Trabajo práctico 12: ANOVA y regresión

AL TERMINAR ESTA GUÍA

- Comparar más de dos medias con el ANOVA de un factor: armar y completar la tabla ANOVA, y leer la prueba F.
- Identificar qué grupos difieren con una comparación post hoc, y entender por qué no se hacen muchas pruebas t.
- Ajustar una recta de mínimos cuadrados, interpretar sus coeficientes y probar si la pendiente es distinta de cero.
- Revisar los supuestos del ANOVA y de la regresión, y reconocer en los residuos un modelo que no ajusta.

PARA EL DOCENTE

Los ejercicios 1 y 3 son los centrales del encuentro práctico, con las tablas de la F y de la t. El 2 es corto y sirve de repaso del 1; el 4 se hace con calculadora y conviene que los residuos los grafique cada grupo.

El 5 necesita computadora. El 6 se puede hacer al final, en voz alta: junta los errores de interpretación más comunes de toda la segunda parte de la materia.

EJERCICIO 1 Los talleres de preparación para el ingreso

A mano

Una facultad ofreció dos talleres de preparación para el examen de ingreso, uno presencial y otro virtual. Se eligieron al azar 6 aspirantes que hicieron el taller presencial, 6 que hicieron el virtual y 6 que no hicieron ninguno, y se registró su nota en el examen (de 0 a 100).

Taller presencial	69	71	65	80	64	86
Taller virtual	80	73	65	56	70	66
Sin taller	52	59	57	61	68	53

Se puede suponer que las notas tienen distribución aproximadamente normal en cada grupo. Se usa α = 0,05.

a) Planteá las hipótesis para responder si la nota media difiere según la preparación.

RESOLUCIÓN

$H_0 : \mu_1 = \mu_2 = \mu_3 \qquad H_1 :$ al menos una media es distinta

> La alternativa no es «las tres medias son distintas»: alcanza con que una se aparte de las otras. Por eso, si se rechaza H_0, todavía no se sabe cuáles difieren.

b) Calculá la media y el desvío de cada grupo y la media general. ¿Es razonable suponer varianzas iguales?

> RESOLUCIÓN
>
> | Grupo | Media | Desvío |
> |---|---|---|
> | Taller presencial | 72,50 | 8,73 |
> | Taller virtual | 68,33 | 8,12 |
> | Sin taller | 58,33 | 5,85 |
>
> $$\bar{x} = \frac{1\,195}{18} = 66{,}39$$
>
> Sí: el desvío mayor es 1,49 veces el menor, lejos del doble. Como los tres grupos tienen el mismo tamaño, además, el ANOVA es poco sensible a diferencias moderadas entre las varianzas.

c) Calculá la suma de cuadrados entre grupos, la suma de cuadrados dentro de los grupos y la total.

> RESOLUCIÓN
>
> $$SC_E = \sum n_i(\bar{x}_i - \bar{x})^2 = 6\left[(72{,}50 - 66{,}39)^2 + (68{,}33 - 66{,}39)^2 + (58{,}33 - 66{,}39)^2\right] = 636{,}11$$
>
> $$SC_D = \sum (n_i - 1)\, s_i^2 = 5\,(8{,}73^2 + 8{,}12^2 + 5{,}85^2) = 882{,}17$$
>
> $$SC_T = SC_E + SC_D = 636{,}11 + 882{,}17 = 1\,518{,}28$$
>
> La suma de cuadrados total mide cuánto varían las 18 notas alrededor de la media general. El ANOVA la parte en dos: lo que varían las medias de los grupos entre sí (entre) y lo que varían las notas dentro de cada grupo (dentro). La total también se puede calcular directamente, como suma de los cuadrados de las 18 desviaciones a la media general, y tiene que dar lo mismo.

d) Armá la tabla ANOVA y tomá la decisión con la tabla de la F. Acotá el p-valor.

