ESTADISTICA.AR

ANOVA y regresión

Estadística · Semana 12 · Unidad 7: ANOVA y regresión

PREGUNTA PARA LA CLASE

Se quiere comparar el empleo no registrado de las cinco regiones del interior del país. ¿Por qué no hacer una prueba t para cada par de regiones?

Hoy

1

Más de dos medias

La idea del análisis de la varianza

2

La tabla ANOVA

Sumas de cuadrados y la prueba F

3

Supuestos y post hoc

Qué revisar y cuáles grupos difieren

4

La regresión lineal simple

El modelo y sus supuestos

5

Inferencia y residuos

La prueba para la pendiente y el diagnóstico

1

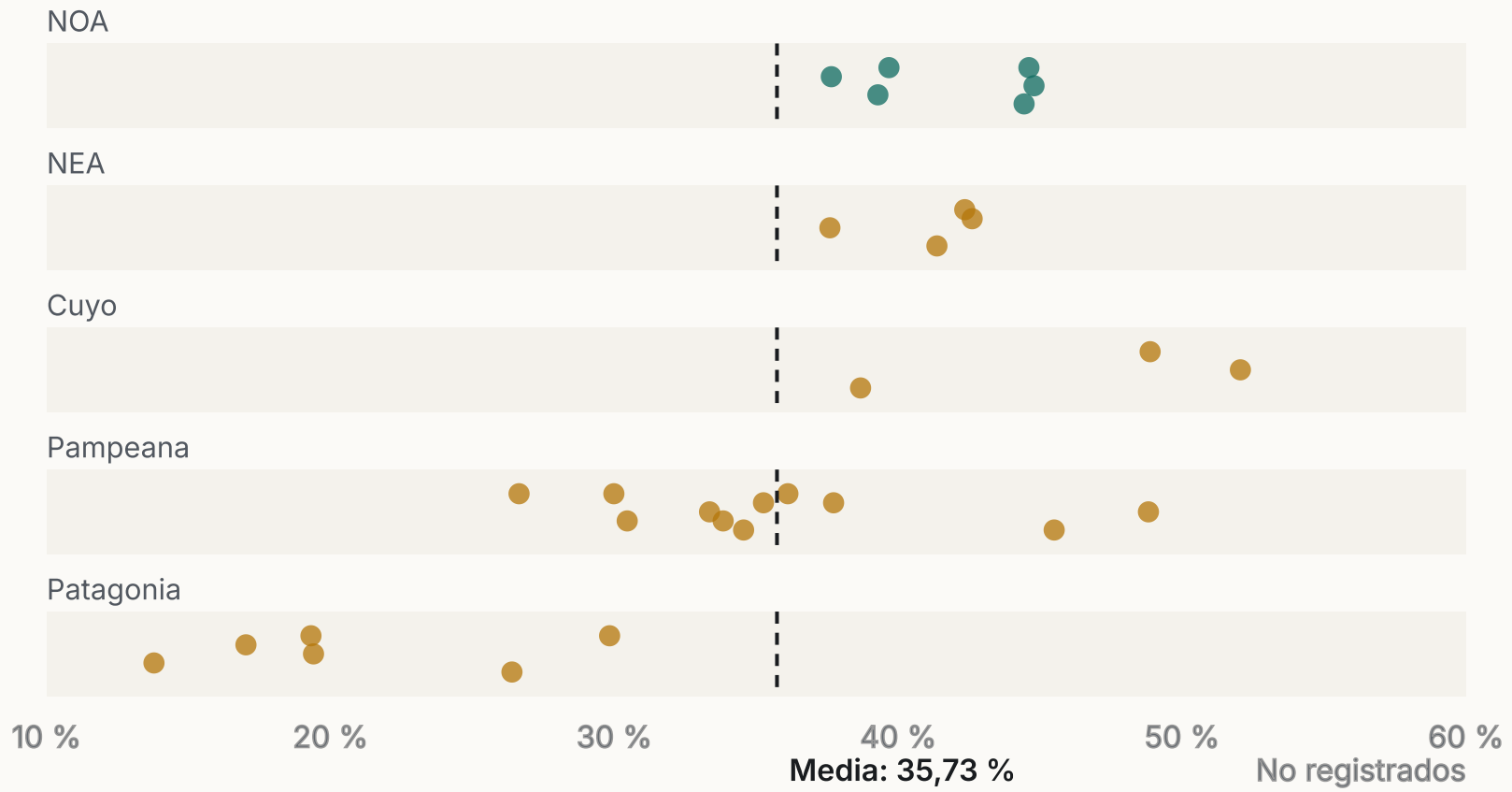
Más de dos medias

La idea del análisis de la varianza

SEMANA 12 · MÁS DE DOS MEDIAS

El empleo no registrado por región

Cada punto es un aglomerado: el porcentaje de sus asalariados sin descuento jubilatorio.



Fuente: INDEC, Encuesta Permanente de Hogares, primer trimestre de 2026. Los 30 aglomerados del interior; base en estadistica.ar/datos/indicadores-aglomerados.

Los resúmenes de cada región

Región	Aglomerados	Media (%)	Desvío (%)
