Minería de Datos para el Análisis de Big Data

Por: Carlos Carreño

ccarrenovi@Gmail.com

Abril,2021

Modulo 5 Introducción al Modelamiento Estadístico

- Prueba de Hipótesis
- El valor p (p-value)
- Covarianza
- Coeficiente de Correlación de Pearson
- Modelo de Regresión Lineal Simple

Prueba o Contraste de Hipótesis

- Es una prueba estadística que se hace en los parámetros de interés como la media poblacional, distribución de probabilidad, la varianza.
- se plantean dos hipótesis mutuamente excluyentes; la **hipótesis nula Ho** y la **hipótesis alternativa H1 (Hipótesis de estudio o de investigación)**
- Es un procedimiento basado en la evidencia muestral y en la teoría de probabilidad que se emplea para determinar si la hipótesis es un enunciado racional y no debe rechazarse o si es irracional y debe ser rechazada.

- Se conoce que la media poblacional es $\mu = 48$
- Se toma una muestra de *N*=80 elementos y se obtiene que $\bar{X} = 52$
- Con estos resultados muéstrales, podemos plantear las siguientes hipótesis:

$$\mu \neq 48$$

$$\mu > 48$$

Prueba de Hipótesis de una Sola Cola y Dos Colas

- La prueba de hipótesis de una sola cola se utiliza cuando en la hipótesis alternativa plateamos que el parámetro es mayor o menor aun valor.
- Cuando en la hipótesis alternativa plateamos que el parámetro es diferente a un valor; utilizamos la prueba de hipótesis de dos colas

Ejemplo: Prueba de Hipótesis Caso 1

- Una cadena de autoservicios, genera su propia tarjeta de crédito, el gerente desea averiguar si el saldo medio adeudado por los clientes es mayor a usd$400, el nivel de significación es de 5%, una revisión aleatoria de 172 saldos deudores revelo que la media muestral es usd$407 y la desviación estándar usd$38. ¿Debería concluir que la media poblacional es mayor que usd$400 o es razonable suponer que la diferencia de usd7 se debe al azar de la selección?

¿Qué tipo de prueba de hipótesis aplicarías de una cola o dos colas?

Primer Caso : Prueba de hipótesis de una sola cola

- El caso es de una sola cola.

n= 172
$\bar{X}$ = $407
S = $38
μ = 400
α = 5%

- 1) Planteamiento de Hipótesis

$$H_0: \mu = \$400$$
$$H_1: \mu > \$400$$

...

- 2) Nivel de significación y valor critico (Z)

$$\alpha = 5\%$$

$$f(x) = 0.50 + 0.45$$

$$Z = 1.6448$$

Nota: Para hallar Z puedes usar la tabla de distribución normal o la función **qnorm(0.95)** en R.

```
> qnorm(0.95)
[1] 1.644854
```

- 3) Identificar el estadístico de prueba

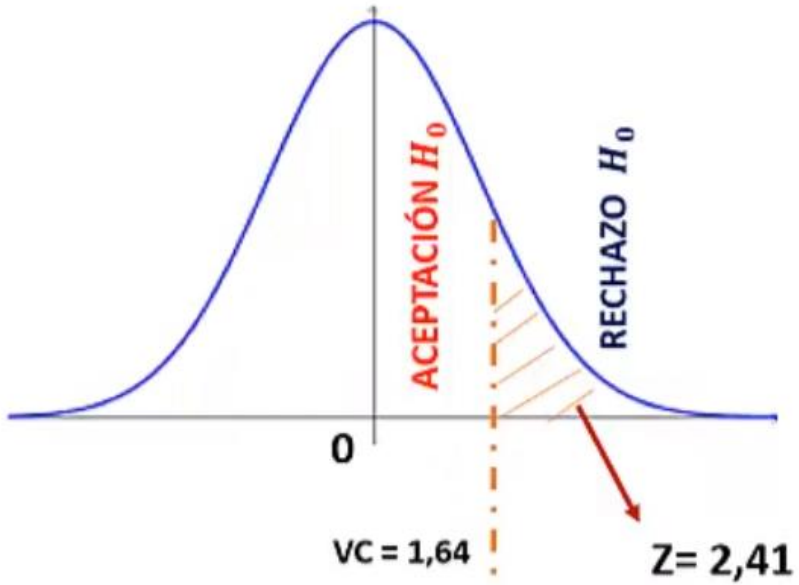
$$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} \qquad z = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$

Nota: La muestra es grande, podemos decir que sigma es cercano a S

$$z = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}} \Rightarrow z = \frac{407 - 400}{\frac{38}{\sqrt{172}}} = 2,41$$

...

- 4) Formular la regla de decisión



...

- 5) Tomar la decisión
- Como el valor critico (Z) esta en el área de rechazo; rechazamos la hipótesis nula Ho y aceptamos la hipótesis alternativa, esto es; existe una alta probabilidad que la media poblacional sea mayor a usd$400

$$H_0: \mu = \$400$$
$$H_1: \mu > \$400$$

Ejemplo: Prueba de Hipótesis Caso 2

- Una cadena de autoservicios, genera su propia tarjeta de crédito, el gerente desea averiguar si el saldo medio adeudado por los clientes es mayor a usd$400, el nivel de significación es de 5%, una revisión aleatoria de 172 saldos deudores revelo que la media muestral es usd$407 y la desviación estándar usd$38. ¿Debería concluir que la media poblacional no es usd$400 o es razonable suponer que la diferencia de usd7 se debe al azar de la selección?

¿Qué tipo de prueba de hipótesis aplicarías de una cola o dos colas?

