Análisis e Interpretación de Datos

MÁSTER UNIVERSITARIO EN ANÁLISIS Y VISUALIZACIÓN DE DATOS MASIVOS / VISUAL ANALYTICS AND BIG DATA

Miller Janny Ariza Garzón

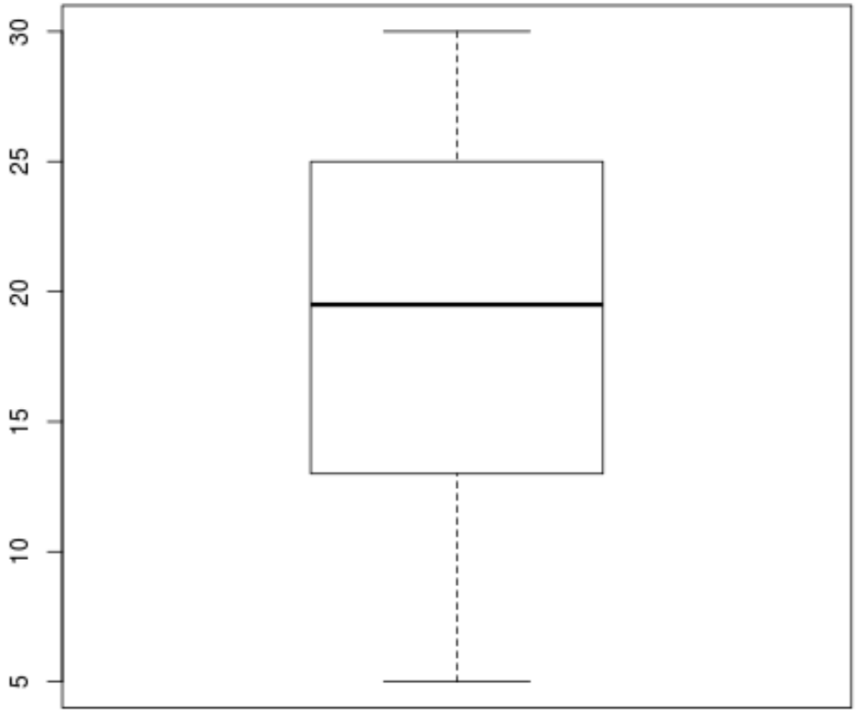
Tema 4. Regresión y correlación I

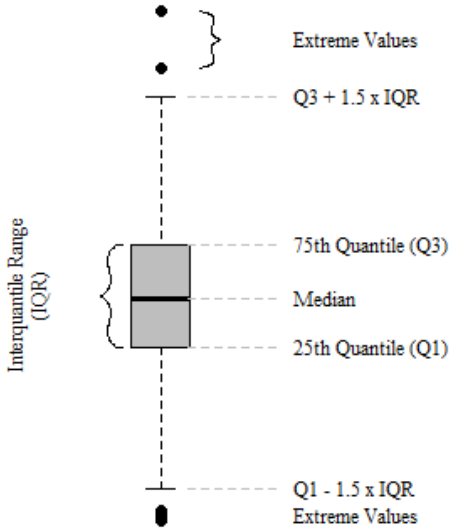
unir LA UNIVERSIDAD EN INTERNET

Universidad Internacional de La Rioja

Repaso y temas pendientes

Boxplot:



