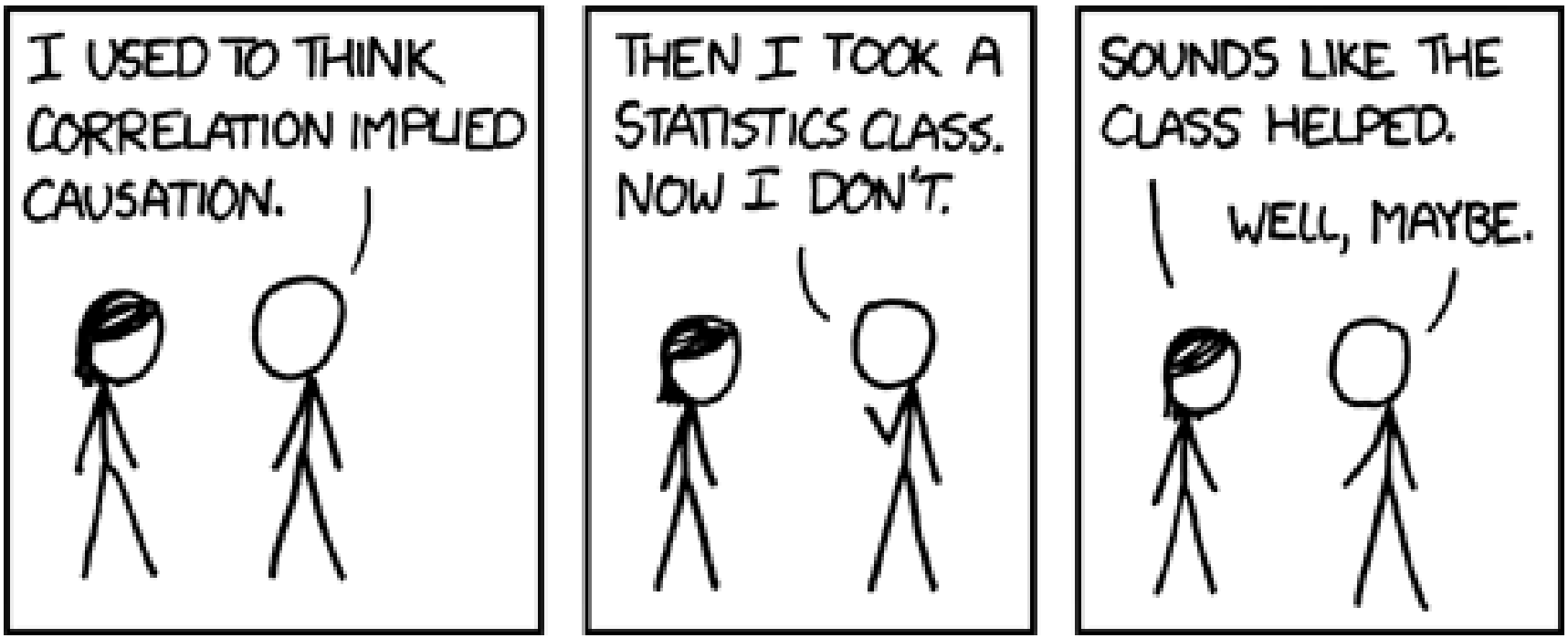
Agenda de hoy

1. El Nokia 3310 de la estadística: Regresión univariada

2. El mensaje del día

Para los que no han visto un Nokia 3310

Previously...


I USED TO THINK CORRELATION IMPLIED CAUSATION.
THEN I TOOK A STATISTICS CLASS. NOW I DON'T.
SOUNDS LIKE THE CLASS HELPED.
WELL, MAYBE.

Previously...

- En una base de datos, nos interesa saber qué hay cosas que se mueven juntas o están correlacionadas.
- Pero, dichas asociaciones son informativas hasta cierto punto, ya que correlación $\neq$ causalidad.
- ¿Entonces cómo podemos saber si realmente dos variables están relacionadas de manera causal?

El problema

- El desafío del análisis cuantitativo es separar la contribución de cada factor ($x_1, \ldots, x_n$) sobre un resultado de interés (y).

El problema

- El desafío del análisis cuantitativo es separar la contribución de cada factor ($x_1, \ldots, x_n$) sobre un resultado de interés (y).
- Hoy nos vamos a concentrar en el caso **univariado**, es decir, el efecto de una variable x sobre el resultado y.

El problema

- El desafío del análisis cuantitativo es separar la contribución de cada factor ($x_1, \ldots, x_n$) sobre un resultado de interés (y).
- Hoy nos vamos a concentrar en el caso **univariado**, es decir, el efecto de una variable x sobre el resultado y.
- Vamos a ver la mecánica del procedimiento utilizando nuestros datos del SABER 11 para colegios distritales en Bogotá.

Correlaciones en los datos del ICFES

- Ya habíamos calculado algunas asociaciones entre variables sociodemográficas y el puntaje en el ICFES.