Primer Caso : Prueba de hipótesis de dos colas

- El caso es de dos colas.

n= 172
$\bar{X}$ = $407
S = $38
μ = 400
α = 5%

- 1) Planteamiento de Hipótesis

$$H_0: \mu = \$400$$
$$H_1: \mu \neq \$400$$

...

- 2) Nivel de significación y valor critico (Z)

$$\alpha = 5\%$$

$$f(x) = (1 - 0.05)/2 = 0.475$$

$$Z = 1.96$$

Nota: Para hallar Z puedes usar la tabla de distribución normal o la función **qnorm(0.975)** en R.

```
> qnorm(0.5 + (1-0.05)/2 )
[1] 1.959964
```

- 3) Identificar el estadístico de prueba

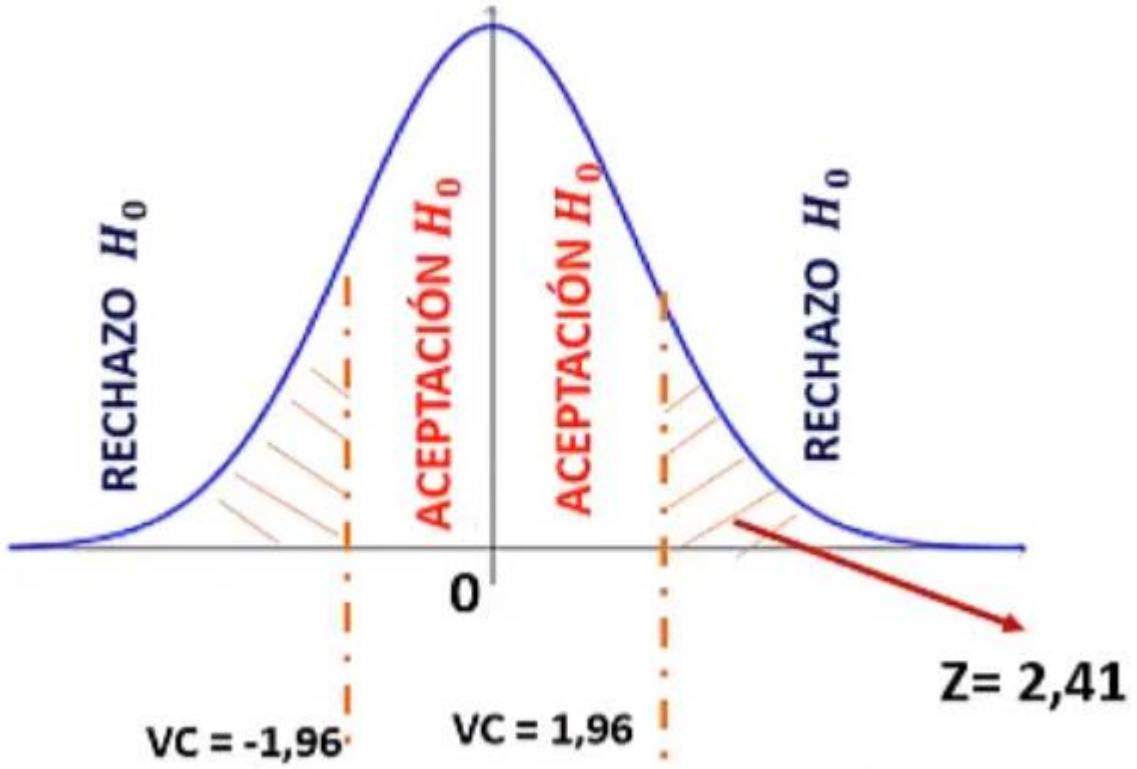
$$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} \qquad z = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$

Nota: La muestra es grande, podemos decir que sigma es cercano a S

$$z = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}} \Rightarrow z = \frac{407 - 400}{\frac{38}{\sqrt{172}}} = 2,41$$

...

- 4) Formular la regla de decisión



...

- 5) Tomar la decisión
- Como el valor critico (Z) esta en el área de rechazo; rechazamos la hipótesis nula Ho y aceptamos la hipótesis alternativa, esto es; existe una alta probabilidad que la media poblacional sea diferente de usd$400

$$H_0: \mu = \$400$$
$$H_1: \mu \neq \$400$$

¿Cual seria el intervalo de confianza para la media poblacional?

Errores estadísticos en las pruebas de hipótesis

- Al rechazar la hipótesis nula lo hacemos con una probabilidad de equivocarnos (menor de α o con el valor p) que no elimina la posibilidad de equivocarnos.
- Cuando se rechaza la H0, pero esta es correcta, entonces cometemos un error Tipo I. Este tipo de error es menor cuando el valor (α) es menor.
- ¿Por qué no llevar el valor de α a un mínimo cercano a 0? Al disminuir la probabilidad de equivocarnos al rechazar H0 estamos, por otro lado, aumentando la probabilidad de aceptar esta hipótesis, siendo realmente incorrecta; a este error se le llama error Tipo II (β).