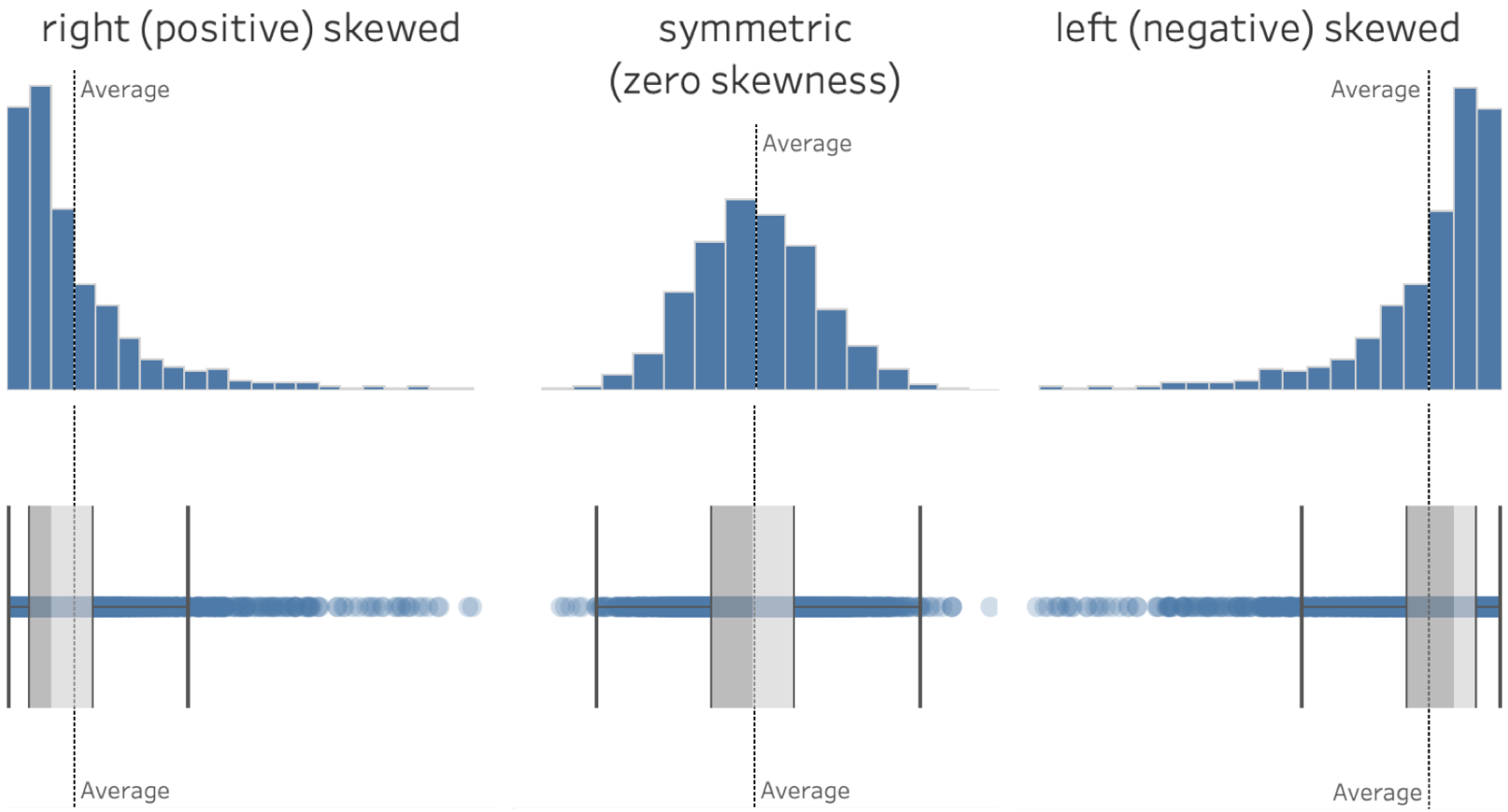


Usos:

- Dispersión
- Valores atípicos
- Comparar distribuciones
- Sesgo o asimetría

Repaso y temas pendientes

Boxplot:



https://campus.datacamp.com/courses/statistical-techniques-in-tableau/univariate-exploratory-data-analysis?ex=9

Repaso y temas pendientes

Curtosis en R:

```
x<-c(1,2,3,4,5)
e1071::kurtosis(x, type=1) # (criterio es 0)
[1] -1.3
```

type=1, This is the typical definition used in many older textbooks.
type=2, Used in SAS and SPSS.
type=3, Used in MINITAB.

```
moments::kurtosis(x) # Pearson's measure of kurtosis (criterio 3)
moments::kurtosis(x)-3 (criterio es 0)
[1] -1.3
```

Tabla de contenido

Tabla de contenido

Asociación conceptual

Y

Variable Explicada
Variable dependiente
Variable predicha
Variable respuesta
Variable objetivo
Variable endogena

X

Variable explicativa
Variable independiente
Variable predictora
Variable entrada
Variable exogena

- volumen de alcohol en sangre
- calificación

- número de cervezas consumidas
- número de horas de estudio

Introducción

La relación estadística entre dos variables se puede identificar a través de tablas, gráficos o medidas.