NOA	6	41,733	3,219
NEA	4	40,973	2,317
Cuyo	3	46,530	6,990
Pampeana	11	35,653	6,518
Patagonia	6	20,955	6,007
Total	30	35,726	

H_0: las cinco medias son iguales. H_1: al menos una es distinta.

La idea: comparar dos variabilidades

Entre los grupos

Cuánto se alejan las medias de cada región de la media general.

Si H_0 es cierta, solo por azar.

Dentro de los grupos

Cuánto varían los aglomerados alrededor de la media de su región.

El ruido, haya o no diferencias.

El cociente

F = variabilidad entre / variabilidad dentro

Cerca de 1 si H_0 es cierta; grande si las medias difieren.

2

La tabla ANOVA

Sumas de cuadrados y la prueba F

Las sumas de cuadrados

Entre

$SCE = \Sigma\, n_i\, (\bar{y}_i - \bar{y})^2 = 1.986{,}0$

Por ejemplo, la Patagonia aporta $6 \cdot (20{,}955 - 35{,}726)^2 = 1.309{,}1$.

Dentro

$SCD = \Sigma\, (n_i - 1)\, s_i^2 = 770{,}9$

El NOA aporta $5 \cdot 3{,}219^2 = 51{,}8$.

Total

$SCT = SCE + SCD = 2.756{,}9$

La variabilidad de los 30 aglomerados alrededor de la media general.

