TEMA 10
Correlación y regresión.
El modelo de regresión simple

Karl Pearson (1857-1936)

1. Introducción. Modelos matemáticos
2. Métodos numéricos. Resolución de sistemas lineales y ecuaciones no lineales
3. Aproximación de funciones: interpolación y ajuste
4. Modelos discretos elementales. Ecuaciones en diferencias
5. Estadística descriptiva. Análisis de datos
6. Variable aleatoria. Distribuciones de probabilidad
7. Distribuciones de probabilidad importantes
8. Estimación de parámetros por intervalos de confianza
9. Contraste de hipótesis. Introducción al análisis de la varianza
10. **Correlación y regresión. El modelo de regresión simple**

- **Introducción. Modelos de regresión y correlación**
- **El modelo de regresión lineal simple**
- **Análisis de la correlación**

Clases estimadas para este tema: 3 clases

Objetivo: Técnicas estadísticas para analizar relación entre dos variables cuantitativas

Consideraciones:

- predecir una variable a partir de otra
- dos variables ¿aleatorias? $\rightsquigarrow$ modelos distintos
- ¡no relación causa-efecto!
- relación lineal entre ambas

El modelo de regresión lineal simple

modelo de **regresión lineal simple**

$$y = \beta_0 + \beta_1 x + u$$

u son las **perturbaciones**

hipótesis sobre u: $\rightsquigarrow$ y

- media nula $\Rightarrow$ $E(y) = \beta_0 + \beta_1 x$
- varianza constante $\Rightarrow$ $V(y) = \sigma^2$
- distribución normal (TCL) $\Rightarrow$ $y \sim N(\beta_0 + \beta_1 x, \sigma^2)$
- son independientes $\Rightarrow$ y independientes

Objetivo: Estimar los parámetros β_0, β_1 y σ^2

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

Metodología:

- dibujar la **nube de puntos** $\rightsquigarrow$ ¿tendencia?
- modelo lineal $\rightsquigarrow$ estimar parámetros
- método de mínimos cuadrados $\rightsquigarrow$ $\hat{\beta}_0$, $\hat{\beta}_1$

$$\begin{cases} \displaystyle\sum_{i=1}^{n} y_i = nb + a \sum_{i=1}^{n} x_i \\ \displaystyle\sum_{i=1}^{n} x_i y_i = b \sum_{i=1}^{n} x_i + a \sum_{i=1}^{n} x_i^2 \end{cases}$$

ecuaciones normales

- **residuos** $e_i = y_i - \hat{y}_i \rightsquigarrow$ análisis de e_i

Ejercicio: Se desea establecer una ecuación mediante la cual pueda predecirse el tiempo de reproducción en base al conocimiento del fotoperiodo bajo el que se inició la reproducción. Se tienen los siguientes datos en 11 patos buceadores (Aythya).

horas de luz	12.8	13.9	14.1	14.7	15	15.1	16	16.5	16.6	17.2	17.9
tiempo de reproducción	110	54	98	50	67	58	52	50	43	15	28

Aspectos de las ecuaciones normales:

ecuación en términos de valores muestrales $y - \bar{y} = \dfrac{S_{xy}}{S_x^2}(x - \bar{x})$

ecuaciones para los residuos $\Rightarrow$ **varianza residual** $\rightsquigarrow$ σ^2

descomposición de la variación total

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$$

variación total = variación explicada + variación no explicada

coeficiente de determinación medida de precisión o ajuste

$$r^2 = \frac{\text{variación explicada}}{\text{variación total}}$$

$r^2 = 0 \Rightarrow$ ajuste nulo $\qquad$ $r^2 = 1 \Rightarrow$ ajuste perfecto

Ejercicio: ¿verdadero o falso?

1. $r^2 = 0 \Rightarrow$ no hay relación entre x e y.
2. Mayor pendiente en la recta de regresión $\Rightarrow$ mayor coeficiente de determinación.
3. Si la regresión es exacta $\Rightarrow |r| = 1$.
4. $1 - r^2$ representa la fracción de variabilidad no explicada.

Análisis de la correlación

Aspectos:

- x e y son dos variables aleatorias
- normales con $f(x, y)$ normal
- grado de asociación lineal entre ambas
- **coeficiente de correlación lineal** $\rho(x, y) = \dfrac{\mathrm{Cov}(x, y)}{\sigma_x \sigma_y}$

coeficiente de Pearson

$$r_{xy} = \frac{S_{xy}}{S_x S_y} \qquad -1 \leq r_{xy} \leq 1$$

Rectas de regresión

$$y - \bar{y} = \frac{S_{xy}}{S_x^2}(x - \bar{x}) \qquad x - \bar{x} = \frac{S_{xy}}{S_y^2}(y - \bar{y})$$

- $\star$ r_{xy} relacionado con los coeficientes de regresión

$$\frac{S_{xy}}{S_x^2} = r_{xy}\frac{S_y}{S_x} \qquad \frac{S_{xy}}{S_y^2} = r_{xy}\frac{S_x}{S_y}$$

- $\star$ r_{xy} relacionado con el coeficiente de determinación

$$r_{xy}^2 = r^2$$

r_{xy}

$\longrightarrow$ relación lineal

$\longrightarrow$ precisión de ajuste

Ejercicio: Diez pacientes con insuficiencia renal crónica:

caso	1	2	3	4	5	6	7	8	9	10
hemoglobina (g/dl)	9	9.5	8.4	11.7	10.8	8.2	9.2	8.9	10.1	11.5
creatinina (mg/dl)	4.3	2.8	6.5	2.4	3.7	8.0	5.3	7.9	4.3	3.2
BUN (mg/dl)	42.6	30.1	57.9	33.4	36.0	58.7	43.8	65.6	48.9	40.5

Hipótesis: el deterioro de la función renal (aumento de la creatinina o aumento del BUN), debería estar relacionado con una caída de la hemoglobina, puesto que la insuficiencia renal puede provocar anemia (¡pero la anemia no es símbolo de insuficiencia renal!). Analizamos la hipótesis del investigador.

Ejercicio: ¿verdadero o falso?

- El coeficiente de correlación es la media geométrica de los coeficientes de regresión lineal
- Es invariante ante un cambio de coordenadas
- $r = 0 \Rightarrow$ no hay relación entre x e y