> RESOLUCIÓN
>
> | Fuente | gl | SC | CM | F |
> |---|---|---|---|---|
> | Entre grupos | 2 | 636,11 | 318,06 | 5,408 |
> | Dentro de los grupos | 15 | 882,17 | 58,81 | |
> | Total | 17 | 1.518,28 | | |
>
> $$F = \frac{CM_E}{CM_D} = \frac{318{,}06}{58{,}81} = 5{,}408 > F_{2;\,15;\,0{,}95} = 3{,}682$$
>
> Se rechaza H_0. Con α = 0,01 el valor crítico es 6,359 y no se rechazaría: 0,01 < p < 0,05. El valor exacto es p = 0,0170. Los grados de libertad son k − 1 = 2 entre grupos y N − k = 15 dentro.

> La lógica del cociente: si H_0 es cierta, CM entre y CM dentro estiman la misma varianza y F anda cerca de 1. Si las medias difieren, CM entre crece y F se hace grande. Por eso la región crítica está solo a la derecha.

e) ¿Por qué no comparar los grupos de a pares con tres pruebas t al 5 %?

> RESOLUCIÓN
>
> $$P(\text{al menos un falso positivo}) = 1 - 0{,}95^3 = 0{,}143$$
>
> Si las tres medias fueran iguales, con tres pruebas al 5 % la probabilidad de declarar alguna diferencia por azar sería casi del 15 %, no del 5 %. El ANOVA hace una sola prueba para la pregunta global y mantiene el nivel de significación.

f) Aplicá la prueba de Tukey para saber qué grupos difieren. El valor de la tabla del rango estudentizado para 3 grupos y 15 grados de libertad es q = 3,67.

> RESOLUCIÓN
>
> $$DHS = q\sqrt{\frac{CM_D}{n}} = 3{,}67\sqrt{\frac{58{,}81}{6}} = 11{,}49$$
>
Comparación	Diferencia de medias	¿Supera la DHS?
Taller presencial – taller virtual	4,17	No
Taller presencial – sin taller	14,17	Sí
Taller virtual – sin taller	10,00	No
>
> Solo difieren el taller presencial y no hacer taller: 14,2 puntos de diferencia. El taller virtual queda en el medio y no se distingue de ninguno de los otros dos: con seis aspirantes por grupo no alcanza para saber si se parece más a uno o al otro. Es una situación frecuente y no es una contradicción: «no se distingue» no es «es igual».
>
> El estudio es observacional: quienes eligen hacer un taller pueden diferir en otras cosas (motivación, tiempo disponible, escuela de origen) de quienes no lo hacen. La diferencia no prueba que el taller la cause.

EJERCICIO 2 Completar una tabla ANOVA

A mano

Para comparar el tiempo de espera (en minutos) hasta ser atendido en 4 centros de atención primaria de una ciudad, se registró la espera de 10 pacientes elegidos al azar en cada centro. Al imprimir la salida del software se perdieron varias celdas de la tabla ANOVA:

Fuente	gl	SC	CM	F	p-valor
Entre centros	…	2.705,8	…	…	…
Dentro de los centros	…	…	…		
Total	…	7.120,4			

a) ¿Cuántos grupos y cuántas observaciones hay? Completá los grados de libertad.

> RESOLUCIÓN
>
> $gl_E = k - 1 = 3 \qquad gl_D = N - k = 40 - 4 = 36 \qquad gl_T = N - 1 = 39$
>
> Los grados de libertad, igual que las sumas de cuadrados, se suman: el total se reparte entre las dos fuentes.

b) Completá la suma de cuadrados dentro, los cuadrados medios y el estadístico F.

> RESOLUCIÓN
>
> $SC_D = SC_T - SC_E = 7\,120{,}4 - 2\,705{,}8 = 4\,414{,}6$
>
> $CM_E = \frac{2\,705{,}8}{3} = 901{,}93 \qquad CM_D = \frac{4\,414{,}6}{36} = 122{,}63 \qquad F = \frac{901{,}93}{122{,}63} = 7{,}355$
>
> El cuadrado medio dentro es una estimación de la varianza de la espera dentro de cada centro: el desvío común es de unos 11,1 minutos.

c) Tomá la decisión con α = 0,05 y con α = 0,01, y acotá el p-valor.