- Tipos de error al rechazar la hipótesis nula Ho

	H_0 True	H_0 False
Reject H_0	Type I Error	Correct Rejection
Fail to Reject H_0	Correct Decision	Type II Error

El valor-p (p-value)

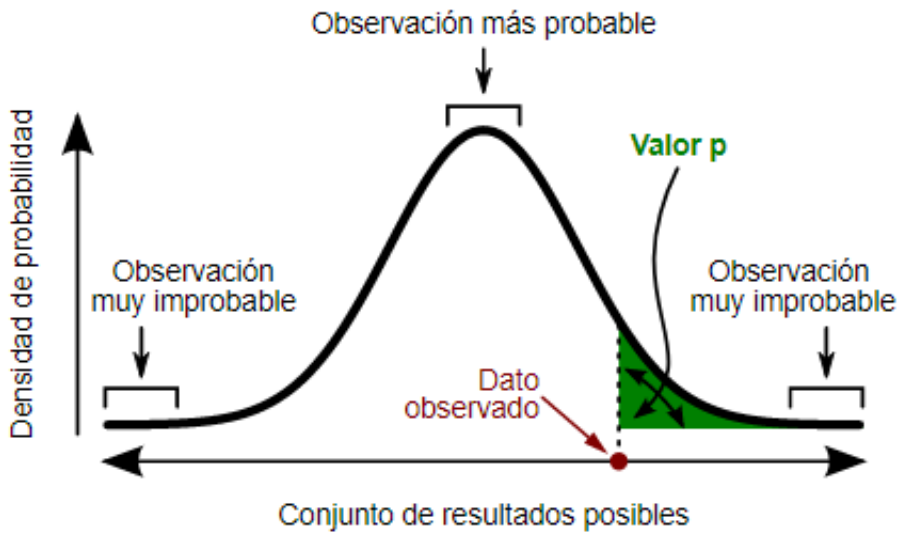
- En estadística general y contrastes de hipótesis, el **valor *p*** (en inglés *p-value*) se define como la probabilidad de que un valor estadístico calculado sea posible dada una hipótesis nula cierta.
- El valor *p* ayuda a diferenciar resultados que son producto del azar del muestreo, de resultados que son estadísticamente significativos.
- Si el valor *p* cumple con la condición de ser menor que un nivel de significancia impuesto arbitrariamente, este se considera como un resultado estadísticamente significativo y, por lo tanto, permite rechazar la hipótesis nula.

$$\text{valor p} = \text{Probabilidad}(\text{resultado tan extremo o más} \mid \text{hipótesis nula}) = \mathbb{P}(\text{resultado tan extremo o más} \mid H_0)$$

Interpretación del valor p

- La probabilidad de observar un resultado dada una cierta hipótesis cierta no es equivalente a la probabilidad de que una hipótesis sea cierta dado un resultado observado

Pr (observación | hipótesis) ≠ Pr (hipótesis | observación)



El **valor p** (área de color verde) es la probabilidad de que un valor observado sea igual o más extremo que un cierto valor, asumiento que la hipótesis nula es cierta.

…

- El valor *p* es un valor de probabilidad, por lo que oscila entre 0 y 1. El valor *p* nos muestra la probabilidad de haber obtenido el resultado que hemos obtenido suponiendo que la hipótesis nula H_0 es cierta. Se suele decir que valores altos de *p* no permiten rechazar la $H_{0,}$ mientras que valores bajos de *p* sí permiten rechazar la H_0.

...

- R.A. Fisher planteo que 1/20=0.05 representa un suceso inusual (muy raro).

Ronald Aylmer Fisher (1890-1962)

Estadístico y biólogo británico

STATISTICAL METHODS FOR RESEARCH WORKERS (1925)

"una de cada veinte (1/20=0.05) oportunidades representa un suceso muestral inusual"

p-valor < 0.05

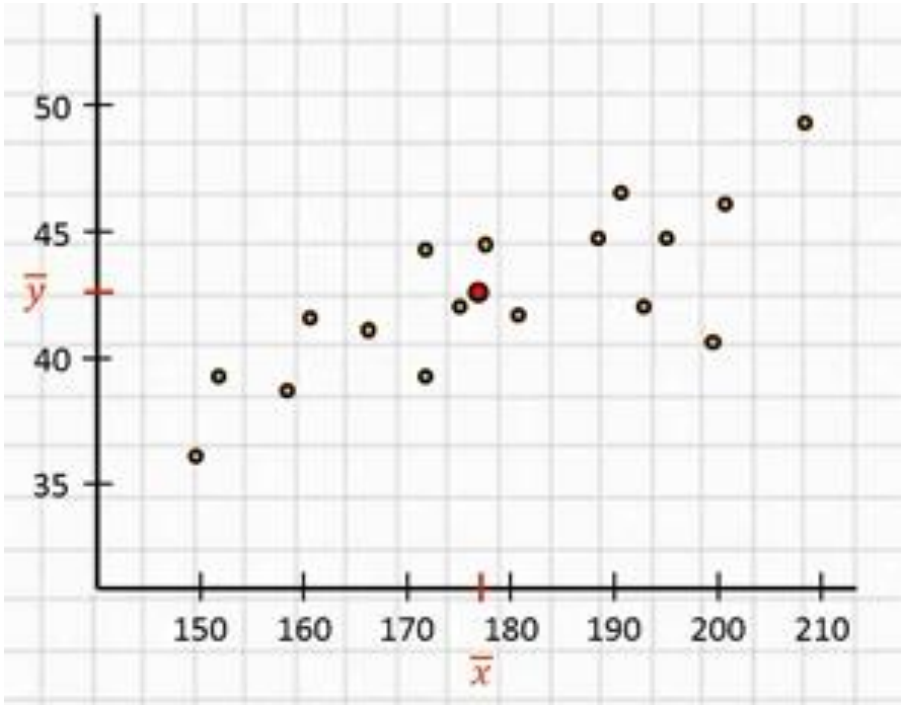
Si el ***p-value < 0.05***, podemos científicamente rechazar la Ho, recuerda que p-value es una probabilidad.