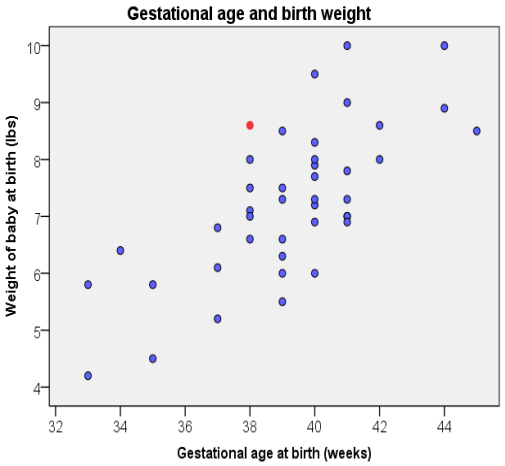
Tablas de frecuencias

Y / X	y_1	y_2	$\cdots$	y_j	$\cdots$	y_p	
x_1	n_{11}	n_{12}	$\cdots$	n_{1j}	$\cdots$	n_{1p}	$n_{1\bullet}$
x_2	n_{21}	n_{22}	$\cdots$	n_{2j}	$\cdots$	n_{2p}	$n_{2\bullet}$
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
x_i	n_{i1}	n_{i2}	$\cdots$	n_{ij}	$\cdots$	n_{ip}	$n_{i\bullet}$
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
x_k	n_{k1}	n_{k2}	$\cdots$	n_{kj}	$\cdots$	n_{kp}	$n_{k\bullet}$
	$n_{\bullet 1}$	$n_{\bullet 2}$	$\cdots$	$n_{\bullet j}$	$\cdots$	$n_{\bullet p}$	$n_{\bullet\bullet}$

Frecuencias absolutas conjuntas

Frecuencias absolutas marginales

Gráficos de dispersión



Medidas

Covarianza,
Correlaciónes,
….

Tablas de frecuencias

ALTURA (X)	Peso (Y)	ALTURA (X)	Peso (Y)
171	110,5	183	66,5
158	79	155	77,5
185	67,5	192	71
172	111	159	79,5
167	108,5	189	94,5
164	107	182	91
178	89	161	105,5
166	58	195	97,5
177	63,5	180	90
184	92	162	56
188	119	176	88
178	64	175	87,5
161	55,5	191	70,5
187	93,5	155	77,5

Para variables cuantitativas categorizadas.

	Peso (Y)					Totales Altura
Altura (X)	55-68	68-81	81-94	94-107	107-120	
155-165	2	4	1	0	1	8
165-175	1	0	0	0	3	4
175-185	3	6	0	0	0	9
185-195	1	2	1	2	1	7
Totales Peso	7	12	2	2	5	**28**

Para variables cualitativas.