> RESOLUCIÓN
>
> Con 3 y 36 grados de libertad, los valores críticos son 2,866 (α = 0,05) y 4,377 (α = 0,01). Como F = 7,355 supera a los dos, se rechaza H_0 en ambos casos: p < 0,01. El valor exacto es p = 0,0006, el que habría que escribir en la celda que falta.
>
> Muchas tablas impresas no traen la fila de 36 grados de libertad; se usa la más cercana hacia abajo (30), que es la más conservadora, o se calcula el valor exacto con software.

d) Calculá la proporción de la variabilidad total explicada por el centro (η^2). ¿Cómo la interpretás?

> RESOLUCIÓN
>
> $\eta^2 = \frac{SC_E}{SC_T} = \frac{2\,705{,}8}{7\,120{,}4} = 0{,}380$
>
> El 38,0 % de la variabilidad de los tiempos de espera se asocia al centro donde se atiende el paciente; el resto es variabilidad entre pacientes del mismo centro. Es la misma idea que el R^2 de la regresión, y cumple el papel del tamaño del efecto: el p-valor dice que hay diferencias, η^2 dice cuánto pesan.

e) Un directivo lee la tabla y concluye: «los cuatro centros tienen tiempos de espera distintos». ¿Es correcto?

> RESOLUCIÓN
>
> No. El ANOVA rechaza que las cuatro medias sean iguales: al menos una difiere. Puede ser que un solo centro se aparte de los otros tres, o que haya dos grupos de dos. Para saber cuáles difieren hacen falta las medias de cada centro y una comparación post hoc, como la de Tukey, que controla el error al hacer las seis comparaciones de a pares.

f) ¿Qué supuestos necesita el ANOVA? ¿Cómo los revisarías con los datos completos?

RESOLUCIÓN

Independencia: cada espera es de un paciente distinto, elegido al azar. Se rompe, por ejemplo, si se toman pacientes del mismo día y la misma hora, porque comparten la congestión del momento. Se asegura con el diseño, no con una prueba.

Normalidad en cada grupo: se mira con un histograma o un gráfico cuantil-cuantil de los residuos (cada espera menos la media de su centro). Los tiempos de espera suelen ser asimétricos a la derecha; si la asimetría es fuerte, conviene usar el logaritmo o la prueba de Kruskal-Wallis.

Varianzas iguales: se comparan los desvíos de los grupos (que el mayor no llegue al doble del menor) o se usa la prueba de Levene. Con grupos del mismo tamaño, como acá, el ANOVA tolera bastante bien varianzas algo distintas; si no, está el ANOVA de Welch.