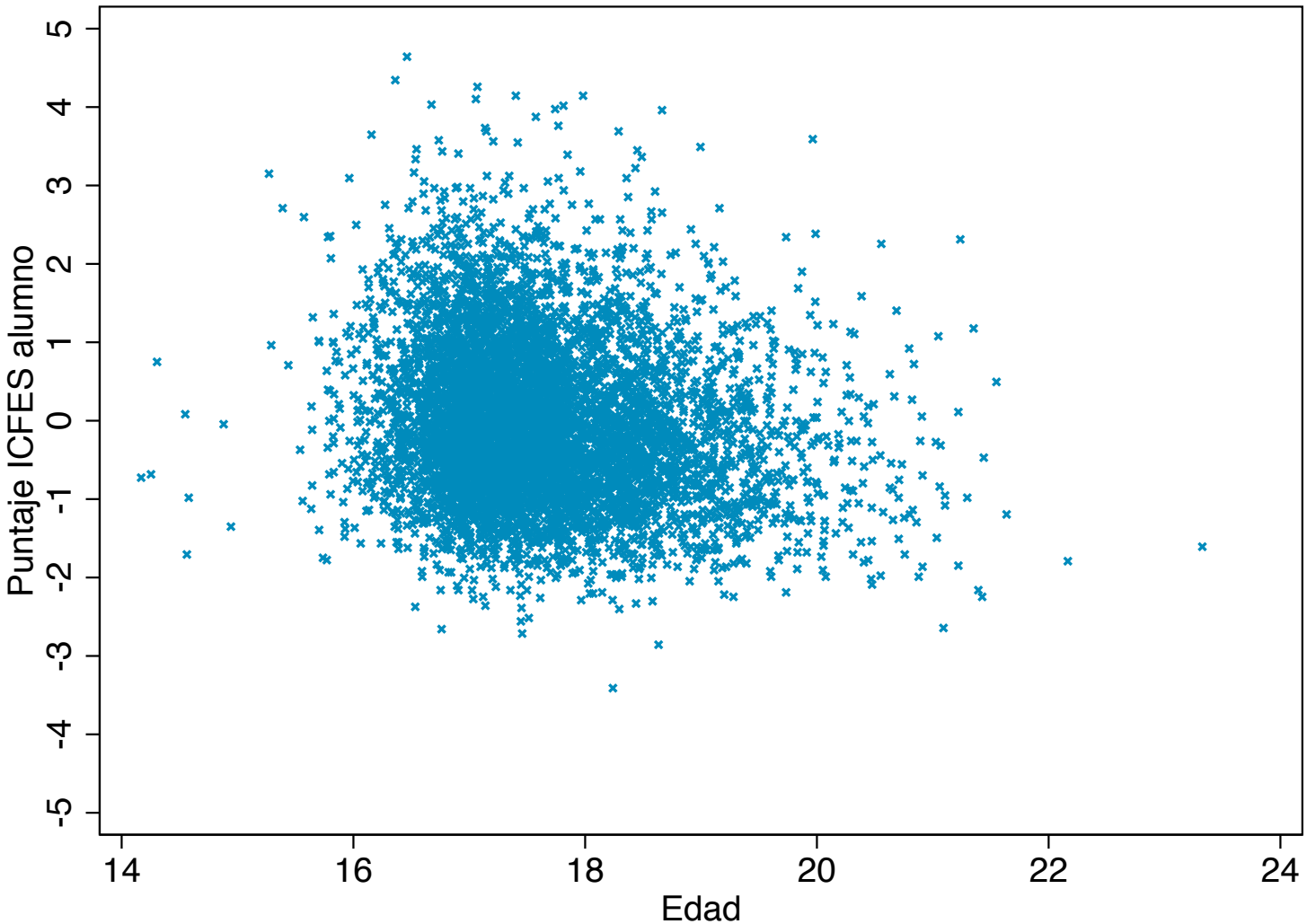
	Coeficiente de correlación
Hombres	0.172
Edad	-0.161
Educación de los padres	0.207
Ingreso familiar	0.162
Jornada matutina	0.047
Puntaje ICFES del colegio	0.259
Número de alumnos en el colegio	0.021

Calculos realizados utilizando datos de Bonilla, Bottan y Ham (2017).

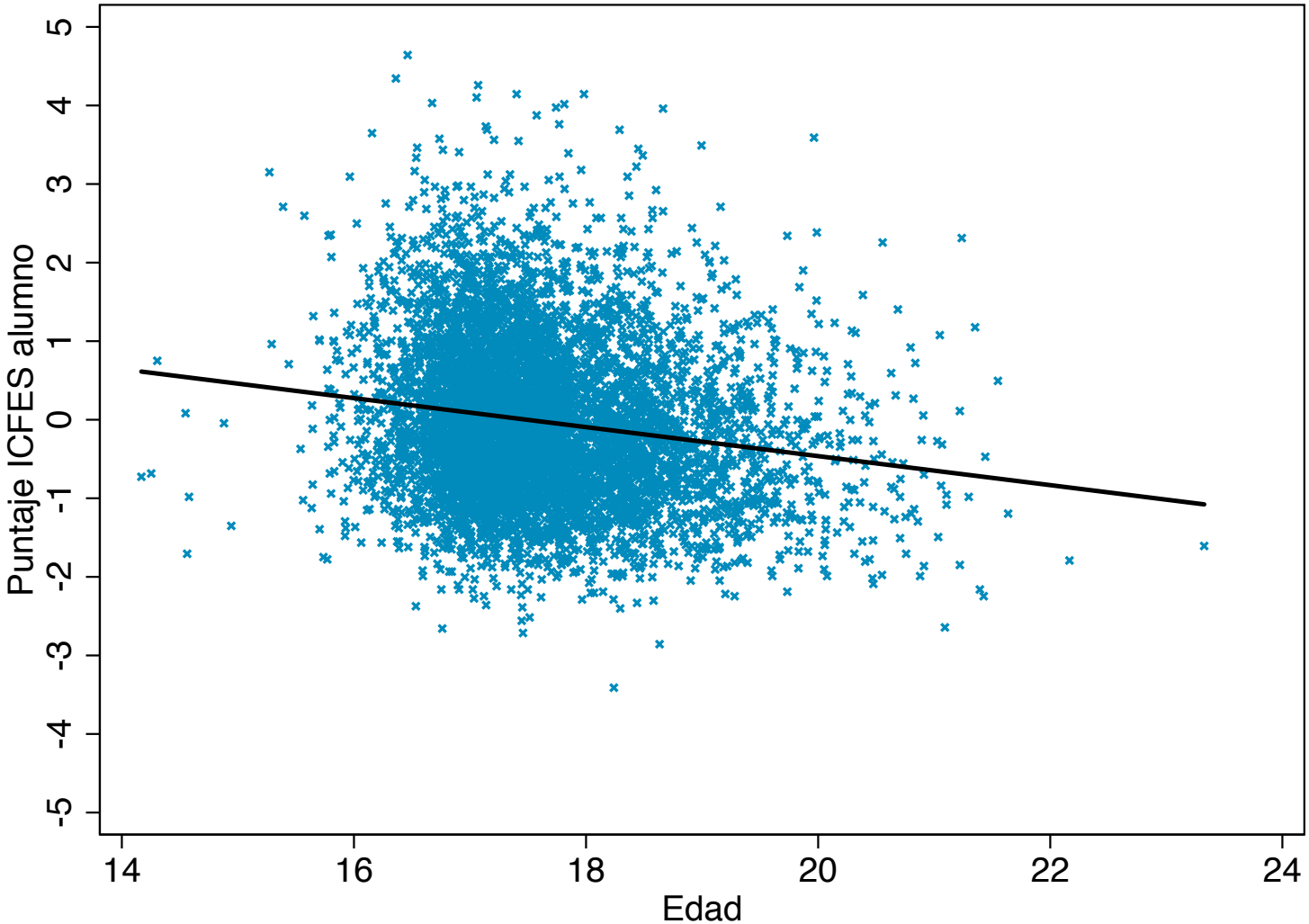
Una pregunta interesante (entre muchas)