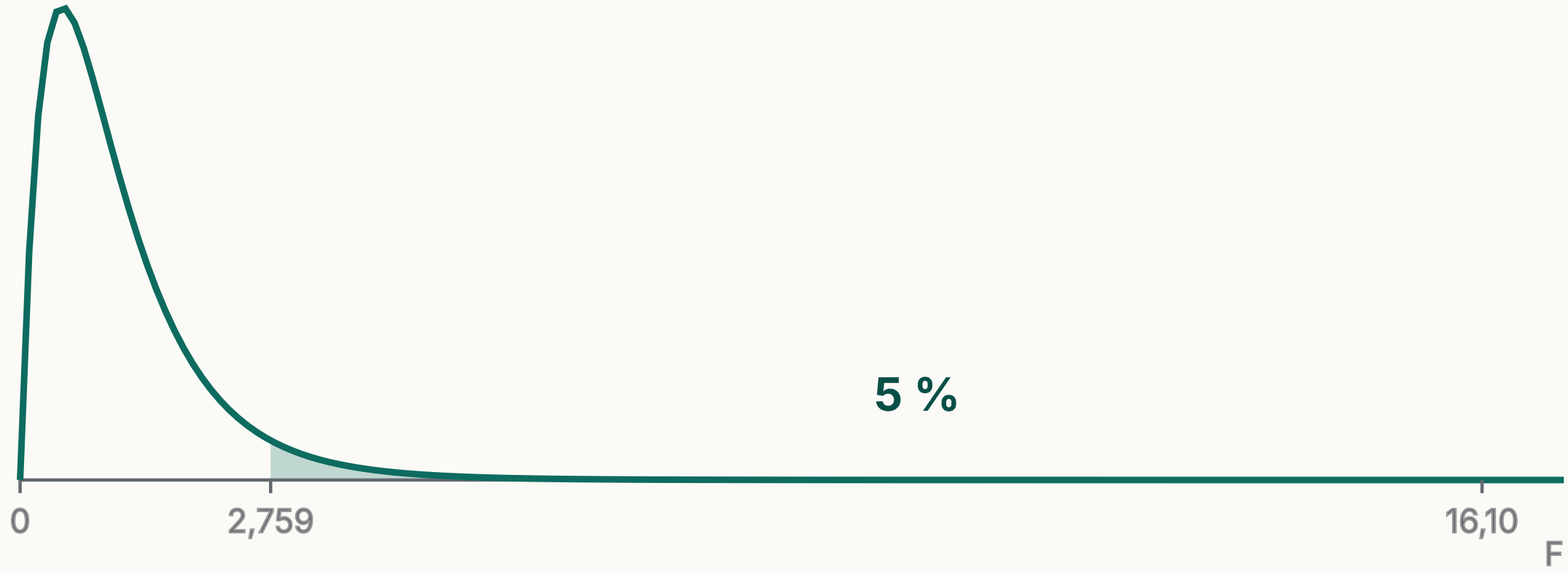
La tabla ANOVA

Fuente	SC	gl	CM = SC / gl	F
Entre regiones	1.986,0	4	496,49	16,10
Dentro de las regiones	770,9	25	30,84	
Total	2.756,9	29		

F = CM entre / CM dentro = 496,49 / 30,84 = 16,10

La distribución F con 4 y 25 grados de libertad

Con α = 0,05, el valor crítico es 2,759. El F de la tabla es 16,10.



p = P(F ≥ 16,10) < 0,001: se rechaza H_0. Al menos una región tiene otra media.

Cuánto explica la región

$\eta^2 = 0{,}72$

SCE / SCT = 1.986,0 / 2.756,9

72 %

de la variabilidad del empleo no registrado entre aglomerados se asocia con la región

Es el tamaño del efecto del ANOVA, como la V de Cramér para la chi-cuadrado.

3

Supuestos y post hoc

Qué revisar y cuáles grupos difieren

Los supuestos del ANOVA

Independencia

Cada dato, independiente de los demás.

Lo garantiza el diseño, no una prueba.

Normalidad

Los datos de cada grupo, aproximadamente normales.

Importa poco con grupos grandes.

Varianzas iguales

Acá el desvío mayor es 3,0 veces el menor: más que el doble.

Levene no rechaza ($p = 0{,}69$), pero con grupos de 3 a 11 tiene poca potencia.

Por las dudas, Welch

El ANOVA de Welch no supone varianzas iguales: $F = 13{,}78$, $p < 0{,}001$.

La conclusión no cambia.

¿Cuáles regiones difieren? Tukey

Regiones	Diferencia de medias	Mínima diferencia significativa (q = 4,15)	¿Difieren?
NOA y NEA	0,76	10,52	No
NOA y Cuyo	4,80	11,52	No
NOA y Pampeana	6,08	8,27	No
NOA y Patagonia	20,78	9,41	Sí
NEA y Cuyo	5,56	12,45	No
NEA y Pampeana	5,32	9,52	No
NEA y Patagonia	20,02	10,52	Sí
Cuyo y Pampeana	10,88	10,61	Sí
Cuyo y Patagonia	25,58	11,52	Sí
Pampeana y Patagonia	14,70	8,27	Sí

4

La regresión lineal simple

El modelo y sus supuestos

PREGUNTA PARA LA CLASE

Un padre mide 10 cm más que el promedio de los padres. ¿Cuánto más que el promedio de los hijos se espera que mida su hijo?

