ESTADISTICA.AR

ANOVA y regresión

Estadística · Semana 12 · Unidad 7: ANOVA y regresión

1 / 29

NOTAS · 1

Pensada para 90 a 120 minutos. Es la última clase de contenidos nuevos: dos herramientas que extienden lo de la semana pasada. El ANOVA compara más de dos medias; la regresión pone a prueba la recta de la semana 4.

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?

2 / 29

NOTAS · 2

Porque con cinco regiones hay 10 pares, y con 10 pruebas al 5 % la probabilidad de al menos un falso positivo, si no hubiera ninguna diferencia, ronda $1 - 0{,}95^{10} = 40\ \%$. Es el problema de las comparaciones múltiples de la semana pasada.

El ANOVA hace una sola prueba para todas las medias juntas. Si rechaza, recién ahí se buscan los pares que difieren, con un método que controla el error.



NOTAS · 3

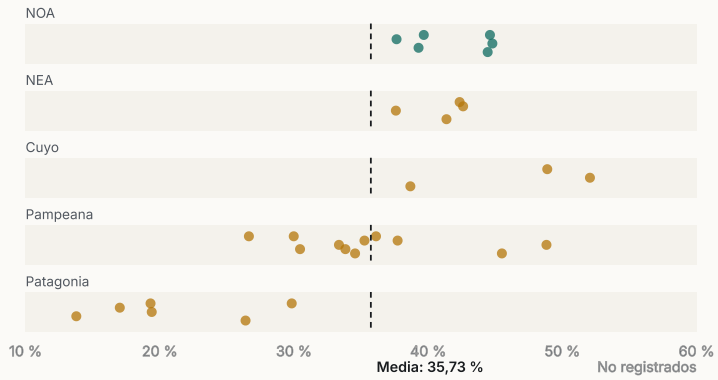
Las dos primeras partes y la tercera son el ANOVA; las dos últimas, la regresión. Las dos son casos del mismo modelo lineal, aunque eso queda para un curso siguiente.



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.

5 / 29

NOTAS · 5

Antes de la prueba, el gráfico: los seis aglomerados de la Patagonia están por debajo del 30 %, y de las otras regiones solo dos, Bahía Blanca y Santa Rosa. Las otras cuatro regiones se superponen bastante.

Gran Catamarca, en el NOA, tiene 39,67 %: cuatro de cada diez asalariados sin aportes.

SEMANA 12 · MÁS DE DOS MEDIAS

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.

6 / 29

NOTAS · 6

H_1 no dice que todas sean distintas, ni cuál lo es: solo que no son todas iguales. Eso es lo que contesta el ANOVA, y por eso después hace falta el post hoc.

La media general es el promedio de los 30 aglomerados, no el de las cinco medias: pesa cada región por su cantidad de aglomerados.

SEMANA 12 · MÁS DE DOS MEDIAS

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.

7 / 29

NOTAS · 7

Se llama análisis de la varianza aunque compara medias: lo hace comparando dos estimaciones de la varianza.

Volver al gráfico: las franjas separadas y los puntos apretados dentro de cada una dan un F grande. Con las mismas medias pero puntos muy dispersos, el F sería chico.

2

La tabla ANOVA

Sumas de cuadrados y la prueba F

8 / 29

SEMANA 12 · LA TABLA ANOVA

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.

9 / 29

NOTAS · 9

Las dos sumas salen de la tabla de resúmenes: no hace falta tener los datos. Es lo que permite completar una tabla ANOVA a partir de lo que publica un artículo.

La suma de cuadrados dentro es la misma idea que la varianza combinada de la prueba t: cada grupo aporta su varianza, pesada por sus grados de libertad.

SEMANA 12 · LA TABLA ANOVA

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		

$F = CM\ entre\ /\ CM\ dentro = 496{,}49\ /\ 30{,}84 = 16{,}10$

10 / 29

NOTAS · 10

Los grados de libertad: entre, la cantidad de grupos menos uno ($5 - 1 = 4$); dentro, la cantidad de datos menos la de grupos ($30 - 5 = 25$). Suman 29, los del total.

