Análisis de regresión

19 de agosto
(semana 2)

Plan de trabajo

1. Estudio de pares de variables cuantitativas
2. Uso de R para el modelo ANOVA a una vía
3. Introducción a la regresión lineal simple

Nota:

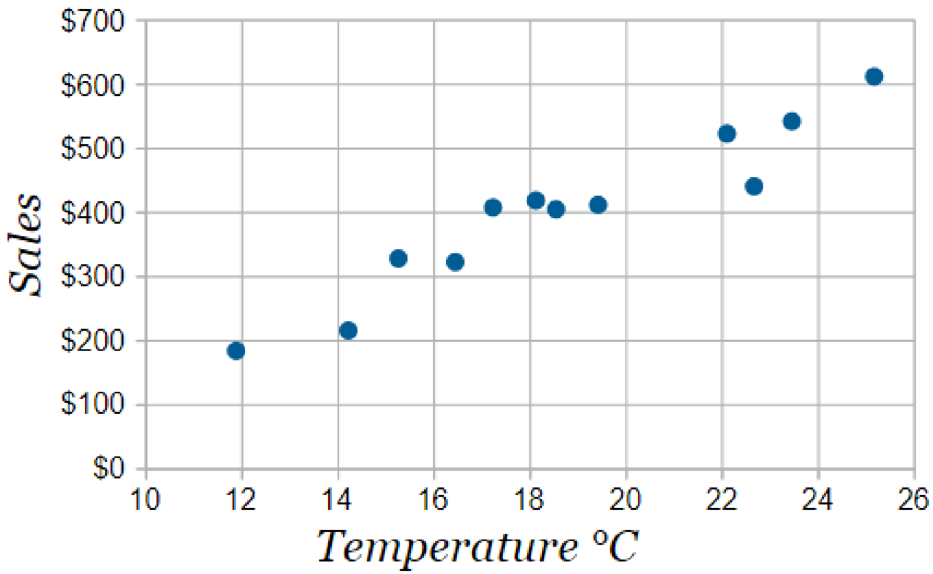
- Ya está disponible el taller 1.
- Este fin de semana restrinjo el Drive a los inscritos y creo el Classroom y el MOODLE.
- La primera entrega del trabajo final será para el domingo 4 de septiembre (ya está en Drive).
- En la semana del 29 de agosto al 2 de septiembre no habrá clase. Pendientes de una actividad a desarrollar.

Relación entre dos variables cuantitativas

Diagrama de dispersión y coeficiente de correlación

- Vimos en estadística descriptiva que el diagrama de dispersión nos permitía visualizar si dos variables cuantitativas tenían o no una relación entre ellas.

Ventas helados



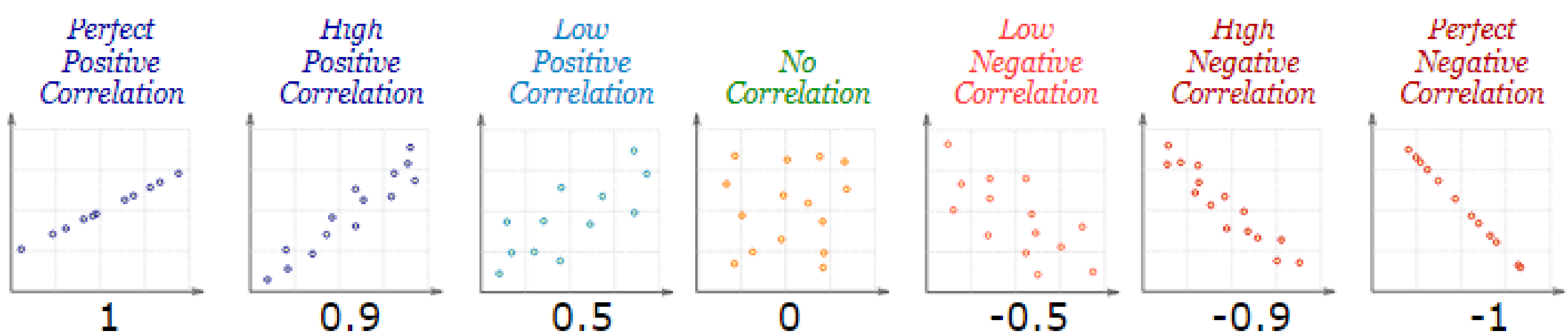
Fuente: https://www.mathsisfun.com/data/scatter-xy-plots.html

- Pero, ¿cómo podemos cuantificar si una relación entre dos variables es fuerte o débil? Para ello, utilizamos el **coeficiente de correlación lineal** de Pearson.
- Recuerden que relación **NO** necesariamente implica causalidad entre las dos variables.