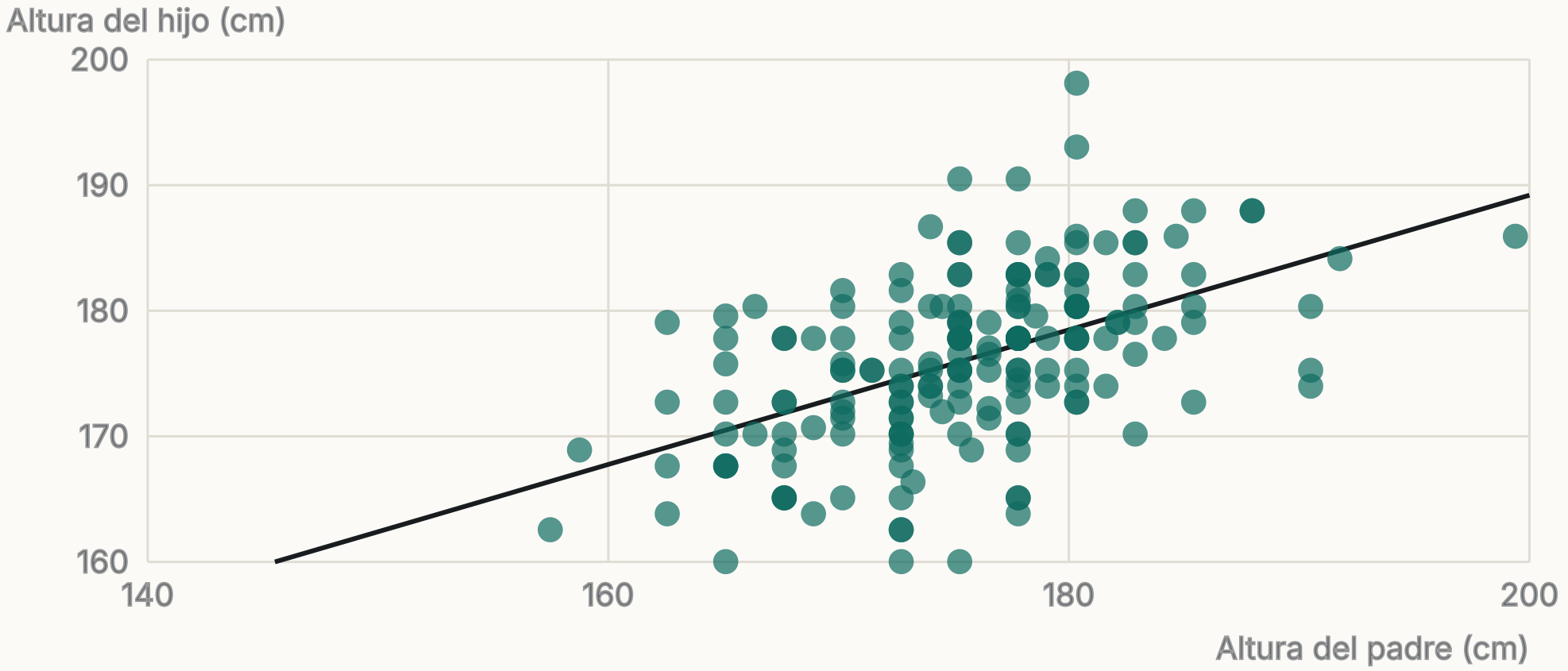
a) Unos 10 cm más

b) Más de 10 cm

c) Menos de 10 cm

Las estaturas de padres e hijos

179 familias: la altura del padre y la de un hijo varón. $\hat{y} = 82{,}0 + 0{,}536 \cdot x$



Fuente: Galton (1886), las 179 familias con algún hijo varón; un hijo por familia, sorteado. Transcripción de Hanley (2004), pasada de pulgadas a centímetros.