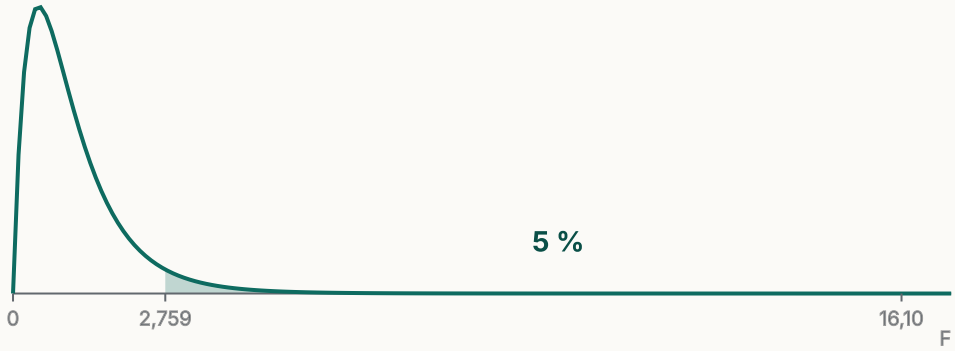
Los cuadrados medios son varianzas: el de dentro estima la varianza común de los grupos; el de entre, también, pero solo si H_0 es cierta. Si las medias difieren, el de entre crece.

El cuadrado medio dentro, 30,84, es una varianza: su raíz, 5,55 puntos, es el desvío típico de un aglomerado alrededor de la media de su región.

SEMANA 12 · LA TABLA ANOVA

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.

11 / 29

NOTAS · 11

Como la chi-cuadrado, la F es asimétrica, solo toma valores positivos, y la prueba es de cola derecha. Tiene dos grados de libertad: los del numerador y los del denominador.

El p-valor exacto es de 1,2 en un millón. En la tabla F alcanza con ver que 16,10 supera largamente al valor crítico, aun al del 1 %.

SEMANA 12 · LA TABLA ANOVA

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.

12 / 29

NOTAS · 12

Eta cuadrado es la proporción de la variabilidad total que está entre los grupos. Es la misma idea que el R^2 de la regresión, que aparece al final de la clase.

Acá es grande: saber la región de un aglomerado dice mucho de su informalidad laboral. Con miles de datos, un F significativo puede venir con un η^2 de 0,01, y ahí la diferencia importa poco.





NOTAS · 14

Es la misma lógica de la semana pasada: el ANOVA de Welch es al ANOVA lo que la prueba de Welch es a la t con varianza combinada.

Con grupos chicos, que una prueba de supuestos no rechace no quiere decir que el supuesto se cumpla: puede ser falta de potencia. Por eso conviene mirar los desvíos (2,3 a 7,0) y, ante la duda, usar el método robusto.

La independencia es la más importante y la que ninguna prueba revisa: datos repetidos de las mismas personas, por ejemplo, necesitan otro análisis.

SEMANA 12 · SUPUESTOS Y POST HOC

¿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

15 / 29

NOTAS · 15

La mínima diferencia significativa de cada par es $q \cdot \sqrt{(CMD / 2 \cdot (1/n_i + 1/n_j))}$, con q = 4,15 de la tabla del rango estudentizado para 5 grupos y 25 grados de libertad. Un ejemplo: Cuyo y Pampeana difieren en 10,88 puntos; la mínima diferencia significativa es $4{,}15 \cdot \sqrt{(30{,}84 / 2 \cdot (1/3 + 1/11))} = 10{,}61$. Como la diferencia la supera, por poco, difieren.

La Patagonia difiere de las otras cuatro, y Cuyo de la Pampeana. El NOA, el NEA y Cuyo no se distinguen entre sí: con estos datos, el norte y Cuyo forman un bloque de informalidad alta.