EJERCICIO 3 La superficie de la vivienda y el consumo eléctrico

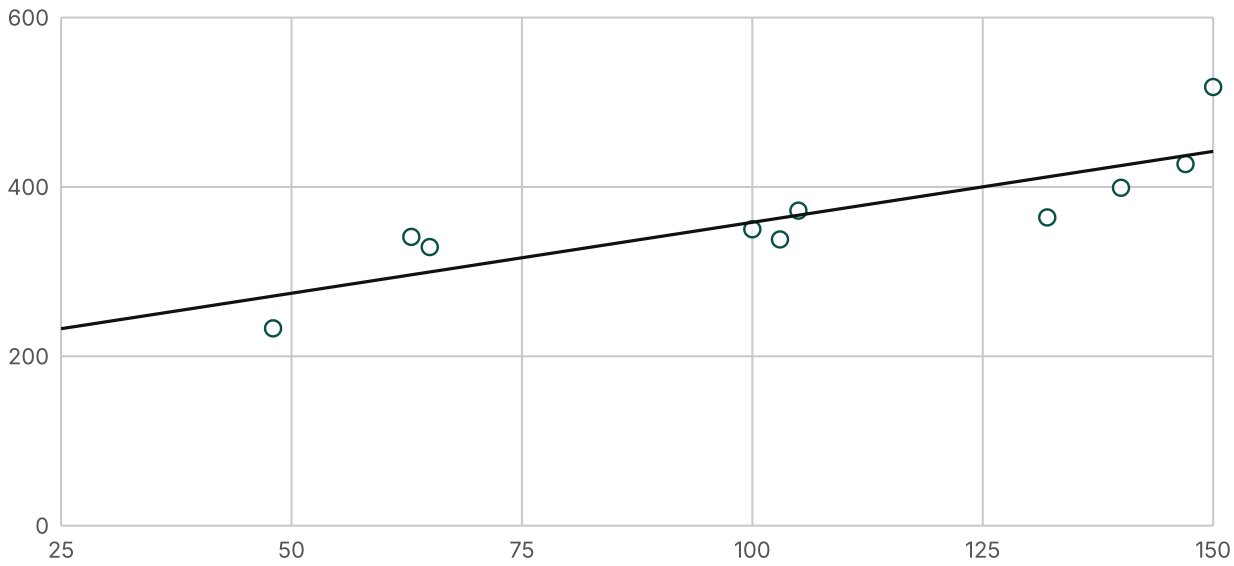
A mano

Una distribuidora eléctrica eligió al azar 10 viviendas de una ciudad y registró la superficie cubierta (en m²) y el consumo de electricidad del último bimestre (en kWh). Se puede usar el modo estadístico de la calculadora para las sumas.

Superficie, viviendas 1 a 5	63	48	140	147	132
Consumo, viviendas 1 a 5	341	233	399	427	364
Superficie, viviendas 6 a 10	150	100	103	105	65
Consumo, viviendas 6 a 10	518	350	338	372	329

a) Hacé el diagrama de dispersión, con la superficie en el eje horizontal. ¿Qué forma tiene la relación?

RESOLUCIÓN



Eje horizontal: superficie (m²). Eje vertical: consumo (kWh). La línea es la recta de mínimos cuadrados.

La relación es creciente y aproximadamente lineal, con bastante dispersión alrededor de la recta: viviendas de superficie parecida consumen distinto, según cuántas personas viven, qué artefactos tienen y cómo se calefaccionan. No se ven puntos muy apartados del resto. La superficie va en el eje horizontal porque es la variable que se usa para explicar o predecir el consumo.

b) Calculá la recta de mínimos cuadrados.

RESOLUCIÓN

$\bar{x} = 105{,}3 \qquad \bar{y} = 367{,}1 \qquad S_{xx} = \sum (x - \bar{x})^2 = 12\,384{,}1 \qquad S_{xy} = \sum (x - \bar{x})(y - \bar{y}) = 20\,746{,}7$

$b_1 = \dfrac{S_{xy}}{S_{xx}} = 1{,}675 \qquad b_0 = \bar{y} - b_1 \bar{x} = 367{,}1 - 1{,}675 \cdot 105{,}3 = 190{,}69$

$\hat{y} = 190{,}69 + 1{,}675\, x$

c) Interpretá la pendiente y la ordenada al origen en el contexto del problema.

RESOLUCIÓN

Pendiente: por cada metro cuadrado más de superficie, el consumo bimestral es, en promedio, 1,68 kWh mayor; o sea, unos 17 kWh más por cada 10 m². Es una diferencia promedio entre viviendas, no lo que pasaría si se ampliara una casa.

Ordenada: sería el consumo medio de una vivienda de 0 m², 191 kWh. No tiene interpretación propia: ninguna vivienda tiene superficie cero, y los datos van de 48 a 150 m². Solo sirve para ubicar la recta.

d) Calculá el coeficiente de determinación y el desvío de los residuos.

RESOLUCIÓN

$S_{yy} = 48\,664{,}9 \qquad SC_{\text{res}} = S_{yy} - b_1 S_{xy} = 13\,908{,}6 \qquad R^2 = 1 - \dfrac{SC_{\text{res}}}{S_{yy}} = 0{,}714$

$s_e = \sqrt{\dfrac{SC_{\text{res}}}{n-2}} = \sqrt{\dfrac{13\,908{,}6}{8}} = 41{,}70$

La superficie explica el 71,4 % de la variabilidad del consumo entre estas viviendas. El desvío de los residuos dice cuánto se equivoca típicamente la recta: unos 42 kWh. Se divide por n − 2 porque se estimaron dos coeficientes.

e) Probá si la pendiente poblacional es distinta de cero, con α = 0,05, y acotá el p-valor.