¿Cuánto importa la edad del alumno(a) para determinar su rendimiento académico?

La nube de puntos



Un regresión ajusta una línea a los datos



¿Qué hace la regresión?

- Cuantifica la relación entre una variable explicativa (x_1) y una variable dependiente (y), controlando por otros factores ($x_2, \ldots, x_n$).

¿Qué hace la regresión?

- Cuantifica la relación entre una variable explicativa (x_1) y una variable dependiente (y), controlando por otros factores ($x_2, \ldots, x_n$).
- En nuestro ejemplo, proveería **la mejor aproximación lineal** entre puntaje ICFES y edad del alumno(a).

¿Cómo hace eso?

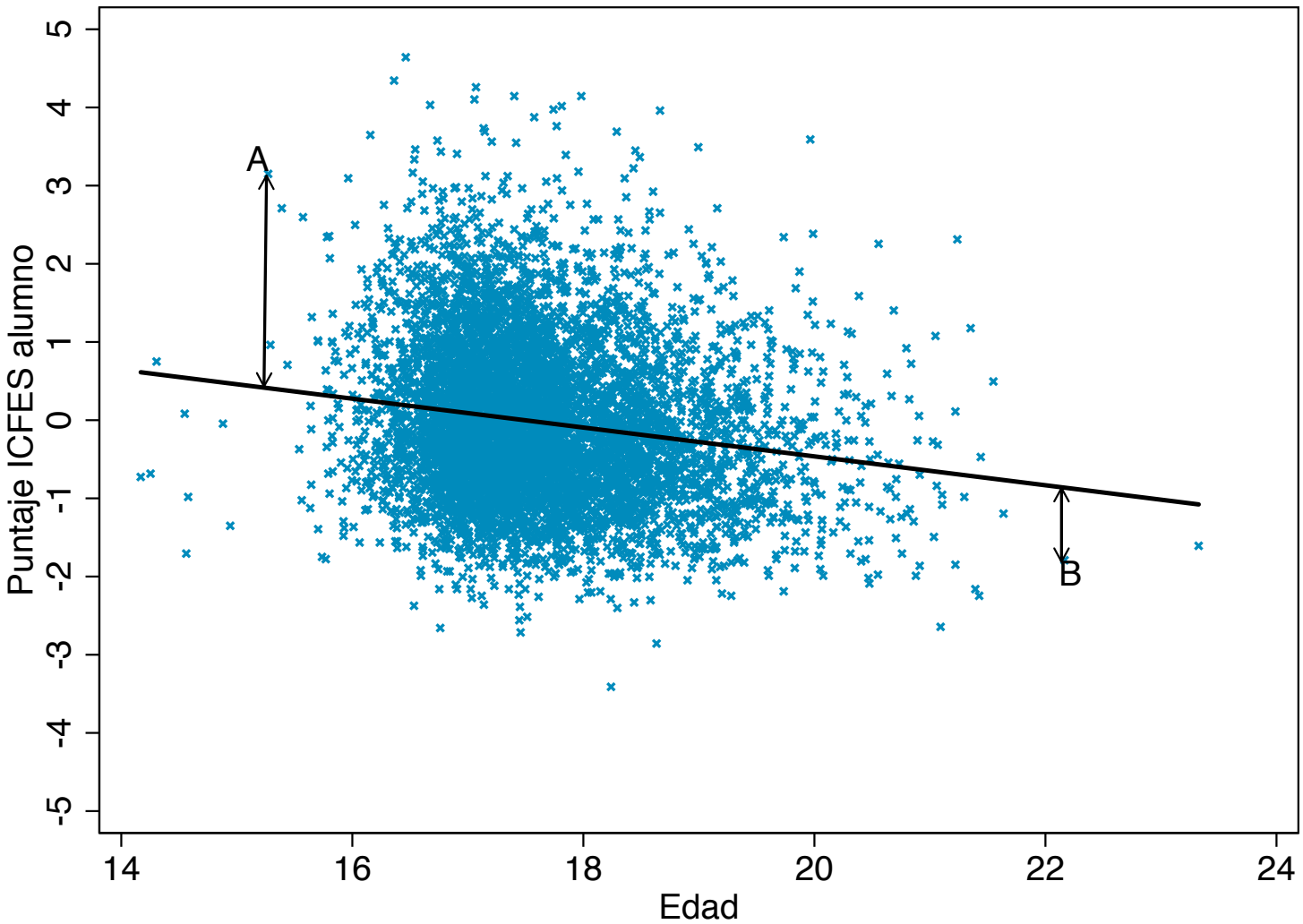
- Encuentra la línea que minimiza los errores cuadrados.