Tukey controla la probabilidad de al menos un falso positivo en las diez comparaciones juntas: por eso su umbral es más exigente que el de una t suelta. La q sale de la tabla del rango estudentizado.

4

La regresión lineal simple

El modelo y sus supuestos

16 / 29

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

17 / 29

NOTAS · 17

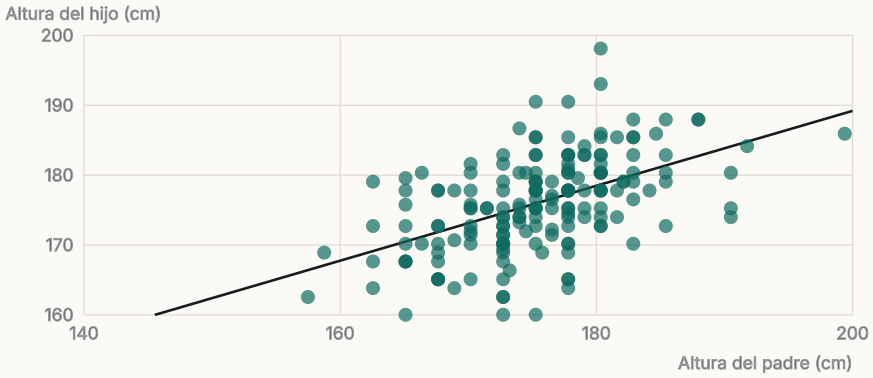
Menos: unos 5,4 cm, según la recta de la próxima diapositiva. Los hijos de padres altos son altos, pero en promedio menos extremos que sus padres. Con los padres bajos pasa lo mismo, al revés.

Galton lo llamó «regresión hacia la mediocridad», y de ahí viene el nombre de la técnica. No es una fuerza biológica: pasa siempre que la correlación no es perfecta, y también al revés, de los hijos a los padres.

SEMANA 12 · LA REGRESIÓN LINEAL SIMPLE

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.

18 / 29

NOTAS · 18

Galton midió en pulgadas, y anotó casi todas las alturas de los padres redondeadas a la media pulgada: por eso los puntos forman columnas. Se toma un solo hijo por familia porque los hermanos se parecen entre sí, y dos hermanos no son dos observaciones independientes.

Antes de cualquier cuenta, el gráfico: una nube alargada que sube, sin curvas y con un ancho parecido en todo el recorrido. Una recta parece razonable.

SEMANA 12 · LA REGRESIÓN LINEAL SIMPLE

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.

19 / 29

NOTAS · 19

La pendiente es menor que 1 porque r es menor que 1 y los dos desvíos son parecidos: esa es la regresión a la media. Con sy = sx, la pendiente sería exactamente r.

Las medias de padres e hijos son casi iguales (175,5 y 176,1): la regresión a la media no achica la variabilidad de una generación a la otra. Los hijos de los extremos se acercan al centro, y los de padres del centro se reparten hacia los dos lados.

SEMANA 12 · LA REGRESIÓN LINEAL SIMPLE

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.

20 / 29

NOTAS · 20

La semana 4 la recta describía una nube. Ahora es un modelo para la población: β_0 y β_1 son parámetros, y b_0 y b_1, sus estimadores. Cambian de muestra en muestra, como la media muestral.

ε es todo lo que la altura del padre no explica: la madre, la alimentación, el azar genético. Los supuestos son sobre ε, y se revisan con los residuos, que son sus estimaciones.

5

Inferencia y residuos

La prueba para la pendiente y el diagnóstico

21 / 29

SEMANA 12 · INFERENCIA Y RESIDUOS

¿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: $p < 0{,}001$.

22 / 29

NOTAS · 22

Los grados de libertad son n − 2 porque se estimaron dos parámetros, la ordenada y la pendiente.

