Relación entre variables cuantitativas

Leccion 03

Omar E. Barrantes Sotela

Regresión lineal de variables

La **correlación** indica la **fuerza** y la **dirección** de una relación lineal y **proporcionalidad** entre dos variables estadísticas.
Se considera que dos variables cuantitativas están correlacionadas cuando los valores de una de ellas varían sistemáticamente con respecto a los valores homónimos de la otra.

> Si tenemos dos variables (A y B) existe correlación entre ellas, sí al disminuir los valores de A lo hacen también los de B y viceversa.

Importante

La **correlación** entre dos variables **no implica**, por sí misma, ninguna relación de **causalidad**.

El Coeficiente de correlación

El **coeficiente de correlación** de Pearson (R), se usa para cuantificar la fuerza de la relación lineal entre dos variables cuantitativas.
El **coeficiente** R oscila entre $\{-1:1\}$

Características de R

1. Es independiente de cualquier unidad usada para medir las variables.
2. Su valor se altera de forma importante ante la presencia de un valor extremo, como sucede con la desviación típica.
3. Solo establece la relación a una línea recta.
4. El coeficiente de correlación no se debe extrapolar más allá del rango de valores observado de las variables.
5. La correlación no implica causalidad.

Ecuación

Existen diversos coeficientes que miden el grado de correlación, adaptados a la naturaleza de los datos. El más conocido es el *coeficiente de correlación de Pearson*, que se obtiene dividiendo la covarianza de dos variables entre el producto de sus desviaciones estándar.

$$R_{xy} = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{(n-1)s_x s_y} = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{n\sum x_i^2 - (\sum x_i)^2}\ \sqrt{n\sum y_i^2 - (\sum y_i)^2}}$$

Otras formas de cálculo

- Coeficiente de correlación de Spearman
- Correlación de Kendall
- Correlación canónica

Consideraciones

- Las dos variables deben proceder de una muestra aleatoria de individuos.
- Al menos una de las variables debe tener una distribución normal en la población de la cual la muestra procede.

Interpretación de la fuerza

Valor	Grado de relación lineal
$R = 1$	Es lineal perfecta.
$R = \{0.8 : 0.99\}$	Es muy fuerte.
$R = \{0.65 : 0.8\}$	Es fuerte.
$R = \{0.45 : 0.64\}$	Es moderada.
$R = \{0.2 : 0.44\}$	Es débil.
$R = \{0.01 : 0.20\}$	Es muy débil
$R = 0$	No existe asociación lineal

Interpretación del sentido

La **dirección** se indica por el signo (+/-) y puede observarse en la pendiente de la ecuación de la recta o en el gráfico.

Signo	Dirección
$+$	Es directa. Al aumentar una variable la otra aumenta.
$-$	Es inversa. Al aumentar una variable la otra decrece (o viceversa).

El Coeficiente de determinación

Al elevar al cuadrado el coeficiente de correlación se obtiene el **coeficiente de determinación** (R^2).

> Indica el **%** de **variabilidad** de la **variable respuesta** que se explica por la relación con la **variable explicativa**. Utilizar más variables y mediante análisis multivariado permite identificar el efecto de estas variables en la **variable respuesta** (**Bosque & Moreno, 1994**).

El **coeficiente** R^2 oscila entre $\{0:1\}$.

Demostración del R.

A continuación se realiza una demostración simple en R: En este caso son datos de una estación meteorológica con 111 registros de las variables: ozono, radiación solar, temperatura y viento.

- Se cargan las librerías y los datos:

```
library(car)      # funciones para el ajuste de modelos de reg
library(visreg)   # visualiza regresiones
library(lmtest)   # prueba modelos lineales
library(corrplot) # gráfico correlaciones
library(corrgram) # más gráficos de correlaciones

datos <- read.table("ozono.txt", header = TRUE,
                    sep="\t", na.strings="NA", dec=".", strip.
```

Visualización de los datos

```r
scatter.smooth(x=datos$ozono, y = datos$temperatura, main = "O
               lpars = list(col = "red", lwd = 3, lty = 3))
```

Figura 1: Gráfico de dispersión entre las variables

El modelo de regresión lineal

```r
m.reg1 <- lm(temperatura ~ ozono  , data = datos)
summary(m.reg1)
```

```

Call:
lm(formula = temperatura ~ ozono, data = datos)

Residuals:
    Min      1Q  Median      3Q     Max 
-21.980  -4.775   1.825   4.228  12.425 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 69.37059    1.05151   65.97   <2e-16 ***
ozono        0.20006    0.01963   10.19   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.851 on 109 degrees of freedom
```

Gráficos para el análisis de regresión lineal

```
avPlots(m.reg1, id.n = 2 ,  id.cex = 0.7 ) # Grafico de regres
```

Figura 2: Gráfico de regresión

Gráfico de residuales

```r
residualPlots(m.reg1)
```

```
           Test stat Pr(>|Test stat|)    
ozono        -5.8489        5.353e-08 ***
Tukey test   -5.8489        4.948e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figura 3: Gráfico de residuales

Gráficos de diagnóstico

```r
influenceIndexPlot(m.reg1,id=TRUE)
```

Figura 4: Gráfico de diagnóstico

Gráficos de influencia

```r
influencePlot(m.reg1, id=list(method="identify"))
```

Figura 5: Gráfico de influencia

Referencias

Bosque, J., & Moreno, A. (1994). *Prácticas de análisis exploratorio y multivariante de datos* (1.ª ed.). oikos-tau.