La recta, a partir de los resúmenes

Los resúmenes

$\bar{x} = 175{,}5$, $\bar{y} = 176{,}1$, $s_x = 6{,}47$, $s_y = 6{,}79$, $r = 0{,}511$

En centímetros.

La pendiente

$b_1 = r \cdot s_y / s_x = 0{,}511 \cdot 6{,}79 / 6{,}47 =$ **0,536**

Por cada centímetro más del padre, el hijo mide en promedio eso más.

La ordenada

$b_0 = \bar{y} - b_1 \cdot \bar{x} = 176{,}1 - 0{,}536 \cdot 175{,}5 = 82{,}0$

Sin sentido propio: nadie mide 0 cm.

El modelo de regresión lineal simple

$$Y = \beta_0 + \beta_1 \cdot x + \varepsilon$$

Linealidad

La media de Y crece en línea recta con x.

Independencia

Los errores ε de distintas observaciones no se relacionan.

Normalidad

Para cada x, los valores de Y son normales alrededor de la recta.

Varianza constante

El ancho de la nube es el mismo para cualquier x.

5

Inferencia y residuos

La prueba para la pendiente y el diagnóstico

¿La pendiente es distinta de cero?

Hipótesis

H_0: $\beta_1 = 0$. La altura del padre no sirve para predecir la del hijo.

El error estándar

$EE(b_1) = s_e / (sx \cdot \sqrt{(n - 1)}) = 5{,}85 / (6{,}47 \cdot \sqrt{178}) = 0{,}0678$

s_e = 5,85 cm: el desvío de los puntos alrededor de la recta.

El estadístico

$t = b_1 / EE(b_1) = 0{,}536 / 0{,}0678 =$ **7,91**

Con n − 2 = 177 grados de libertad: p < 0,001.