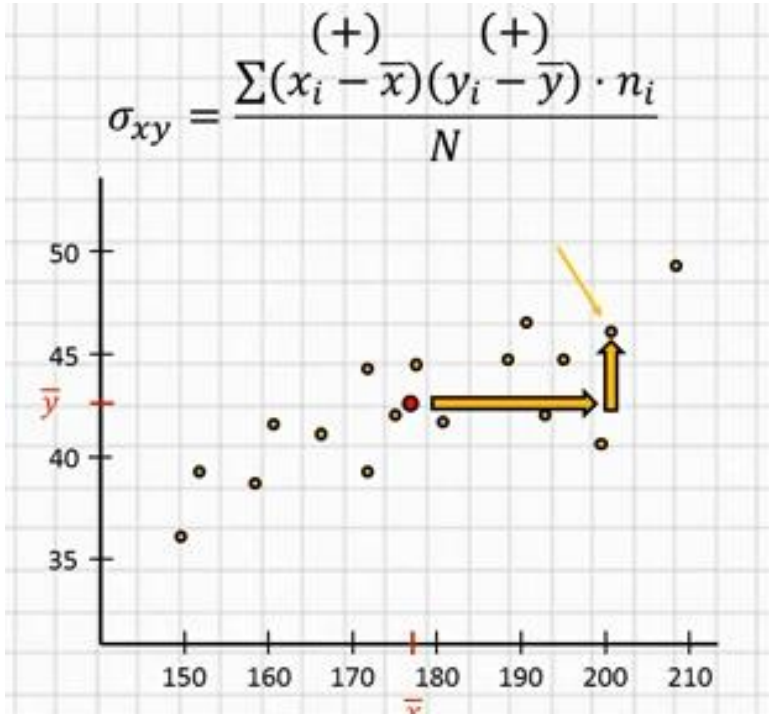
Covarianza

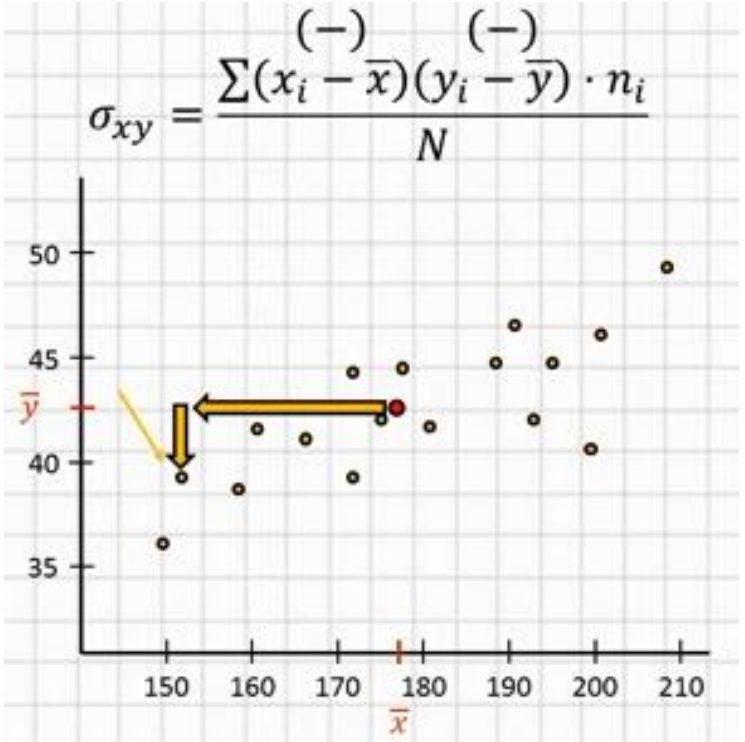
- Significado de la covarianza

$$\sigma_{xy} = \frac{\sum(x_i - \overline{x})(y_i - \overline{y}) \cdot n_i}{N}$$