RESOLUCIÓN

$H_0 : \beta_1 = 0 \qquad H_1 : \beta_1 \neq 0$

$EE(b_1) = \dfrac{s_e}{\sqrt{S_{xx}}} = \dfrac{41{,}70}{\sqrt{12\,384{,}1}} = 0{,}3747 \qquad t = \dfrac{b_1}{EE(b_1)} = \dfrac{1{,}675}{0{,}3747} = 4{,}471$

Con 8 grados de libertad, t = 4,471 supera tanto a 2,306 (cola de 0,025) como a 3,355 (cola de 0,005): se rechaza H_0 y $p < 0{,}01$. El valor exacto es p = 0,0021. Hay evidencia de que el consumo medio aumenta con la superficie.

> $\beta_1 = 0$ es la hipótesis de que la recta poblacional es horizontal: que conocer la superficie no ayuda a predecir el consumo.

f) Construí el intervalo de confianza del 95 % para la pendiente.

> RESOLUCIÓN
>
> $1{,}675 \pm 2{,}306 \cdot 0{,}3747 = 1{,}675 \pm 0{,}864 \;\Rightarrow\; (0{,}811;\ 2{,}539)$
>
> Con 95 % de confianza, cada metro cuadrado adicional se asocia con entre 0,81 y 2,54 kWh más de consumo bimestral, en promedio. El intervalo no incluye al cero, coherente con la prueba; y es ancho, porque son solo 10 viviendas.

g) Estimá el consumo medio de las viviendas de 100 m². ¿Usarías la recta para una vivienda de 400 m²?

> RESOLUCIÓN
>
> $\hat{y} = 190{,}69 + 1{,}675 \cdot 100 = 358{,}2\ \text{kWh}$
>
> Para 100 m², sí: está dentro del rango de los datos. Para 400 m², no: es extrapolar muy lejos del rango observado (de 48 a 150 m²), y nada garantiza que la relación siga siendo lineal ahí. Una casa de 400 m² puede tener pileta climatizada o estar casi vacía.

EJERCICIO 4 El precio de los autos usados

A mano

En un sitio de ventas se eligieron 12 autos usados del mismo modelo y versión, en buen estado, y se registró su antigüedad (en años) y el precio pedido (en millones de pesos).

Antigüedad	1	2	3	4	6	7	8	9	12	13	14	15

Precio, autos 1 a 6	27,1	22,7	21,7	18,6	14,4	13,1

Precio, autos 7 a 12	11,3	10,3	6,8	5,4	5,0	4,7

El software ajustó la recta de mínimos cuadrados: precio = 25,477 − 1,539 · antigüedad, con R^2 = 0,955. La pendiente es muy significativa ($p < 0{,}001$).

a) Calculá el precio predicho por la recta y el residuo de cada auto.