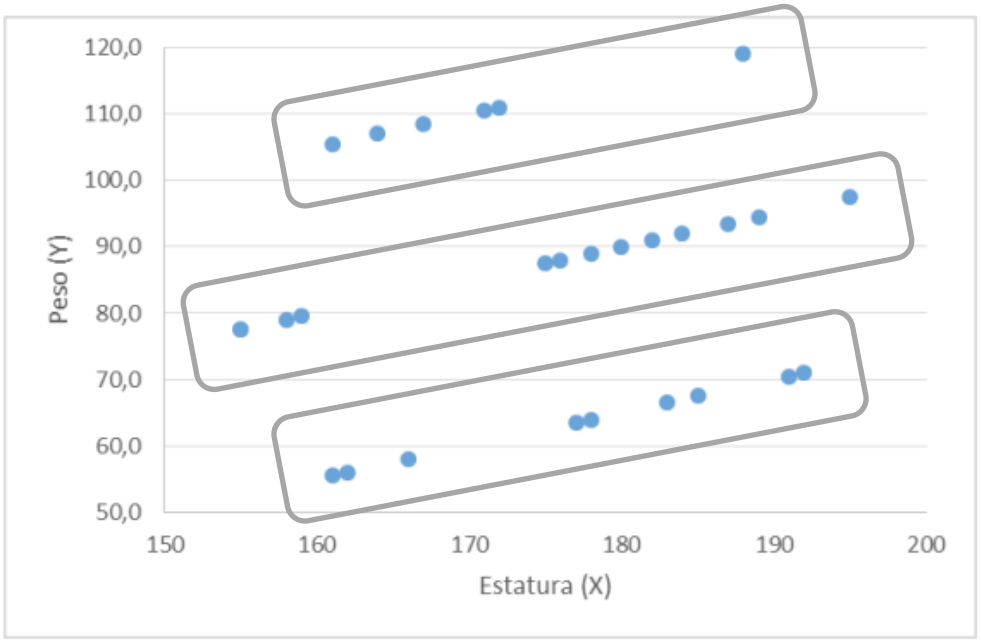
Class * Survived? Crosstabulation

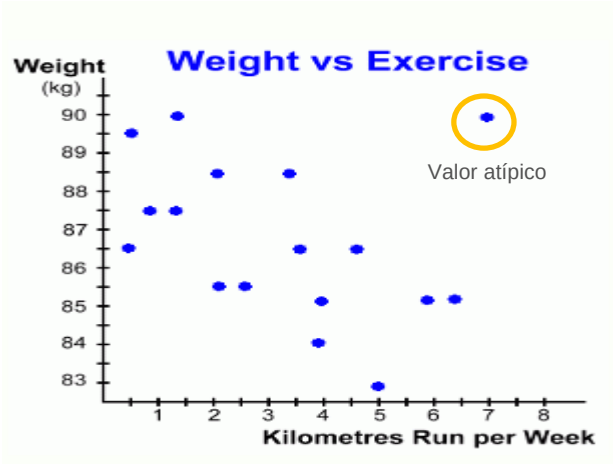
Count

		Survived?		
		Died	Survived	Total
Class	1st	123	200	323
	2nd	158	119	277
	3rd	528	181	709
Total		809	500	1309

Gráfico de dispersión

Permite:





- Los patrones de **asociación** que se manifiestan (en el caso del ejemplo líneal).
- La forma, dirección y fuerza de tal patrón (fuertemente positiva).
- Fijarse en los individuos que se alejan del patrón de la mayoría: los valores atípicos (no parece haber ninguno).
- También podemos identificar clusters de **asociación** (podríamos considerar a cada uno de los tres grupos como un conglomerado)

Gráfico de dispersión

Identificar fuerza de asociación y dirección

A paper in the New England Journal of Medicine: relationship between chocolate and Nobel Prize winners

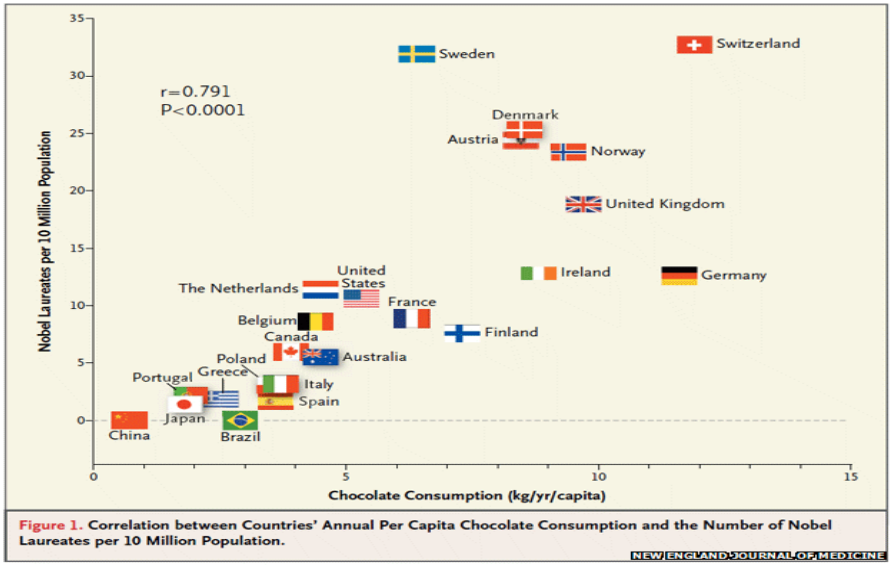
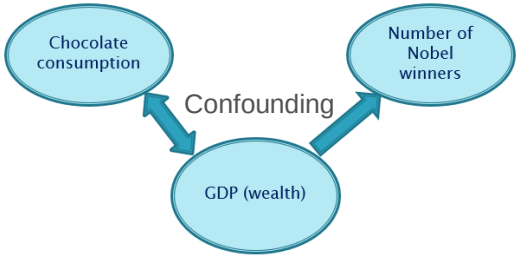