ni =frecuencia de la observación y N = numero de sucesos de la muestra

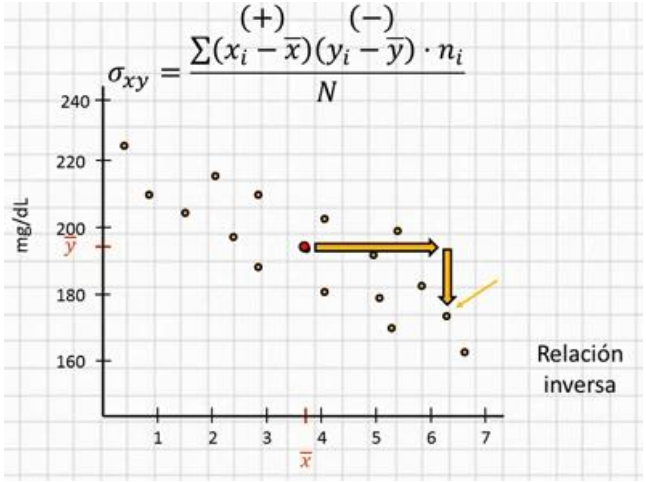


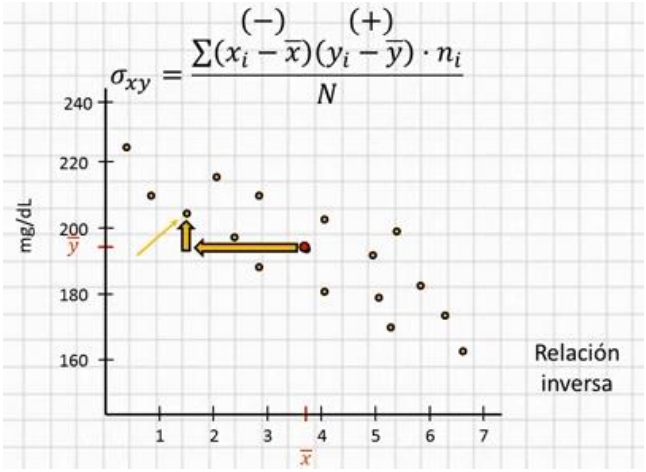




Covarianza

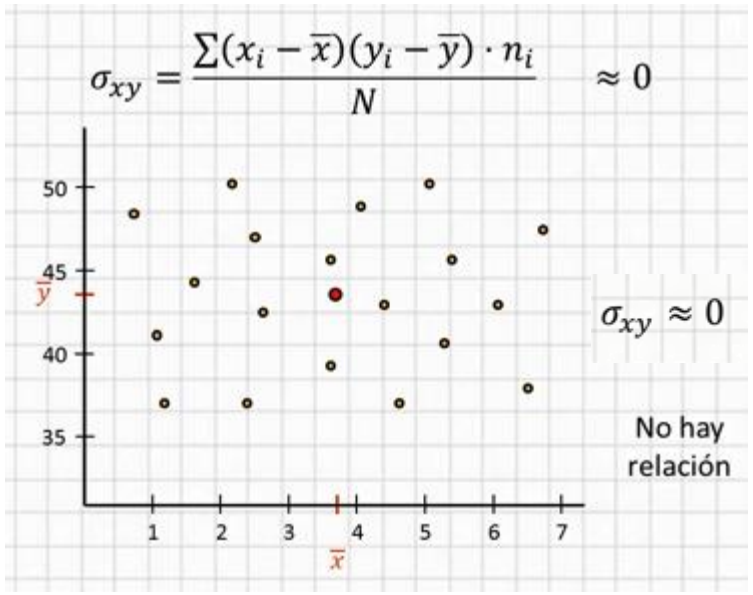
- Cuando entre las dos variables hay una relación directa, la covarianza da un valor positivo.
- Cuando entre las dos variables hay una relación inversa, la covarianza da un valor negativo





Covarianza

- Cuando entre las dos variables no hay una relación, la covarianza da un valor cercano a CERO



Coeficiente de Correlación de Pearson

- Formula del coeficiente de correlación de Pearson

$$r = \frac{\sigma_{xy}}{\sigma_x \cdot \sigma_y}$$

σ_{xy}: covarianza

σ_x: desviación típica de x

σ_y: desviación típica de y

→ r es un coeficiente (no tiene unidades)

→ $-1 \leq r \leq 1$

Si $r \approx 1$ existe una correlación directa fuerte

Si $r \approx -1$ existe una correlación inversa fuerte

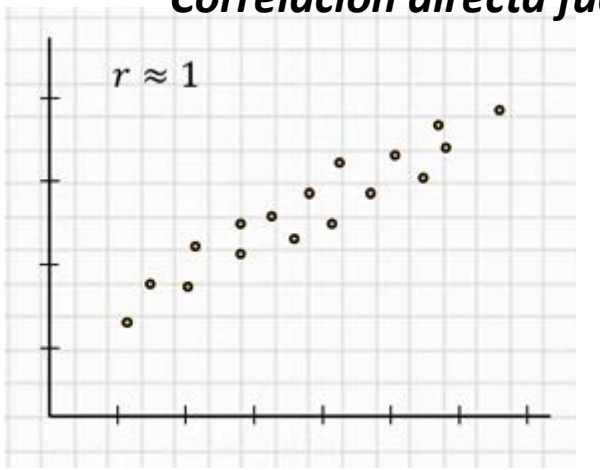
Si $r = 1$ o $r = -1$ hay una correlación funcional