¿Cómo hace eso?

- Encuentra la línea que minimiza los errores cuadrados. ¿El qué?

¿Cómo hace eso?

- Encuentra la línea que minimiza los errores cuadrados. ¿El qué?



¿Cómo hace eso?

- Una regresión con una sola variable explicativa se puede escribir de la siguiente manera:

¿Cómo hace eso?

- Una regresión con una sola variable explicativa se puede escribir de la siguiente manera:

$$y_i = \alpha + \beta x_i + u_i$$

¿Cómo hace eso?

- Una regresión con una sola variable explicativa se puede escribir de la siguiente manera:

$$y_i = \alpha + \beta x_i + u_i$$

- Donde:
  - y es la variable dependiente o explicada: puntaje ICFES
  - x la variable independiente o explicativa: edad.
  - α es el intercepto de la línea de regresión.
  - β es la pendiente de la línea de regresión, que nos dice cómo cambia y si aumenta x en una unidad.
  - u incluye todos los otros factores que afectan a y.

¿Cómo hace eso? Un paréntesis matemático

- Olvidemos las i y pongamos la constante dentro del vector β.

$$y = \beta x + u$$

¿Cómo hace eso? Un paréntesis matemático

- Olvidemos las i y pongamos la constante dentro del vector β.

$$y = \beta x + u$$

- Una regresión minimiza los errores cuadrados:

$$\min_{\beta} u^2 = (y - \beta x)(y - \beta x)$$

¿Cómo hace eso? Un paréntesis matemático

- Olvidemos las i y pongamos la constante dentro del vector β.