Diagrama de dispersión y coeficiente de correlación

Coeficiente de correlación lineal de Pearson

Es un parámetro que mide qué tan fuerte es la relación lineal entre dos variables. Toma valores entre -1 y 1, donde -1 indica una relación lineal inversa y 1 indica una perfecta relación lineal directa.



Fuente: https://www.mathsisfun.com/data/correlation.html

- ¡Ojo! El coeficiente de correlación solo mide relaciones lineales. Si las dos variables están relacionadas de manera no lineal, es posible que este coeficiente no detecte esa relación.
- http://guessthecorrelation.com/

Otras medidas de asociación

Medidas de asociación en la población

Sea $(X_1,Y_1),(X_2,Y_2),\ldots,(X_n,Y_n)$ una muestra aleatoria de una población bivariada continua (X,Y) con parámetros de asociación:

$$\rho_{X,Y}=\frac{\operatorname{cov}(X,Y)}{\sqrt{Var(X)Var(Y)}},\ \textbf{(Pearson)}$$

$\rho \in [-1,1]$

$$\tau_{X,Y}=p\left[(X_1-X_2)(Y_1-Y_2)>0\right]-p\left[(X_1-X_2)(Y_1-Y_2)<0\right],$$

(Kendall)

$\tau \in [-1,1]$

$$\mathrm{R}_{X,Y}=\frac{\operatorname{cov}(\mathrm{R}(X),\mathrm{R}(Y))}{\sqrt{Var(\mathrm{R}(X))Var(\mathrm{R}(Y))}},\ \textbf{(Spearman)}$$

$R \in [-1,1]$

- El primero solo mide relaciones lineales, los otros relaciones monótonas

Repaso: operador correlación

Teorema: propiedades de la correlación

(1) $\left|\operatorname{cor}\left(X,Y\right)\right| \leq 1;$

(2) $\operatorname{cor}\left(X,Y\right) = \operatorname{cor}\left(Y,X\right);$

(3) $\operatorname{cor}\left(X,X\right) = 1;$

(3) $\operatorname{cor}\left(X,-X\right) = -1;$

(5) $\operatorname{cor}\left(aX+b,Y\right) = \operatorname{cor}\left(X,Y\right)\ \forall a,b \in \mathbb{R},\ a>0;$

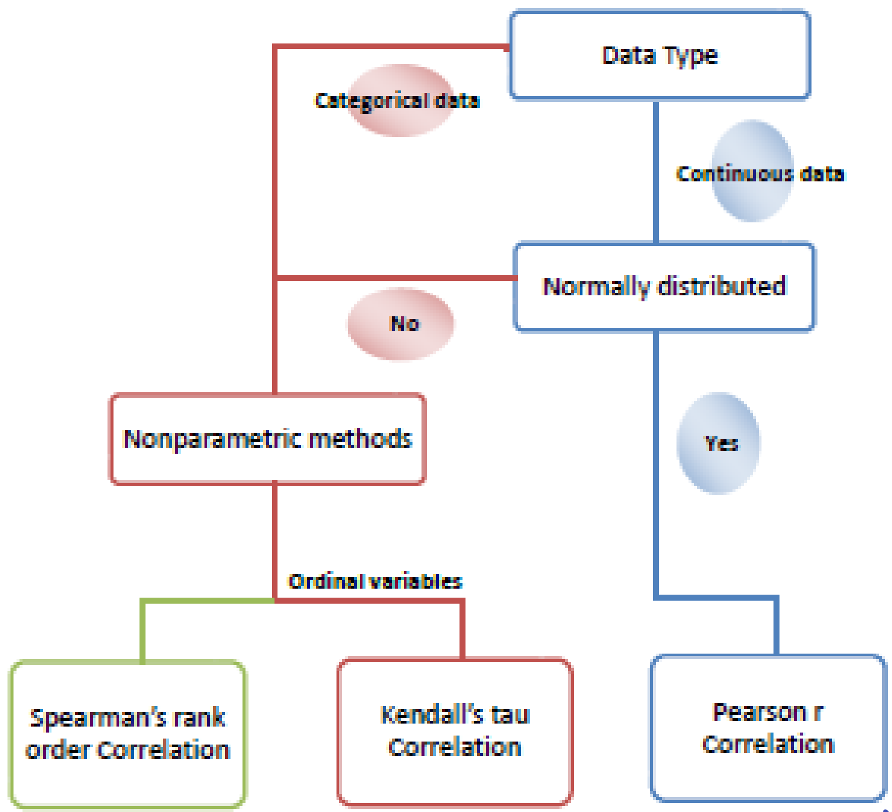
(6) $\left|\operatorname{cor}\left(X,Y\right)\right| = 1$ si y sólo si $\exists\ a,b \in \mathbb{R}$ (no simultáneamente 0) tales que $p\left(aX+bY=0\right)=1.$