Cuánto, y cuánto explica

0,40 a 0,67

el intervalo del 95 % para β_1: $b_1 \pm 1{,}973 \cdot EE(b_1)$

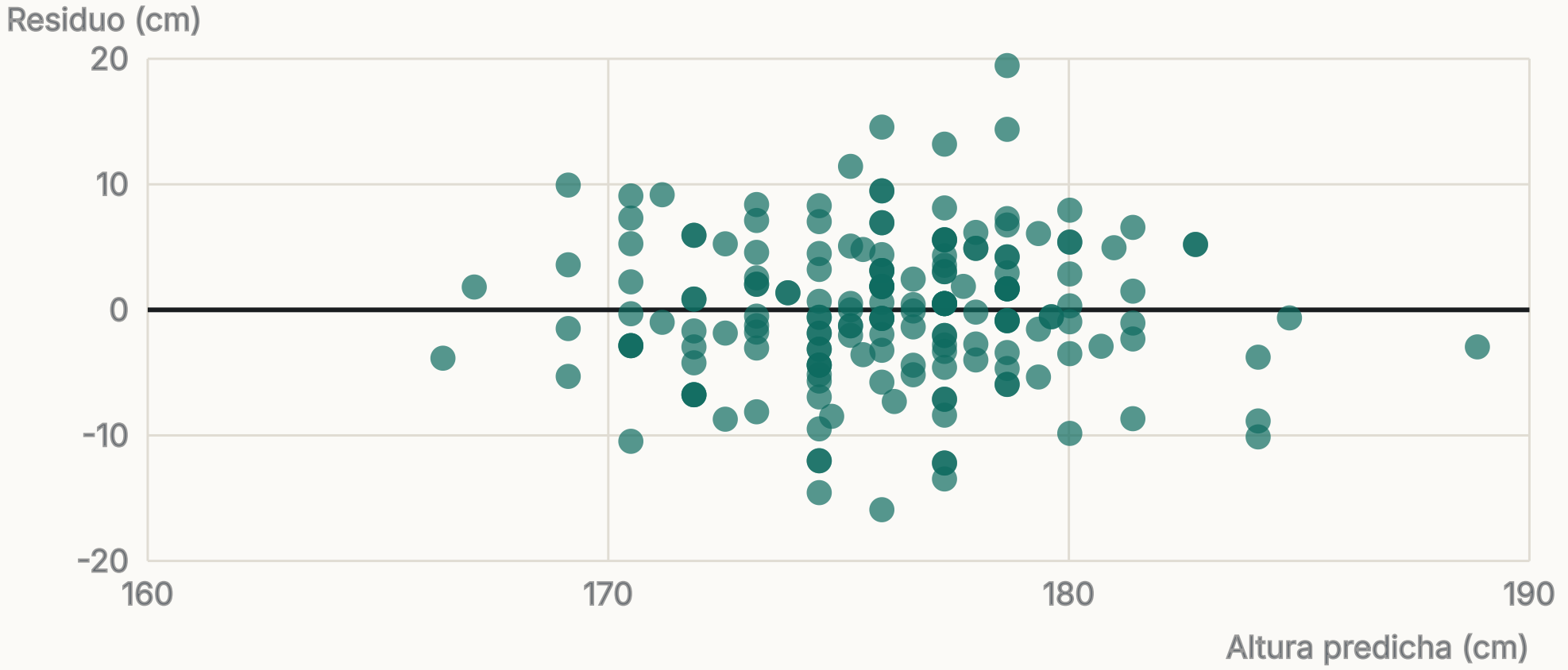
$R^2 = 0{,}26$

$r^2 = 0{,}511^2$: la altura del padre explica el 26 % de la variabilidad de la del hijo

Significativa no es lo mismo que precisa: la recta predice el promedio de los hijos, no la altura de cada uno.

Los residuos: lo que la recta no explica

Cada punto es un hijo: su altura predicha y su residuo, la altura observada menos la predicha.



Fuente: Galton (1886), las 179 familias con algún hijo varón; un hijo por familia, sorteado. Transcripción de Hanley (2004), pasada de pulgadas a centímetros.