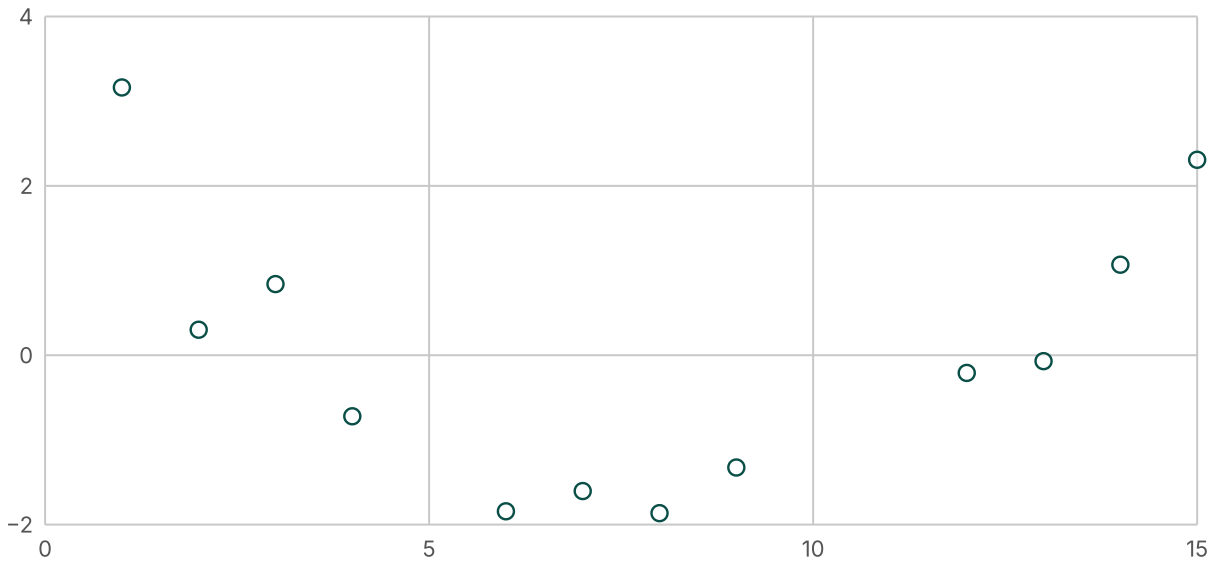
RESOLUCIÓN

Antigüedad	Precio	Predicho	Residuo
1	27,1	23,94	3,16
2	22,7	22,40	0,30
3	21,7	20,86	0,84
4	18,6	19,32	−0,72
6	14,4	16,24	−1,84
7	13,1	14,70	−1,60
8	11,3	13,17	−1,87
9	10,3	11,63	−1,33
12	6,8	7,01	−0,21
13	5,4	5,47	−0,07
14	5,0	3,93	1,07
15	4,7	2,39	2,31

El residuo es lo observado menos lo predicho: positivo si el auto cuesta más de lo que dice la recta, negativo si cuesta menos. Los residuos de una recta de mínimos cuadrados suman cero (salvo por el redondeo de los coeficientes).

b) Graficá los residuos contra la antigüedad. ¿Se ve algún patrón?

RESOLUCIÓN



Eje horizontal: antigüedad (años). Eje vertical: residuo (millones).

Sí: los residuos son positivos en los autos más nuevos y en los más viejos, y negativos en los del medio. Dibujan una U. Si el modelo fuera adecuado, los residuos deberían repartirse al azar alrededor de cero, sin ninguna forma.

c) ¿Qué dice ese patrón sobre el modelo? ¿No alcanzaba con un R^2 tan alto?

> RESOLUCIÓN
>
> Que la relación no es lineal: el precio cae rápido en los primeros años y cada vez más despacio después. La recta subestima el precio de los autos nuevos y de los viejos, y lo sobreestima en el medio. El R^2 de 0,955 es alto porque el precio baja con la antigüedad de manera muy consistente, pero no dice nada sobre la forma: una curva clara también da un R^2 alto con una recta. Por eso siempre hay que mirar los residuos.

d) Según la recta, ¿cuánto costaría un auto de 20 años?

> RESOLUCIÓN
>
> $\hat{y} = 25{,}477 - 1{,}539 \cdot 20 = -5{,}30$
>
> Un precio negativo: absurdo. Es una extrapolación, fuera del rango de antigüedades observado, y además con un modelo que ya se veía mal dentro del rango. La curva que sugieren los residuos nunca llegaría a cero.

e) Un compañero propone usar el logaritmo del precio. Con software, ajustá esa recta. ¿Qué cambia?

> RESOLUCIÓN
>
> $\ln(\text{precio}) = 3{,}436 - 0{,}1282 \cdot \text{antigüedad} \qquad R^2 = 0{,}996$
>
> El R^2 sube a 0,996 y los residuos ya no forman una U: van de −0,083 a 0,050 sin un patrón claro. El modelo dice que el auto pierde un porcentaje fijo de su valor por año: e^(−0,1282) = 0,8797, o sea, alrededor del 12,0 % anual. A los 20 años predice 2,4 millones: positivo, aunque sigue siendo una extrapolación.

f) Además de la linealidad, ¿qué otros supuestos de la regresión se revisan con los residuos?

> RESOLUCIÓN
>
> Varianza constante: que la dispersión de los residuos no crezca ni se achique a lo largo del eje horizontal (si dibujan un embudo, no se cumple). Normalidad: un histograma o un gráfico cuantil-cuantil de los residuos; importa sobre todo para las pruebas y los intervalos con muestras chicas. Independencia: que cada observación no dependa de otra; con datos en el tiempo se grafican los residuos en el orden en que se registraron. Y los puntos influyentes: un solo dato extremo puede mover mucho la recta.