(7) ($Independencia \Rightarrow \rho = 0$) Si X,Y son independientes, $\operatorname{cor}\left(X,Y\right)=0.$

- La mayoría de las propiedades se tienen para el coeficiente de Kendall y de Spearman.

Otras medidas de asociación

Making a Decision of the Correlation Methods



Fuente: https://www.originlab.com/doc/Origin-Help/Correlation-Coefficient

- $Independencia \Rightarrow \rho = 0,\ \ \tau = 0,\ \ \mathrm{R} = 0.$
- $\rho = 0$ y normalidad bivariada $\Rightarrow$ *Independencia.*
- $|\rho| = 1 \Rightarrow |\tau| = 1,\ \ |\mathrm{R}| = 1.$

En distrib. Normal multiv.:
① hay relación lineal ($\rho \neq 0$)
② Hay independ. ($\rho = 0$)

Coeficiente de correlación (Bonnett & Wright, 2000)

- **Estimación puntual:**

$$\hat{\rho} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2 \cdot \sum_{i=1}^{n}(y_i - \overline{y})^2}}$$

- **Estimación asintótica por intervalo** (Asumiendo que los datos vienen de una distribución normal)**:**
  - Paso 1. Calcule: $$z_r = \frac{1}{2}\ln\left(\frac{1+\hat{\rho}}{1-\hat{\rho}}\right)$$
  - Paso 2. Identifique el límite inferior (l) y superior (u) mediante la fórmula: $$z_r \mp z_{1-\frac{\alpha}{2}}\sqrt{\frac{1}{n-3}}$$
  - Paso 3. Transforme l y u de vuelta a la escala original del coeficiente $$\rho_L = \frac{\exp(2l)-1}{\exp(2l)+1},\ \rho_U = \frac{\exp(2u)-1}{\exp(2u)+1}$$
- **Pruebas de hipótesis y código R:** `cor.test` (IC e hipótesis =/≠ cero)
  - Usar Bootstrap si no hay normalidad o para otros sistemas

Coeficiente Tau de Kendall (Hogg et al, 2005)

- **Estimación puntual:** $$\hat{\tau} = \frac{2}{n(n-1)} \sum_{i<j} signo(x_i - x_j) signo(y_i - y_j)$$

- **Prueba de hipótesis de no asociación:**

```
cor.test(…,method="kendall")
```

Theorem 10.8.2. *Let $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ be a random sample on the bivariate random vector (X, Y) with continuous cdf $F(x, y)$. Under the null hypothesis of independence between X and Y, i.e., $F(x, y) = F_X(x)F_Y(y)$, for all (x, y) in the support of (X, Y), the test statistic K satisfies the following properties:*

$K = \tau$

$$K \text{ is distribution free with a symmetric pmf} \tag{10.8.4}$$

$$E_{H_0}[K] = 0 \tag{10.8.5}$$

$$Var_{H_0}(K) = \frac{2}{9} \frac{2n+5}{n(n-1)} \tag{10.8.6}$$

$$\frac{K}{\sqrt{Var_{H_0}(K)}} \text{ has an asymptotic } N(0,1) \text{ distribution.} \tag{10.8.7}$$

Coeficiente Rho de Spearman (Bonnett & Wright, 2000)

- **Estimación puntual:** $\hat{r}_S = 1 - \dfrac{6\sum_{i=1}^{n} d_i^2}{n\left(n^2-1\right)}$, con $d_i = \mathrm{R}\left(x_i\right) - \mathrm{R}\left(y_i\right)$