Cuatro conjuntos con la misma recta

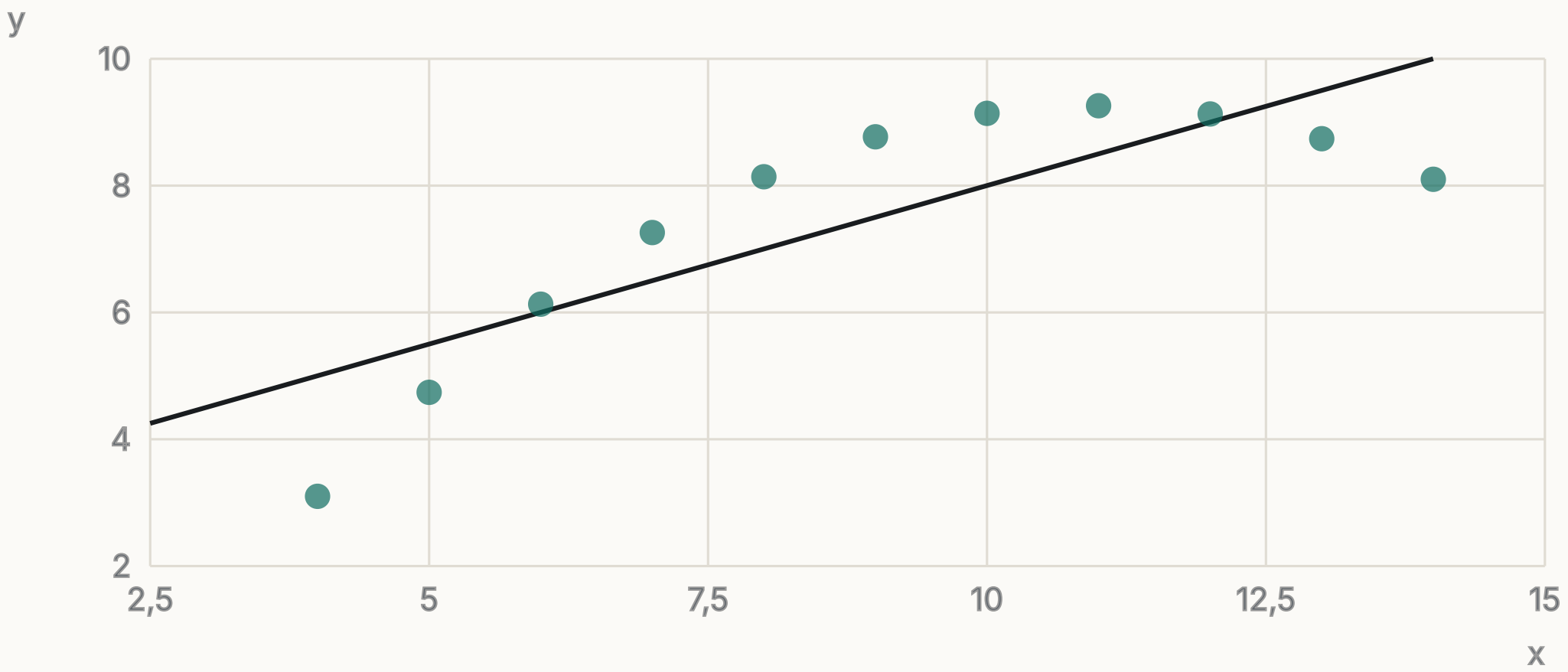
El cuarteto de Anscombe: cuatro conjuntos de 11 puntos, inventados para este ejemplo.

Conjunto	Media de x	Media de y	Desvío de y	r	Recta
I	9,00	7,50	2,03	0,82	$\hat{y} = 3{,}00 + 0{,}500 \cdot x$
II	9,00	7,50	2,03	0,82	$\hat{y} = 3{,}00 + 0{,}500 \cdot x$
III	9,00	7,50	2,03	0,82	$\hat{y} = 3{,}00 + 0{,}500 \cdot x$
IV	9,00	7,50	2,03	0,82	$\hat{y} = 3{,}00 + 0{,}500 \cdot x$

Fuente: Anscombe, F. J. (1973). Graphs in statistical analysis. The American Statistician, 27, 17-21.