Si $r \approx 0$ no existe una correlación lineal

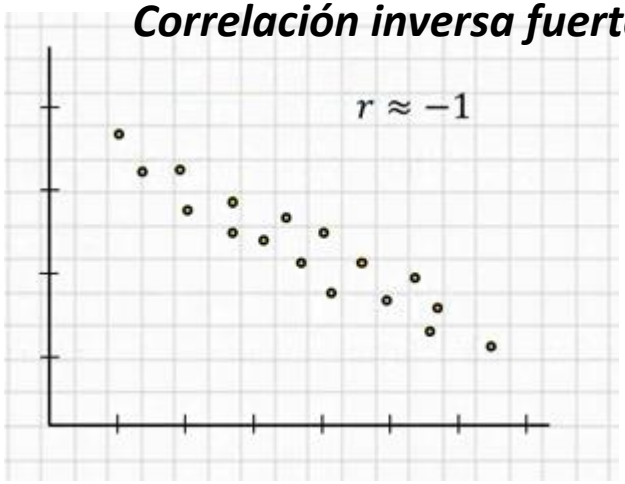
Coeficiente de Correlación de Pearson

- Coeficiente de correlación de Pearson y la relación entre variables

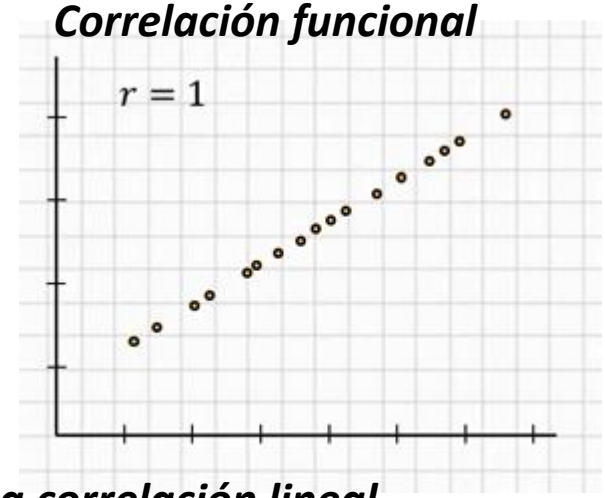
Correlación directa fuerte



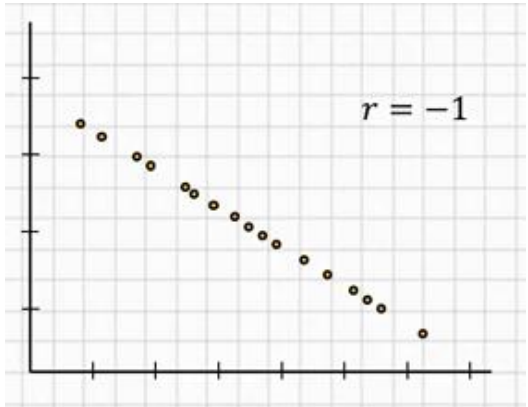
Correlación inversa fuerte



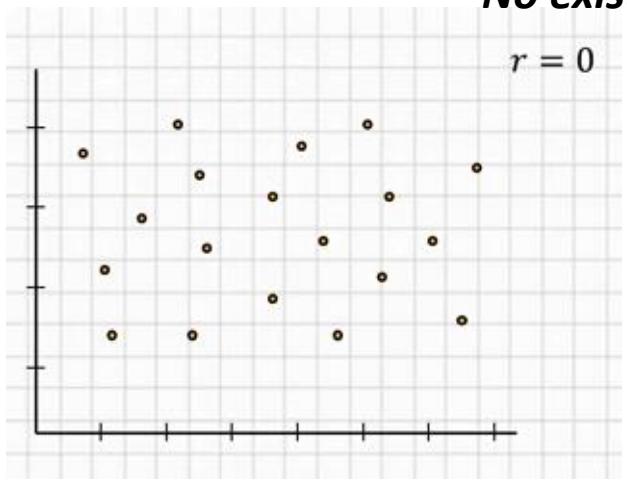
Correlación funcional

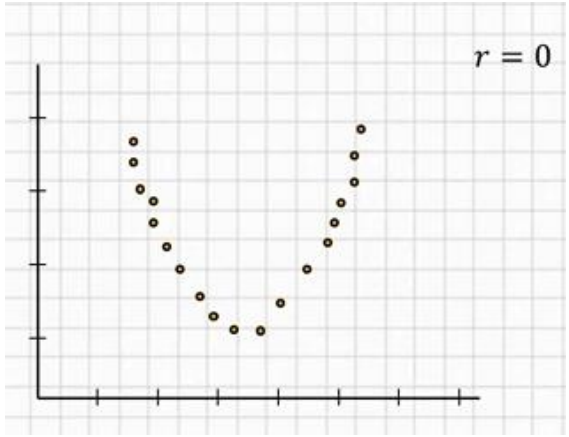


Correlación funcional



No existe una correlación lineal





Ejemplo: Calculo de Coeficientes de Correlación de Pearson con R

RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function | Addins | Project: (None)

Untitled1*

Source on Save | Run | Source

```
1  datos = read.table("c:\\Data\\grasas.txt", header=TRUE)
2  names(datos)
3  pairs(datos)
4  #Calcula los coeficientes de correlacion de pearson entre las variables
5  cor(datos)
6
7
```

4:72 (Top Level) R Script

Environment History Connections Tutorial

Import Dataset | List

R | Global Environment

Data

datos	25 obs. of 3 variables

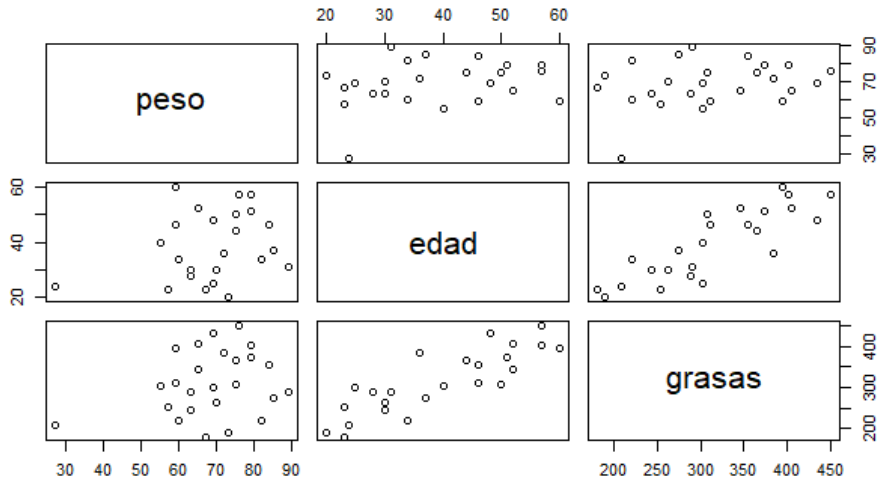
Console Terminal Jobs

~/

```
> names(datos)
[1] "peso"   "edad"   "grasas"
> pairs(datos)
> cor(datos)
            peso      edad    grasas
peso   1.0000000 0.2400133 0.2652935
edad   0.2400133 1.0000000 0.8373534
grasas 0.2652935 0.8373534 1.0000000
> class(datos)
[1] "data.frame"
>
```