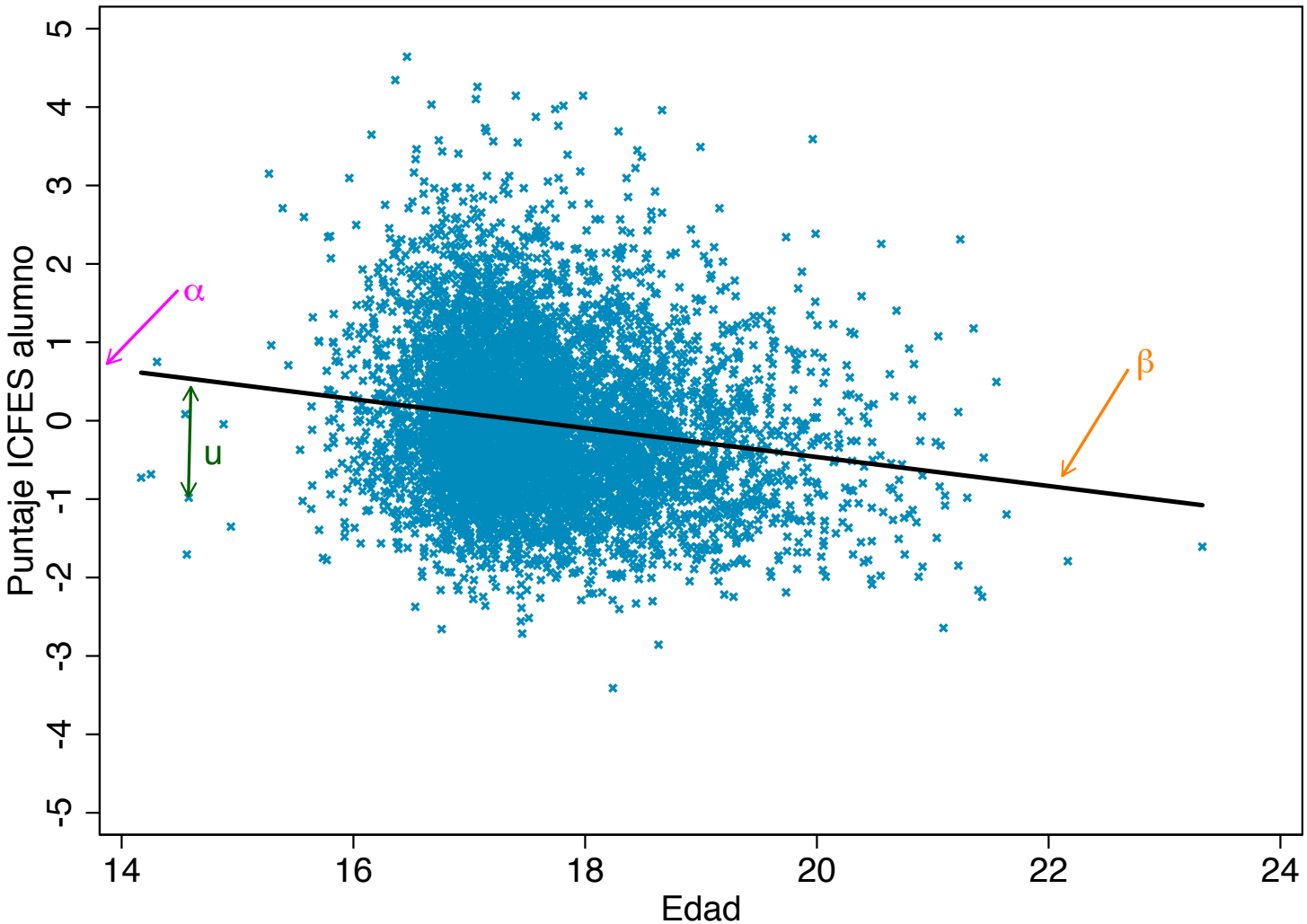
$$y = \beta x + u$$

- Una regresión minimiza los errores cuadrados:

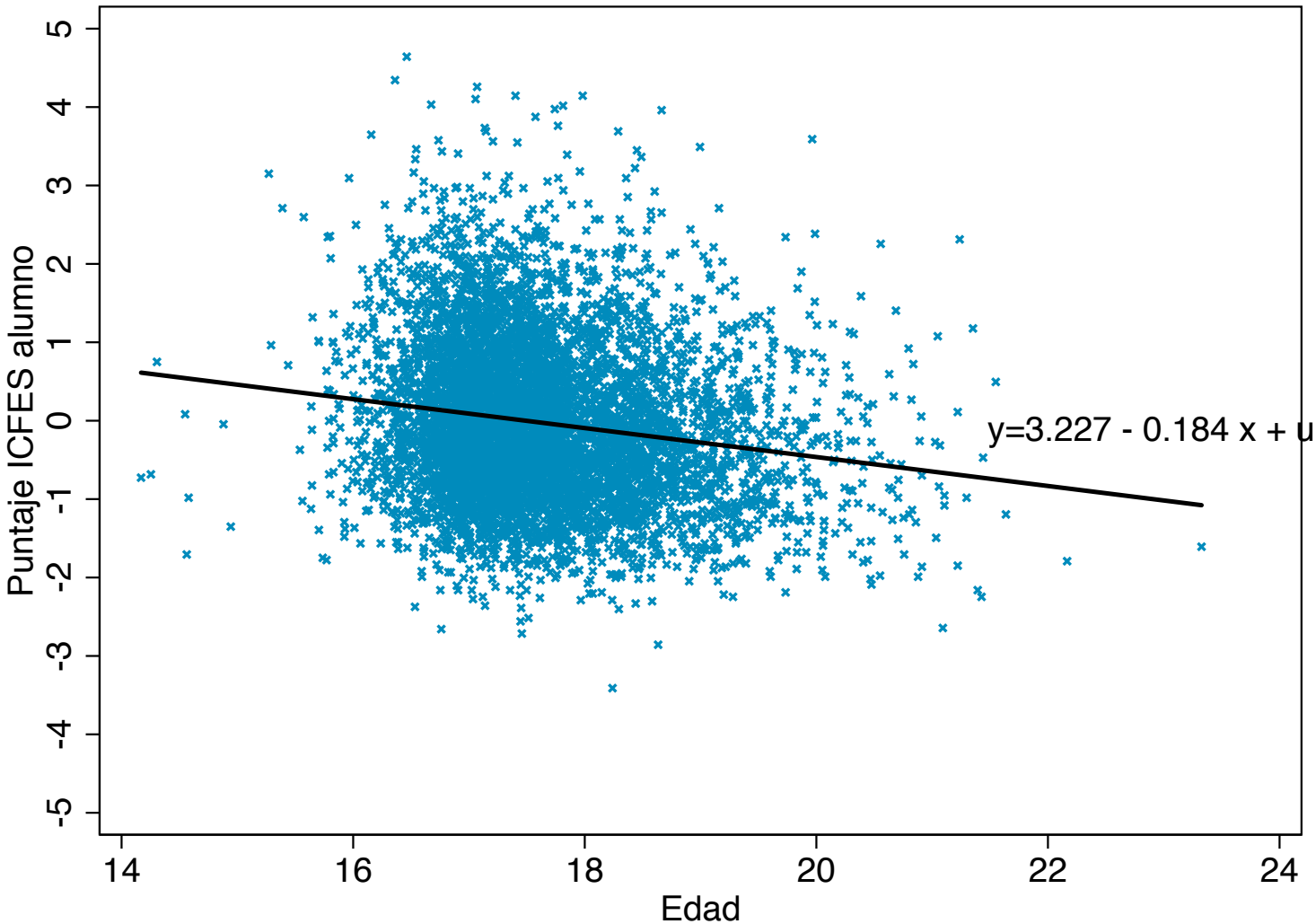
$$\min_{\beta} u^2 = (y - \beta x)(y - \beta x)$$

- Si hacen el álgebra y derivan con respecto a β les debe dar que $\beta^* = \frac{xy}{x^2}$. Este es el estimador de **mínimos cuadrados ordinarios**. Se los dejo de tarea.

¿Cómo hace eso?



El resultado en un gráfico



El resultado en un 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.

¿Cómo desciframos esto?

- Hay tres cosas importantes que podemos sacar de un cuadro que presenta estimaciones de una regresión.

¿Cómo desciframos esto?

- Hay tres cosas importantes que podemos sacar de un cuadro que presenta estimaciones de una regresión.