Figure 1. Correlation between Countries' Annual Per Capita Chocolate Consumption and the Number of Nobel Laureates per 10 Million Population.

NEW ENGLAND JOURNAL OF MEDICINE

http://www.nejm.org/doi/full/10.1056/NEJMon1211064

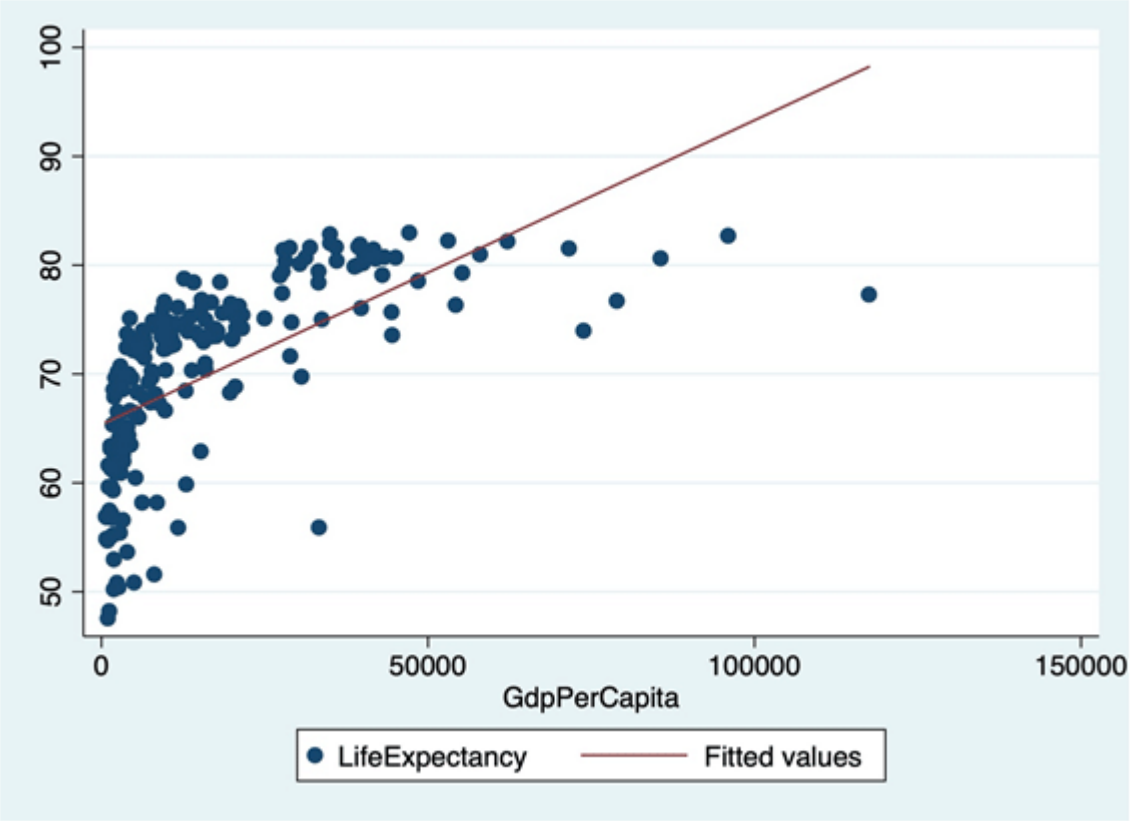


¿Hay algo más que afecte tanto al consumo de chocolate como a los premios Nobel?