Files Plots Packages Help Viewer

Zoom | Export | Publish



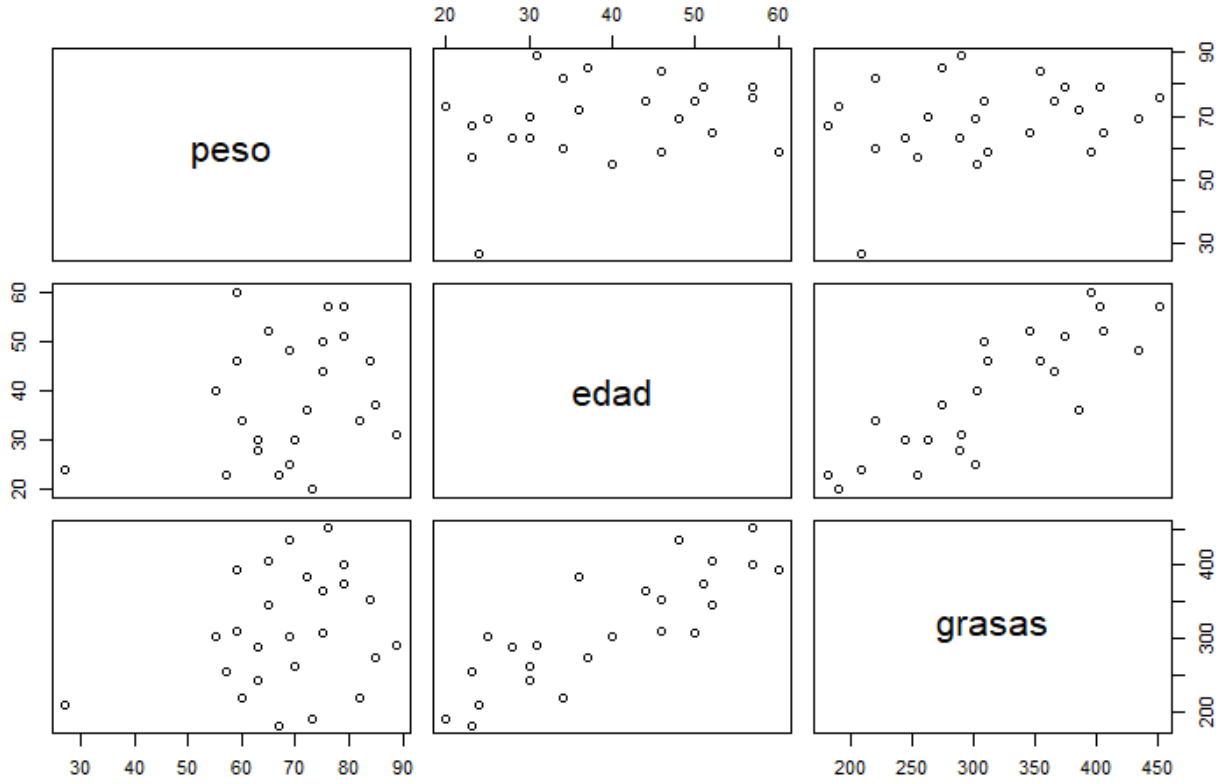
Modelo de Regresión Lineal Simple

- El objetivo de un modelo de regresión es tratar de explicar la relación que existe entre una variable dependiente (variable respuesta) Y un conjunto de variables independientes (variables explicativas) X1,..., Xn.
- En un modelo de regresión lineal simple tratamos de explicar la relación que existe entre la variable respuesta Y y una única variable explicativa X.

Ejemplo: Grasas vs Edad

- Podemos sospechar que existe una relación directa fuerte entre la cantidad acumulada de grasa y la edad de una persona.

```
> pairs(datos)
```

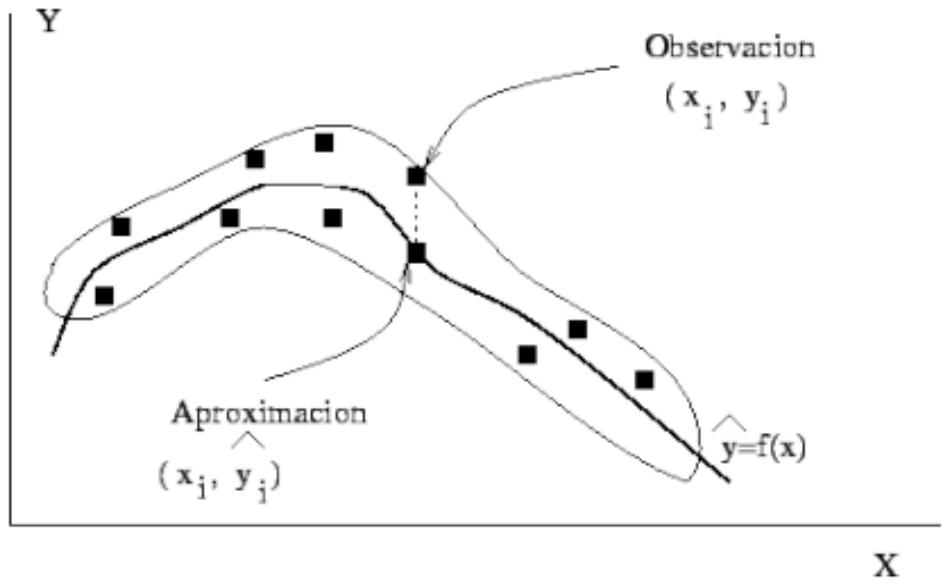


	peso	edad	grasas
1	84	46	354
2	73	20	190
3	65	52	405
4	70	30	263
5	76	57	451
6	69	25	302
7	63	28	288
8	72	36	385
9	79	57	402
10	75	44	365
11	27	24	209
12	89	31	290
13	65	52	346
14	57	23	254

...