    1. **Signo**: ¿es positivo o negativo el coeficiente de la pendiente?

¿Cómo desciframos esto?

- Hay tres cosas importantes que podemos sacar de un cuadro que presenta estimaciones de una regresión.

  1. **Signo**: ¿es positivo o negativo el coeficiente de la pendiente?
  2. **Tamaño**: ¿qué tan grande es ese coeficiente?

¿Cómo desciframos esto?

- Hay tres cosas importantes que podemos sacar de un cuadro que presenta estimaciones de una regresión.

  1. **Signo**: ¿es positivo o negativo el coeficiente de la pendiente?
  2. **Tamaño**: ¿qué tan grande es ese coeficiente?
  3. **Significancia**: ¿qué tan confiados estamos que ese coeficiente es distinto de cero?

Signo

- ¿Cómo le explicarían este valor estimado a alguien?

	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.

Tamaño o Significancia económica

- ¿Eso les parece grande o chico?

	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.

- Necesitamos contexto. El puntaje promedio para toda la muestra es -0.027 con desvío estándar 1.00.

Significancia

- ¿Estamos seguro que la edad afecta al ICFES y nuestro resultado no es pura casualidad?

Significancia

- ¿Estamos seguro que la edad afecta al ICFES y nuestro resultado no es pura casualidad?
- ¡Para contestar esta pregunta toca hacer una **prueba de hipótesis**!

Significancia

- ¿Estamos seguro que la edad afecta al ICFES y nuestro resultado no es pura casualidad?
- ¡Para contestar esta pregunta toca hacer una **prueba de hipótesis**!
- Si recordamos bien, para eso necesitabamos estimar un promedio y su error estándar. Despues definimos la prueba.

Significancia

- ¿Estamos seguro que la edad afecta al ICFES y nuestro resultado no es pura casualidad?
- ¡Para contestar esta pregunta toca hacer una **prueba de hipótesis**!
- Si recordamos bien, para eso necesitabamos estimar un promedio y su error estándar. Despues definimos la prueba.
- En este caso, ¿cuál es nuestra hipótesis nula?

Significancia

- ¿Estamos seguro que la edad afecta al ICFES y nuestro resultado no es pura casualidad?
- ¡Para contestar esta pregunta toca hacer una **prueba de hipótesis**!
- Si recordamos bien, para eso necesitabamos estimar un promedio y su error estándar. Despues definimos la prueba.
- En este caso, ¿cuál es nuestra hipótesis nula? Y la alternativa?

Significancia estadística

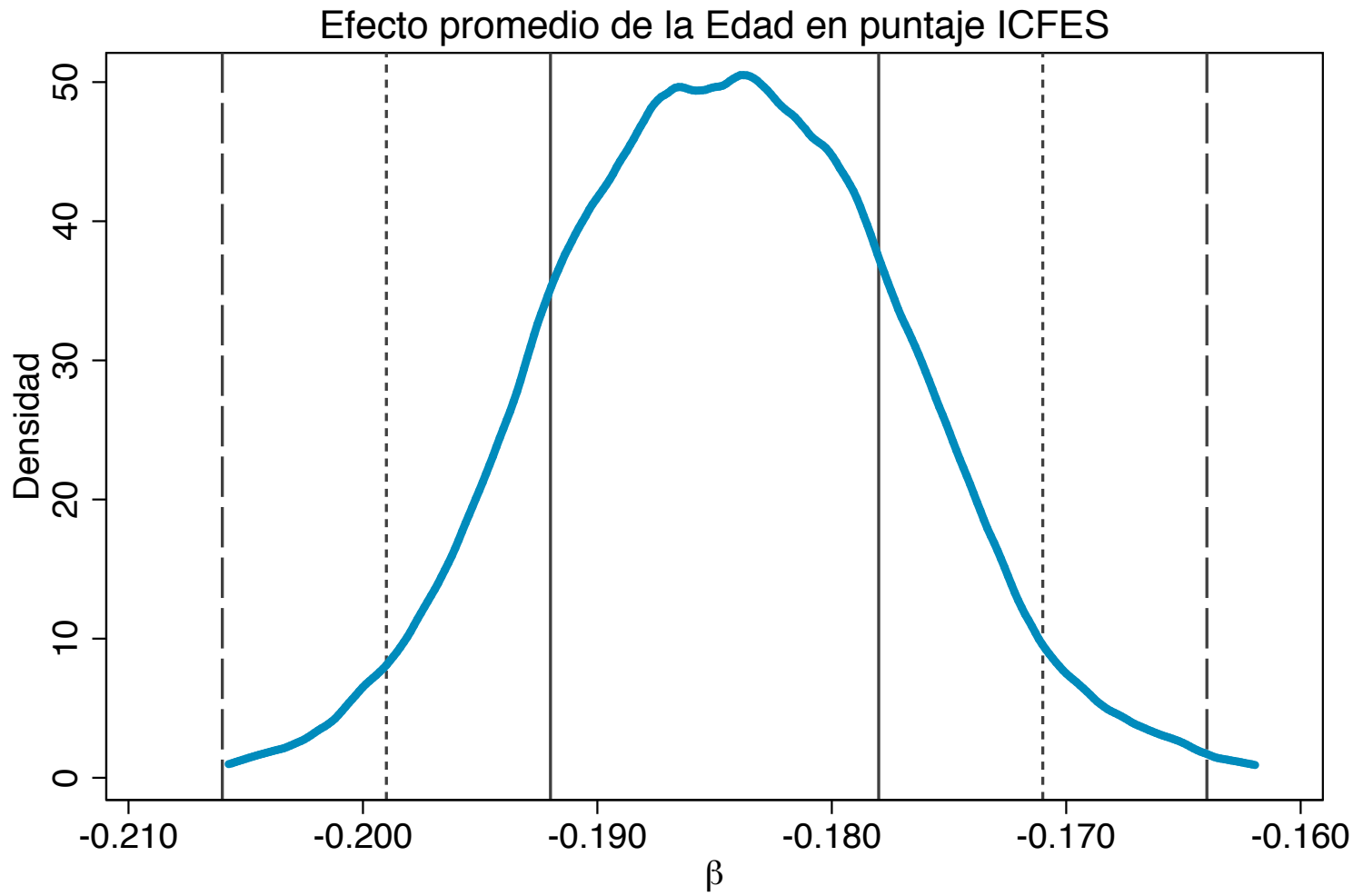
	Puntaje ICFES
Edad ($\hat{\beta}$)	-0.184
	(0.014)***
Intercepto ($\hat{\alpha}$)	3.227
	(0.255)***
R^2	0.027
Observaciones	6,316