Es la misma prueba que la de la correlación: H_0: $\rho = 0$ da el mismo t y el mismo p. La salida de cualquier software trae esta prueba en la fila de la pendiente.

s_e sale de los resúmenes: $s_e = sy \cdot \sqrt{((1 - r^2) \cdot (n - 1) / (n - 2))} = 5{,}85$. En la práctica lo da el software, como «error estándar residual».

SEMANA 12 · INFERENCIA Y RESIDUOS

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.

23 / 29

NOTAS · 23

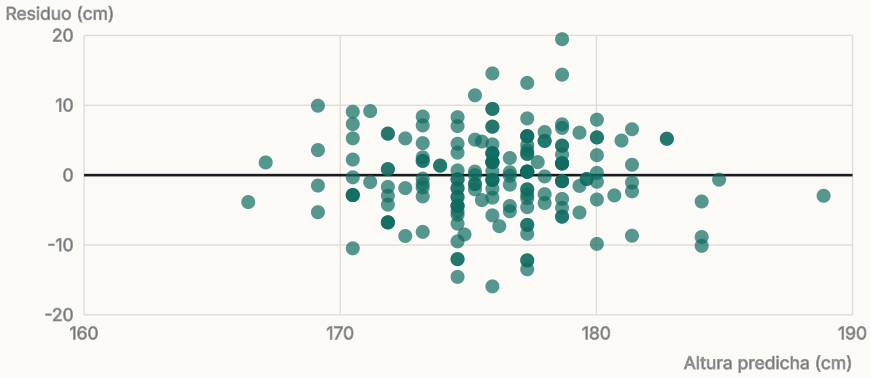
Por cada 10 cm más del padre, el hijo mide en promedio entre 4,0 y 6,7 cm más. El intervalo no contiene al 0, que es lo que dijo la prueba, y tampoco al 1: la regresión a la media no es casualidad de la muestra.

El 74 % restante es lo que la recta no explica. Para un padre dado, los hijos se reparten con un desvío de unos 5,8 cm alrededor de la recta.

SEMANA 12 · INFERENCIA Y RESIDUOS

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.

24 / 29

NOTAS · 24

Así tiene que verse un gráfico de residuos: una franja horizontal alrededor del cero, sin curvas, sin embudos y sin puntos muy alejados. Los supuestos son razonables.

El gráfico de residuos agranda lo que en la dispersión original es difícil de ver, porque saca la tendencia y deja solo lo que sobra.

SEMANA 12 · INFERENCIA Y RESIDUOS

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	ŷ = 3,00 + 0,500 · x
II	9,00	7,50	2,03	0,82	ŷ = 3,00 + 0,500 · x
III	9,00	7,50	2,03	0,82	ŷ = 3,00 + 0,500 · x
IV	9,00	7,50	2,03	0,82	ŷ = 3,00 + 0,500 · x

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

25 / 29

NOTAS · 25

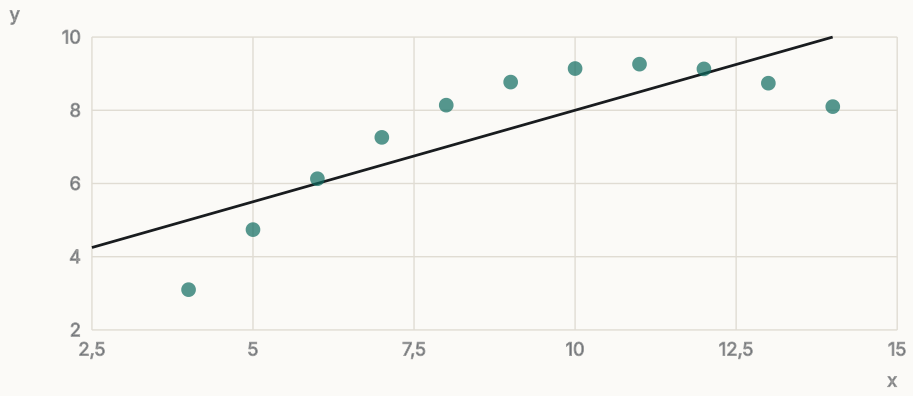
Los mismos resúmenes, la misma correlación, la misma recta y, aunque no está en la tabla, el mismo t para la pendiente. Si solo se mira la salida del software, los cuatro conjuntos son iguales.