- Mediante las técnicas de regresión de una variable Y sobre una variable X, buscamos una función que sea una buena aproximación de una nube de puntos (xi,yi), mediante una curva del tipo:

$$\hat{Y} = f(X)$$



- El modelo de regresión lineal simple tiene la siguiente expresión:

$$Y = \alpha + \beta X + \varepsilon$$

- En donde α es la ordenada en el origen (el valor que toma Y cuando X vale 0), β es la pendiente de la recta (e indica cómo cambia Y al incrementar X en una unidad) y ε una variable que incluye un conjunto grande de factores, cada uno de los cuales influye en la respuesta sólo en pequeña magnitud, a la que llamaremos error

Método de Mínimos Cuadrados

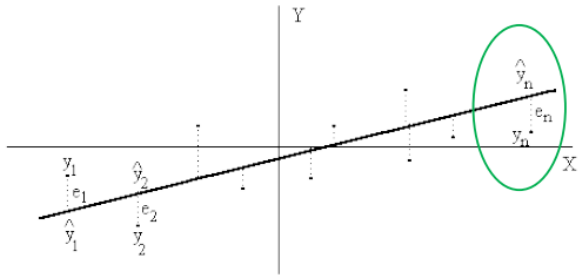
- Para hacer una estimación del modelo de regresión lineal simple, trataremos de buscar una recta de la forma:

$$\hat{Y} = \hat{\alpha} + \hat{\beta}X = a + bX$$

- Para esto utilizaremos el método de mínimos cuadrados. Este método consiste en minimizarla suma de los cuadrados de los errores:

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

- Es decir, la suma de los cuadrados de las diferencias entre los valores reales observados y_i y los valores estimados ($\widehat{y_i}$) .



- Con este método, las expresiones que se obtiene para a y b son las siguientes:

$$a = \overline{y} - b\overline{x}, \quad b = \frac{S_{XY}}{S_X^2}$$

Nota:

En donde $\overline{x}$ e $\overline{y}$ denotan las medias muestrales de *X* e *Y* (respectivamente), S_X^2 es la varianza muestral de *X* y S_{XY} es la covarianza muestral entre *X* e *Y*.

...

- Estos parámetros se calculan como:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}, \quad \bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}, \quad S_X^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n}, \quad S_Y^2 = \frac{\sum_{i=1}^{n} (y_i - \bar{y})^2}{n}, \quad S_{XY} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{n}.$$

- La cantidad b se denomina coeficiente de regresión de Y sobre X se denota por $b_{Y/X}$

$$a = \bar{y} - b\bar{x}, \quad b = \frac{S_{XY}}{S_X^2}$$

$$\tilde{y} = a + bX$$

Coeficiente de Regresión

- El coeficiente de regresión nos da información sobre el comportamiento de la variable Y frente a la variable X, de manera que:

a) Si $b_{Y/X} = 0$, para cualquier valor de *X* la variable *Y* es constante (es decir, no cambia).

b) Si $b_{Y/X} > 0$, esto nos indica que al aumentar el valor de *X*, también aumenta el valor de *Y*.

c) Si $b_{Y/X} < 0$, esto nos indica que al aumentar el valor de *X*, el valor de *Y* disminuye.

Componentes de la Variabilidad

- La variabilidad las podemos descomponer en las siguientes partes:

$$\sum (y_i - \bar{y})^2 = \sum (\hat{y}_i - \bar{y})^2 + \sum (y_i - \hat{y}_i)^2 \quad \Leftrightarrow$$
$$SCtot \quad = \quad SCR \quad + \quad SCres$$

Coeficiente de Determinación r^2

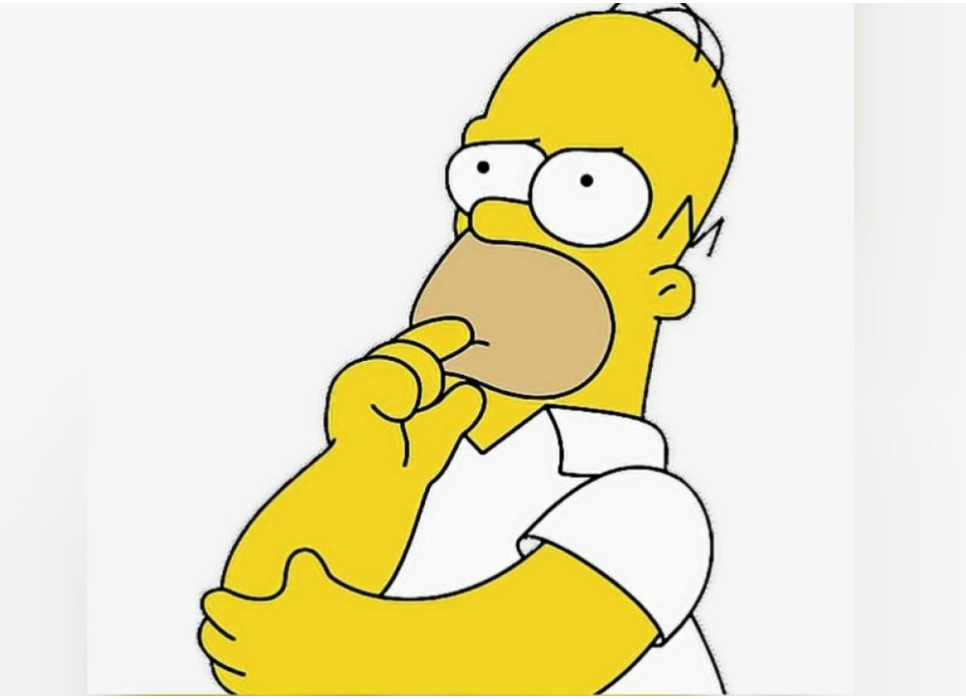
- El coeficiente de determinación esta dado por la siguiente formula

$$r^2 = \frac{\sum(\hat{y}_i - \overline{y})^2}{\sum(y_i - \overline{y})^2} = \frac{SCR}{SCtot}$$

- El coeficiente de determinación r^2 puede interpretarse como la proporción de variabilidad de Y que es explicada por X. Mide la proximidad de la recta ajustada a los valores observados de Y.

Preguntas

- Alguna pregunta?



Demo

- Crear y validar un modelo predictivo basado en regresión lineal simple para estudiar el caso de acumulación de grasa en una persona respecto a la edad de la misma.