EJERCICIO 5 Ingreso, nivel educativo y edad en la EPH

Con software

Se usa la base «Ocupación e ingresos en Gran Catamarca», extracto de la EPH del primer trimestre de 2026 (estadistica.ar/datos/eph-catamarca-2026t1), con las personas ocupadas que declararon ingreso (ESTADO = 1 y P21 > 0). En R: aov(log(P21) ~ grupo) para el ANOVA, pairwise.t.test(..., p.adjust.method = "bonferroni") para las comparaciones de a pares y lm(log(P21) ~ CH06) para la regresión.

Se agrupa el nivel educativo (NIVEL_ED) en tres categorías: hasta secundario incompleto (1 a 3), secundario completo o superior incompleto (4 y 5) y superior completo (6). Como en las semanas anteriores, se trata a la base como una muestra aleatoria simple, sin el ponderador.

a) Calculá el desvío del ingreso y el de su logaritmo en cada grupo. ¿En qué escala es razonable el supuesto de varianzas iguales?

RESOLUCIÓN

Nivel educativo	n	Desvío del ingreso	Media del logaritmo	Desvío del logaritmo
Hasta secundario incompleto	137	341.497	12,944	0,748
Secundario completo o superior incompleto	294	986.445	13,293	0,735
Superior completo	163	1.012.204	13,880	0,696

En pesos, el desvío mayor es 3,0 veces el menor: los grupos que ganan más también tienen ingresos mucho más dispersos, algo habitual en los ingresos. En logaritmos, la razón es 1,08: las varianzas son parecidas, y además el logaritmo corrige buena parte de la asimetría. Conviene hacer el ANOVA sobre el logaritmo.

b) Hacé el ANOVA del logaritmo del ingreso según el nivel educativo. Informá F, los grados de libertad, el p-valor y η^2.

RESOLUCIÓN

Fuente	gl	SC	CM	F
Nivel educativo	2	68,948	34,474	65,08
Residuos	591	313,077	0,5297	

F = 65,08 con 2 y 591 grados de libertad, $p < 0{,}001$: el ingreso medio (en logaritmos) difiere según el nivel educativo. $\eta^2 = 0{,}180$: el nivel educativo explica el 18,0 % de la variabilidad del logaritmo del ingreso.

c) Compará los grupos de a pares con la corrección de Bonferroni. ¿Cuánto más gana, en proporción, cada grupo que el anterior?

RESOLUCIÓN

Comparación	Diferencia de logaritmos	t	p ajustado	Razón de ingresos
Secundario completo o superior incompleto contra hasta secundario incompleto	0,349	4,64	< 0,001	1,42
Superior completo contra hasta secundario incompleto	0,936	11,09	< 0,001	2,55
Superior completo contra secundario completo o superior incompleto	0,587	8,25	< 0,001	1,80

Los tres pares difieren. Una diferencia de logaritmos se lee como una razón entre ingresos típicos (las medias geométricas): terminar el secundario se asocia con un ingreso 1,42 veces el de quienes no lo terminaron (e^0,349), y terminar el nivel superior, con uno 1,80 veces el del grupo intermedio. Entre los extremos, la razón es 2,55. Bonferroni multiplica cada p-valor por 3, el número de comparaciones; Tukey (TukeyHSD en R) da conclusiones muy parecidas.

d) Ajustá la regresión del logaritmo del ingreso sobre la edad (CH06). Interpretá la pendiente, su prueba y el R^2.

RESOLUCIÓN

$\widehat{\ln(\text{ingreso})} = 12{,}791 + 0{,}01412 \cdot \text{edad} \qquad EE(b_1) = 0{,}00261 \qquad t = 5{,}404 \qquad R^2 = 0{,}0470$

La pendiente es significativa (t = 5,40, gl = 592, p < 0,001): en promedio, cada año más de edad se asocia con un ingreso alrededor de 1,4 % mayor, y cada diez años, con uno 15,2 % mayor. Pero el R^2 es de apenas 0,047: la edad explica menos del 5 % de la variabilidad. Con 594 personas, una relación muy débil ya da significativa.

e) Calculá el residuo medio de la regresión en cada tramo de edad. ¿Qué muestra? ¿Es adecuada la recta?