Que adivinen cómo son los gráficos antes de verlos. El I es una nube común; los otros, no.

SEMANA 12 · INFERENCIA Y RESIDUOS

El conjunto II: una curva

ŷ = 3,00 + 0,500 · x



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

26 / 29

NOTAS · 26

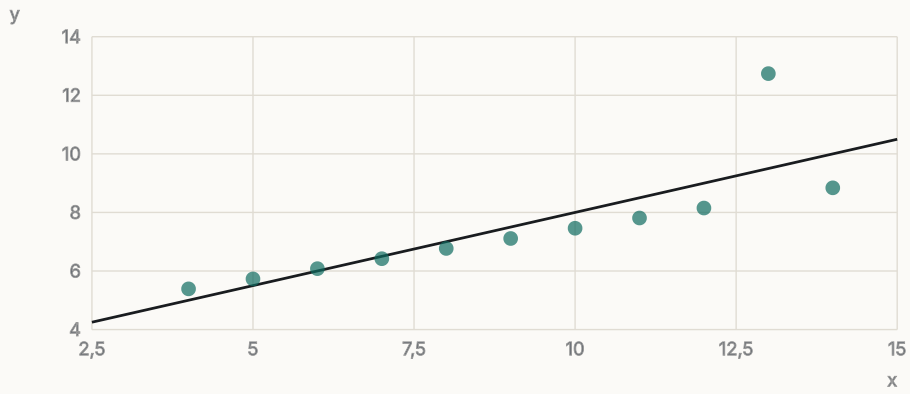
Los puntos siguen una curva perfecta, y la recta pasa por debajo en el centro y por arriba en los extremos. Los residuos dibujarían una U invertida: el supuesto de linealidad no se cumple.

Un r de 0,82 no garantiza que la relación sea lineal: solo dice que una recta resume bastante bien. Acá una parábola la resumiría sin error.

SEMANA 12 · INFERENCIA Y RESIDUOS

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.

27 / 29

NOTAS · 27

Diez puntos sobre una recta exacta y uno muy alejado, que tira de la recta hacia arriba. Sin él, la pendiente sería más chica y r valdría casi 1.

En el conjunto IV es peor: todos los x valen 8 salvo uno, y ese único punto decide toda la recta. Un punto así se llama influyente.

SEMANA 12 · INFERENCIA Y RESIDUOS

Qué buscar en los residuos

Una curva	Un embudo	Puntos alejados	Asimetría
Los residuos suben y bajan con x.	Los residuos se abren a medida que crece x.	Un residuo mucho más grande que los demás.	Un histograma de los residuos muy torcido.
La relación no es lineal.	La varianza no es constante: pasa con los ingresos.	Revisar el dato y ajustar con y sin él.	Importa con muestras chicas.

28 / 29

NOTAS · 28

Siempre, antes de leer los p-valores de una regresión, el gráfico de residuos. Las pruebas de la pendiente suponen que el modelo es razonable; si no lo es, los números no significan lo que parecen.

En el práctico, la regresión del ingreso sobre la edad muestra uno de estos patrones. Que lo encuentren ellos.

SEMANA 12 · ANOVA Y REGRESIÓN

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 ANOVA
estadistica.ar/herramientas/calc-anova

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

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

29 / 29

NOTAS · 29

Para el práctico: las tablas F, t y del rango estudentizado, y calculadora. El ejercicio con software usa la EPH. Las semanas que siguen son de evaluación.