

Guía 2 – Estadística Bivariada y Regresión

Correlación · Recta de Regresión · R^2 · Modelo Exponencial · Salidas de R

Problema 1 – Correlación y Predicción

En un programa de intervención nutricional se estudió la relación entre el **número de meses de seguimiento** (variable independiente) y el **porcentaje de reducción del colesterol LDL** (variable dependiente):

Meses (x)	1	3	5	7	9	11
Reducción LDL % (y)	5	12	18	26	33	40

- ¿Existe evidencia de asociación lineal entre ambas variables? Justifique utilizando un estadístico adecuado.
- Determine la recta de regresión ajustada.
- Si un paciente completara **13 meses** de seguimiento, ¿cuál sería una estimación de su porcentaje de reducción del colesterol?

Problema 2 – Regresión a partir de Resumen Descriptivo

Para estudiar la relación entre los años de experiencia (X) y el sueldo mensual en cientos de miles de pesos (Y) se dispone del siguiente resumen de una muestra de $n = 12$ trabajadores:

$$\bar{x} = 8, \quad \bar{y} = 15, \quad s_x^2 = 4, \quad s_y^2 = 10, \quad \text{Cov}(X, Y) = 6.$$

- Calcule el coeficiente de correlación r e interprételo.
- Determine la recta de regresión $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$.
- Calcule el coeficiente de determinación R^2 e interprételo.
- Prediga el sueldo de un trabajador con 10 años de experiencia.

Recuerde: $r = \frac{\text{Cov}(X, Y)}{s_x s_y}$; $\hat{\beta}_1 = \frac{\text{Cov}(X, Y)}{s_x^2}$; $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$; $R^2 = r^2$ en regresión lineal simple.

Problema 3 – Lectura de Salida en R (Modelo No Significativo)

Se ajustó en R un modelo lineal para explicar el nivel de colesterol (mg/dL) en función del número de hijos, con $n = 15$ personas. La salida entrega:

$$\widehat{\text{colesterol}} = 195,60 + 1,80 \cdot \text{hijos}, \quad R^2 = 0,0103, \quad F = 0,135 \text{ (} p\text{-value} = 0,719\text{)}.$$

- Escriba la ecuación del modelo ajustado e interprete sus coeficientes.
- Interprete el valor de R^2 . ¿Es un buen ajuste?
- Realice el test F de significancia global del modelo con $\alpha = 0,05$ (hipótesis, decisión y conclusión en contexto).

Problema 4 – Modelo Exponencial

Se midió el crecimiento de un cultivo de bacterias (millones de células, Y) a lo largo del tiempo (horas, X):

x_i	0	1	2	3	4
y_i	2	3	5	9	15

Se propone el modelo exponencial $y = \alpha e^{\beta x}$.

- Estime α y β linealizando el modelo con $z = \ln y$.
- Prediga la cantidad de células a las 5 horas.
- ¿El modelo es creciente o decreciente? Justifique con el signo de $\hat{\beta}$.

Recuerde: Si $y = \alpha e^{\beta x}$, entonces $\ln y = \ln \alpha + \beta x$: una recta en $(x, \ln y)$. Estime la pendiente y el intercepto sobre los datos transformados y luego destransforme: $\hat{\alpha} = e^{\ln \alpha}$.

Problema 5 – Comparación de Modelos en R

Para modelar las ventas mensuales (miles de unidades, Y) en función del tiempo (meses, X), con $n = 20$, se ajustaron dos modelos:

Modelo	Ecuación ajustada	R^2	p -value de $\hat{\beta}_1$
A (exponencial)	$\widehat{\ln Y} = 1,6094 + 0,2231 X$	0,9774	$< 2 \times 10^{-16}$
B (lineal)	$\hat{Y} = -10,50 + 2,85 X$	0,4720	$8,1 \times 10^{-4}$

- a) ¿Es $\hat{\beta}_1$ significativa en cada modelo? ($\alpha = 0,05$).
- b) Prediga las ventas para $X = 10$ con cada modelo.
- c) ¿Cuál modelo es preferible? Justifique con R^2 .

Respuestas

Problema 1. a) $r \approx 0,9995$, muy cercano a $+1$: asociación lineal positiva muy fuerte. **b)** $\hat{\beta}_1 = \frac{246}{70} \approx 3,514$; $\hat{\beta}_0 = 22,333 - 3,514 \cdot 6 \approx 1,25$; $\hat{y} = 1,25 + 3,51x$. **c)** $\hat{y}(13) \approx 1,25 + 3,51 \cdot 13 \approx 46,9$ % de reducción.

Problema 2. a) $s_x = 2$, $s_y \approx 3,1623$; $r = \frac{6}{2 \cdot 3,1623} \approx 0,9487$: asociación lineal positiva muy fuerte (a mayor experiencia, mayor sueldo). **b)** $\hat{\beta}_1 = \frac{6}{4} = 1,5$; $\hat{\beta}_0 = 15 - 1,5 \cdot 8 = 3$; $\hat{y} = 3 + 1,5x$. **c)** $R^2 = r^2 \approx 0,90$: el 90 % de la variabilidad del sueldo es explicada por los años de experiencia. **d)** $\hat{y}(10) = 3 + 15 = 18$ cientos de miles = \$1,800,000.

Problema 3. a) $\widehat{\text{col}} = 195,60 + 1,80 \cdot \text{hijos}$: por cada hijo adicional el colesterol estimado aumenta 1,80 mg/dL; con 0 hijos se estima 195,60 mg/dL. **b)** $R^2 = 0,0103$: el modelo explica solo el 1,03 % de la variabilidad; ajuste muy malo. **c)** $H_0 : \beta_1 = 0$ vs $H_1 : \beta_1 \neq 0$; $p = 0,719 > 0,05$: no se rechaza H_0 . El número de hijos no es un predictor significativo del colesterol.

Problema 4. a) $z_i = \ln y_i$: 0,6931, 1,0986, 1,6094, 2,1972, 2,7081. $\hat{\beta} \approx 0,5128$; $\widehat{\ln \alpha} \approx 0,6357$; $\hat{\alpha} \approx 1,89$. Modelo: $\hat{y} = 1,89 e^{0,51x}$. **b)** $\hat{y}(5) = 1,89 e^{2,564} \approx 24,5$ millones de células. **c)** Creciente: $\hat{\beta} \approx 0,51 > 0$ (crece ≈ 67 % por hora).

Problema 5. a) En ambos $p < 0,05$: se rechaza $H_0 : \beta_1 = 0$; X es significativa en A y B. **b)** Modelo A: $\widehat{\ln Y}(10) = 3,8404 \Rightarrow \hat{Y}_A = e^{3,8404} \approx 46,5$ miles. Modelo B: $\hat{Y}_B = -10,50 + 28,50 = 18,0$ miles. **c)** Preferible el Modelo A: $R_A^2 = 0,9774$ vs $R_B^2 = 0,4720$; la relación subyacente es de naturaleza exponencial.