RESOLUCIÓN

Edad	Personas	Residuo medio
Hasta 25	73	−0,257
26 a 35	137	−0,053
36 a 45	154	0,216
46 a 55	141	0,099
56 a 65	79	−0,174
Más de 65	10	−0,748

Los residuos medios son negativos en los más jóvenes, positivos entre los 36 y los 55 años y otra vez negativos después: el ingreso sube con la edad hasta la mitad de la vida laboral y después baja. Es la misma U invertida del gráfico de residuos contra la edad. La recta promedia una subida y una bajada, y por eso la pendiente es chica y el R^2 también. Un modelo mejor agrega la edad al cuadrado (regresión múltiple), aunque el tramo de más de 65 tiene pocos casos (10).

EJERCICIO 6 Lecturas de un ANOVA y de una regresión

Para discutir

Cada afirmación interpreta el resultado de un ANOVA o de una regresión. Decidí si es correcta y, si no lo es, corregila.

a) «El ANOVA dio p = 0,003: los cuatro tratamientos tienen efectos distintos entre sí.»

RESOLUCIÓN

Incorrecta. El ANOVA rechaza que las cuatro medias sean iguales: al menos una difiere. Para saber cuáles hacen falta comparaciones post hoc, y puede que varios tratamientos no se distingan entre sí.

b) «Con dos grupos, da lo mismo hacer un ANOVA o una prueba t con varianza combinada.»

> RESOLUCIÓN
>
> Correcta: con dos grupos el estadístico F es el cuadrado del t, y los p-valores coinciden (el de la t bilateral). El ANOVA generaliza la prueba t a más de dos grupos.

c) «La pendiente de la edad es significativa ($p < 0{,}001$), así que la edad explica buena parte del ingreso.»

> RESOLUCIÓN
>
> Incorrecta. Significativa quiere decir que es improbable que la pendiente poblacional sea cero, no que la relación sea fuerte. En la EPH, la pendiente es muy significativa y el R^2 no llega a 0,05. Cuánto explica lo dice el R^2; con muestras grandes, relaciones muy débiles dan p-valores chicos.

d) «El R^2 dio 0,02: entre las dos variables no hay ninguna relación.»

> RESOLUCIÓN
>
> Incorrecta. El R^2 mide la relación lineal. Puede haber una relación fuerte y curva (una U, como el ingreso y la edad) con un R^2 cercano a cero. Antes de concluir, hay que mirar el diagrama de dispersión.

e) «La recta predice que una vivienda de 0 m^2 consume 150 kWh: el modelo está mal.»

> RESOLUCIÓN
>
> Incorrecta. La ordenada al origen está fuera del rango de los datos y no tiene interpretación; solo ubica la recta. Que el modelo esté bien o mal se juzga dentro del rango observado, con los residuos.

f) «En los departamentos con más centros de salud hay más defunciones por año: los centros de salud no mejoran la salud de la población.»

> RESOLUCIÓN
>
> Incorrecta. Las dos variables crecen con la población del departamento: es una variable de confusión. Habría que comparar tasas (defunciones por cada mil habitantes, centros por cada diez mil) y, aun así, una regresión con datos observacionales muestra asociación, no causa.

g) «Los estudiantes con las peores notas del primer parcial mejoraron mucho en el segundo: la tutoría que se les dio funcionó.»

> RESOLUCIÓN
>
> No se puede saber. Es esperable aunque la tutoría no sirva: quienes sacaron notas extremadamente bajas en parte tuvieron mala suerte ese día, y en la segunda medición tienden a acercarse a la media. Es la regresión a la media. Para evaluar la tutoría hace falta un grupo de comparación con notas igual de bajas que no la haya recibido.

---

estadistica.ar/docentes/trabajos-practicos/semana-12 · Las lecturas de cada tema y la hoja de fórmulas están en el sitio.