- **Prueba de hipótesis de no asociación:**
  `cor.test(…,method="spearman")`

Theorem 10.8.4. *Let $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ be a random sample on the bivariate random vector (X, Y) with continuous cdf $F(x, y)$. Under the null hypothesis of independence between X and Y, i.e., $F(x, y) = F_X(x)F_Y(y)$, for all (x, y) in the support of (X, Y), the test statistic r_S satisfies the following properties:*

$$r_S \text{ is distribution-free, symmetrically distributed about 0} \tag{10.8.11}$$

$$E_{H_0}[r_S] = 0 \tag{10.8.12}$$

$$Var_{H_0}(r_S) = \frac{1}{n-1} \tag{10.8.13}$$

$$\frac{r_S}{\sqrt{Var_{H_0}(r_S)}} \text{ is asymptotically } N(0,1). \tag{10.8.14}$$

Coeficientes Tau y Rho (Hogg et al, 2005)

- **Intervalos de confianza**
  - Es necesario usar técnicas de Bootstrap. Por ejemplo, para intervalos bilaterales:

```
library(boot); library(npsm)
cor.boot.ci(…,method="spearman")
cor.boot.ci(…,method="kendall")
```

  - Para intervalos unilaterales o pruebas de hipótesis a una cola, deberán escribir su propio código.

Coeficiente Xi (Chatterjee, 2021)

- **Parámetro:** Coeficiente Xi de dependencia funcional de Y en función de X, en la población objetivo.

Let (X, Y) be a pair of random variables, where Y is not a constant. Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be i.i.d. pairs with the same law as (X, Y), where $n \geq 2$. The new coefficient has a simpler formula if the X_i's and the Y_i's have no ties. This simpler formula is presented first, and then the general case is given. Suppose that the X_i's and the Y_i's have no ties. Rearrange the data as $(X_{(1)}, Y_{(1)}), \ldots, (X_{(n)}, Y_{(n)})$ such that $X_{(1)} \leq \cdots \leq X_{(n)}$. Since the X_i's have no ties, there is a unique way of doing this. Let r_i be the rank of $Y_{(i)}$, that is, the number of j such that $Y_{(j)} \leq Y_{(i)}$. The new correlation coefficient is defined as

$$\xi_n(X, Y) := 1 - \frac{3 \sum_{i=1}^{n-1} |r_{i+1} - r_i|}{n^2 - 1}. \tag{1.1}$$