El conjunto II: una curva

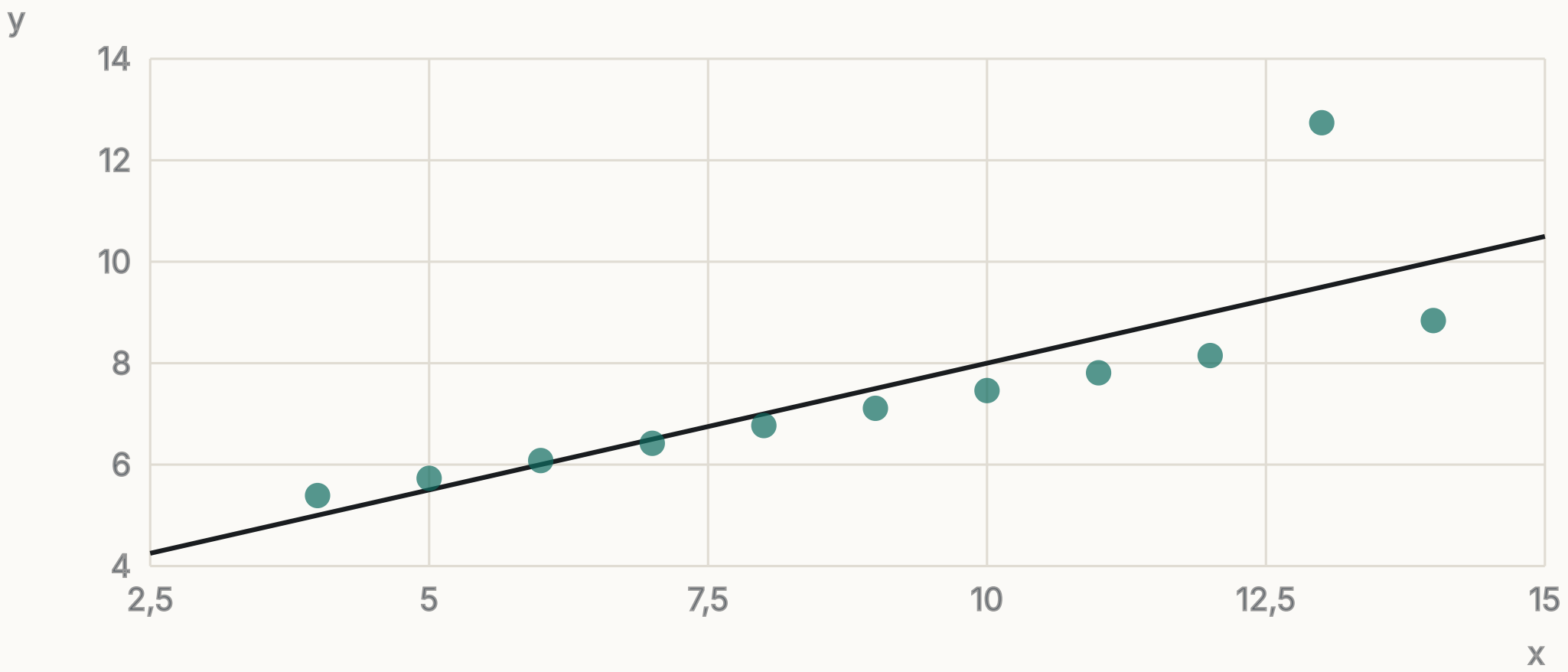
$\hat{y} = 3{,}00 + 0{,}500 \cdot x$



Fuente: Anscombe, F. J. (1973). Graphs in statistical analysis. The American Statistician, 27, 17-21.

El conjunto III: un atípico

$\hat{y} = 3{,}00 + 0{,}500 \cdot x$



Fuente: Anscombe, F. J. (1973). Graphs in statistical analysis. The American Statistician, 27, 17-21.

Qué buscar en los residuos

Una curva

Los residuos suben y bajan con x.

La relación no es lineal.

Un embudo

Los residuos se abren a medida que crece x.

La varianza no es constante: pasa con los ingresos.

Puntos alejados

Un residuo mucho más grande que los demás.

Revisar el dato y ajustar con y sin él.

Asimetría

Un histograma de los residuos muy torcido.

Importa con muestras chicas.

Para llevarse

- El ANOVA compara más de dos medias con una sola prueba: la variabilidad entre los grupos contra la de dentro.
- Si el F es significativo, un post hoc como Tukey dice cuáles grupos difieren.
- La regresión pone a prueba la pendiente: H_0: $\beta_1 = 0$, con $t = b_1 / EE(b_1)$ y $n - 2$ grados de libertad.
- η^2 y R^2 dicen cuánto se explica; el p-valor, solo si hay algo.
- Antes de creerle a un modelo, mirar los residuos.

Guía de trabajo práctico 12
estadistica.ar/docentes/trabajos-practicos/semana-12

Calculadora de correlación y regresión
estadistica.ar/herramientas/calc-correlacion-regresion

Calculadora de ANOVA
estadistica.ar/herramientas/calc-anova

Ejercicios: inferencia sobre la pendiente
estadistica.ar/ejercicios/inferencia-pendiente