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

El teorema central del límite en funcionamiento

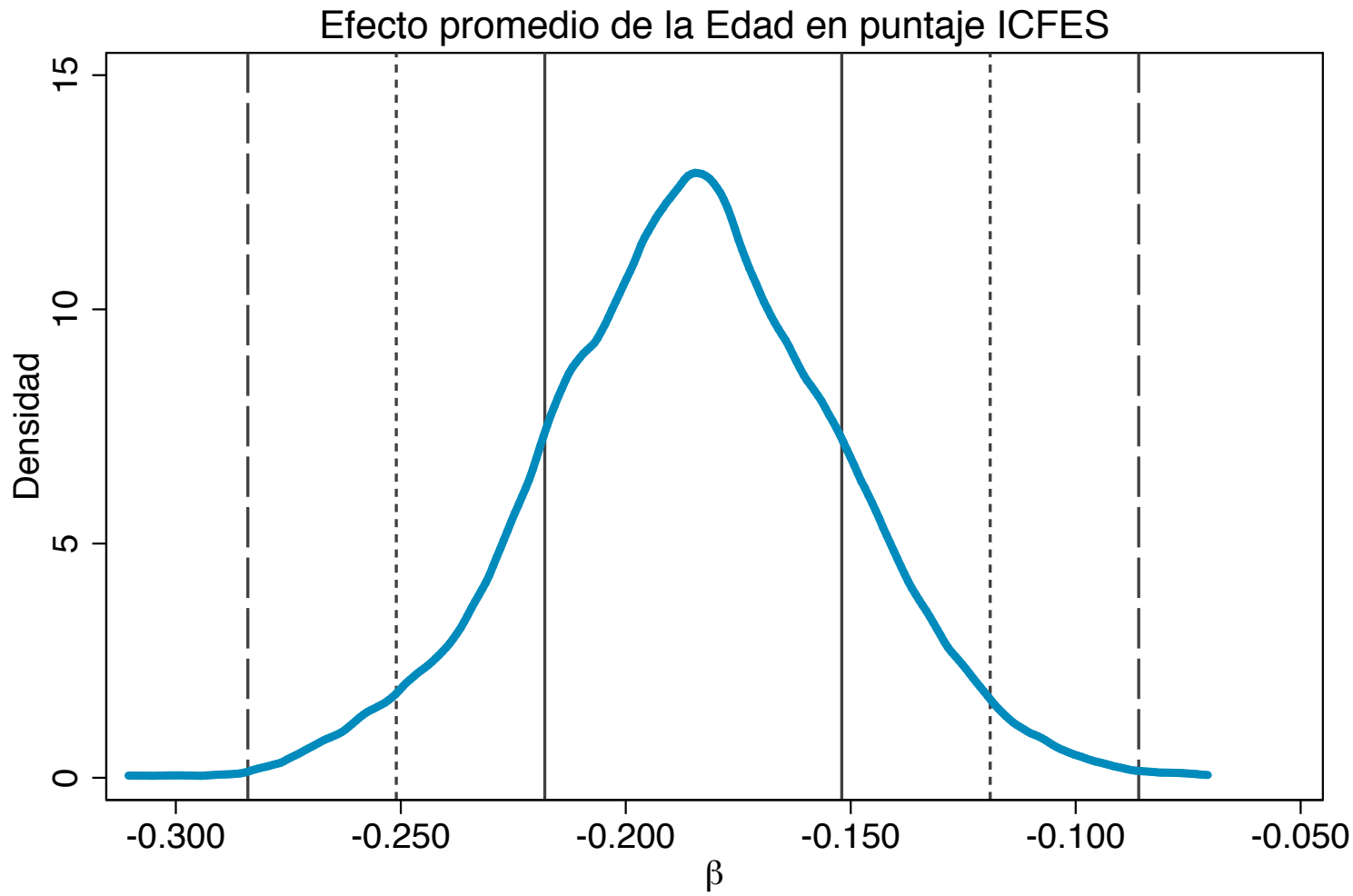
- El TCL nos decía que un promedio tiende a estar entre $\pm$ 2 errores estándar 95 de cada 100 veces.
- Entonces, si nuestra hipótesis nula es que la edad no afecta el puntaje, el cero no debe estar en ese rango.
- Hagamos varios ejercicios tomando muestras de tamaños diferentes.