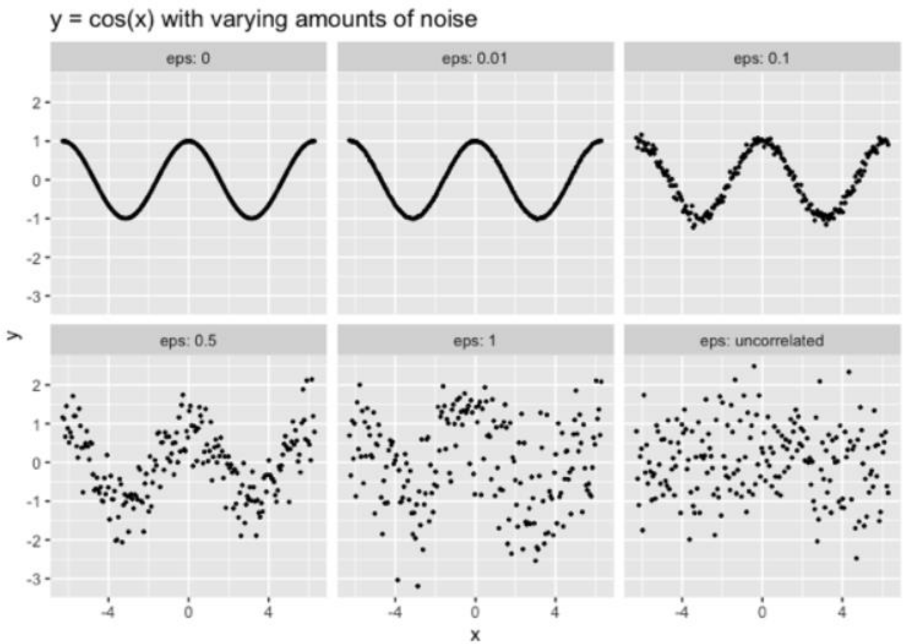
No relac func

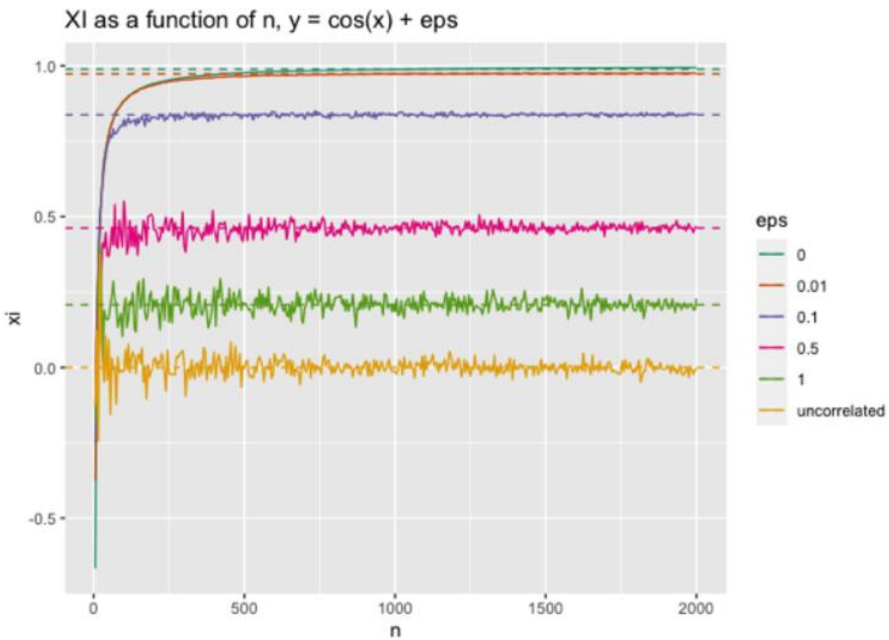
$\xi \in [0, 1]$

Relac Func perf

- No es simétrico. Si su valor es cercano a 1 indica relación funcional entre las variables. Si es 0, da evidencia de independencia.
- Requiere tamaños de muestra grandes para ser concluyente.
- Es más potente cuando la relación funcional es suave y no monótona

Coeficiente Xi (Chatterjee, 2021) (II)





Theorem 2.1. *Suppose that X and Y are independent and Y is continuous. Then $\sqrt{n}\xi_n(X,Y) \to N(0,2/5)$ in distribution as $n \to \infty$.*

- **Código en R:** `calculateXI` o `xicor` (librería `XICOR`)

Implementación de un modelo $\boxed{\text{ANOVA}}$ en R

$\to$ Relac var cualit y cuantit

1. Identificación

gráficos por cada categoría $\begin{cases} \text{Boxplot} \\ \text{Histogramas} \end{cases}$

$\Downarrow$

- Relación que afecte únicamente la tendencia central
- Se mantienen iguales:
  - Variabilidad
  - Simetría
  - Curtosis

ANOVA

$$Y_{ij} = \mu + \underbrace{\tau_i}_{\text{Cambio medio en el grupo } i} + \varepsilon_{ij} \overset{iid}{\sim} N(0, \sigma^2)$$

2. Estimación e inferencia

Prueba F

$$\begin{cases} H_0: \tau_1 = \tau_2 = \dots = \tau_k = 0 \\ H_1: \exists\, \tau_i \neq 0 \end{cases}$$

$$① \quad \underbrace{\sum_i \sum_j (Y_{ij} - \bar{Y}_{..})^2}_{SCTot} = \underbrace{\sum_i \sum_j (Y_{ij} - \bar{Y}_{i.})^2}_{SCerror} + \underbrace{\sum_i n_i (\bar{Y}_{i.} - \bar{Y}_{..})^2}_{SCtrat}$$

	SCerror	SCtrat
Si Ho es cierta:	Grande	Pequeño
Si Ho es falsa:	Pequeño	grande

$$F = \frac{SCTrat/k-1}{SCerror/N-k} \overset{H_0}{\sim} F(k-1, N-k)$$

τ: "Rechazar Ho si $f_c > f_{1-\alpha}(k-1, N-k)$"