5000 observaciones



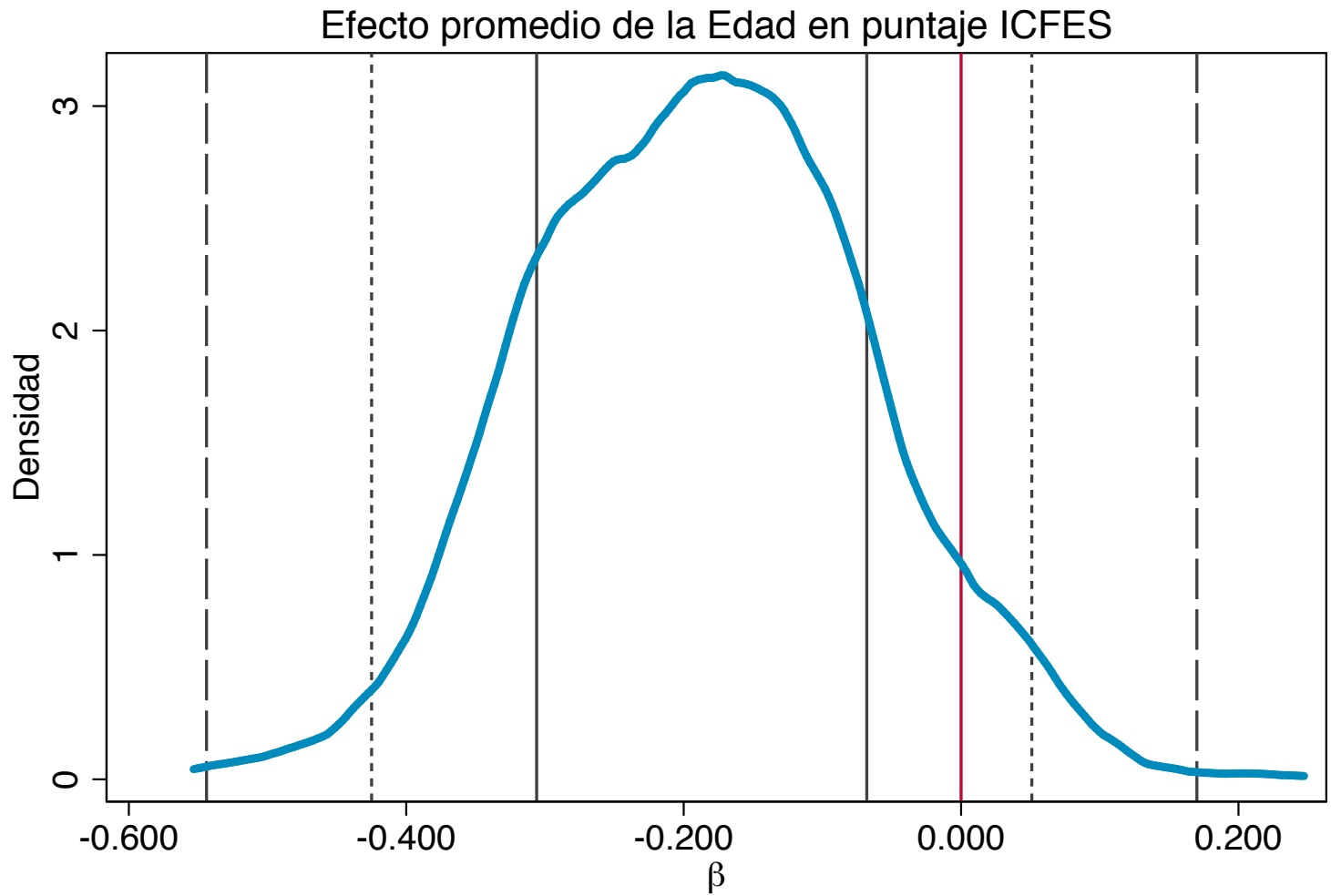
Media=-0.185; Error Estándar=0.007

1000 observaciones



Media=-0.185; Error Estándar=0.033

100 observaciones



Media=-0.187; Error Estándar=0.119

El ajuste de la línea de regresión

- ¿Cuánto explica la edad el puntaje ICFES?

	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.

¿Cuál les parece que es la fortaleza de la regresión univariada? ¿Y las debilidades?

Un vistazo a lo que se viene

	Univariada	Multivariada
Edad	-0.184	-0.165
	(0.014)***	(0.014)***
Hombres		0.350
		(0.024)***
Padres con educación primaria		0.077
		(0.037)**
Padres con educación secundaria		0.162
		(0.037)***
Padres con educación superior		0.373
		(0.046)***
Ingreso medio		0.150
		(0.033)***
Ingreso alto		0.248
		(0.038)***
Jornada matutina		0.068
		(0.025)***
Puntaje ICFES del colegio		0.828
		(0.049)***
Número de alumnos		-0.001
		(0.000)**
Intercepto	3.227	2.307
	(0.255)***	(0.254)***
R^2	0.027	0.142
Observaciones	6,316	6,302

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

El mensaje del día

- Las asociaciones entre variables pueden ser *estimadas* mediante el análisis de regresión.
- El método provee la mejor aproximación lineal de dicha relación, pero su interpretación depende de cosas ajenas a la estadística.
- La información relevante de una regresión tiene que ver con signo, tamaño o significatividad económica y significatividad estadística.
- Permite estimar relaciones causales, pero solamente bajo ciertas condiciones, que iremos viendo más adelante.

En el próximo capítulo...

- ¿Y si controlamos por más factores? Regresión multivariada.

- Lecturas:
  - Capítulo 11 de Wheelan: "*Naked Statistics*", páginas 198-207.
  - Capítulo 2 de Angrist y Pischke: "*Mastering Metrics*", páginas 69-81.