Asociación no implica causalidad

Gráfico de dispersión

Identificar no linealidades



Medidas: Covarianza

σ_{xy} | $S_{xy} = \frac{\sum(X - \bar{X})(Y - \bar{Y})}{N}$ | $S_{xy} = \frac{\sum(X - \bar{X})(Y - \bar{Y})}{n - 1}$

Mide lo que covarían las dos variables. Corresponde a la media aritmética de los productos de las desviaciones de cada variable respecto a su media.

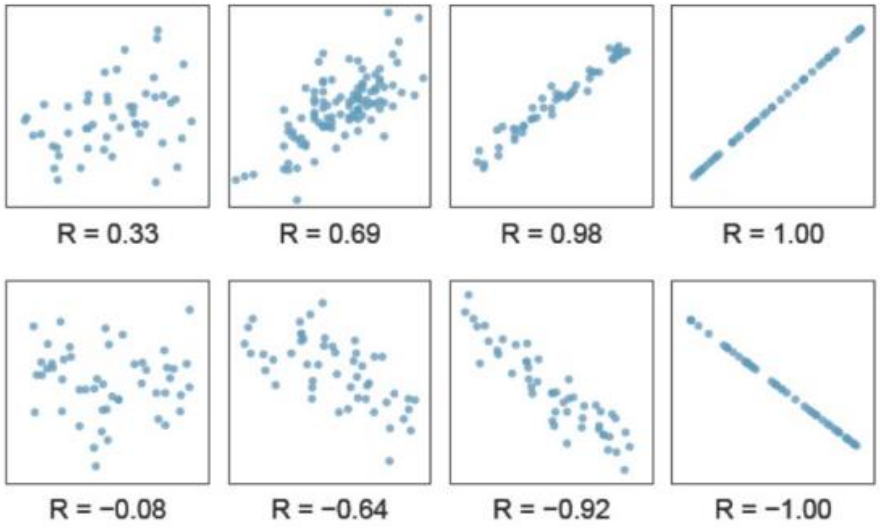
- Un valor positivo indica una relación lineal directa o creciente y un valor negativo indica una relación lineal decreciente. Un valor cercano a cero, poca asociación
- La covarianza nos será muy útil para medir la fuerza de la relación entre las variables
- Es un insumo para calcular otra medida que si mide la fuerza de asociación entre variables

Medidas: Correlación de Pearson

Indica si existe **asociación lineal** entre dos variables.

Medida más útil que la covarianza ya que indica el sentido como el grado (fuerza) de relación.

$$r = \frac{S_{xy}}{S_x S_y}$$



Propiedades:

1. $-1 \leq r_{xy} \leq 1$, rango delimitado.
2. r_{xy} no depende de las unidades de medida.
3. Mide la dirección.
4. Mide el grado; fuerte, débil y no existe.
5. Relación simétrica.
6. No hay relación de causa y efecto.

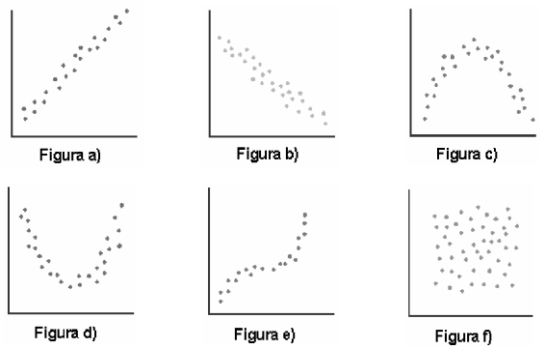
Medidas: Correlación de Pearson

•La ventaja principal es su fácil calculo e interpretación.

Correlation coefficient value			Relationship
−0.3 to +0.3			Weak
−0.5 to −0.3	or	0.3 to 0.5	Moderate
−0.9 to −0.5	or	0.5 to 0.9	Strong
−1.0 to −0.9	or	0.9 to 1.0	Very strong

Cohen, L. (1992). Power Primer. Psychological Bulletin, 112(1) 155-159

•Cuando las variables no presentan relación lineal, **r** no puede medir esta asociación.



•Si X y Y son estadísticamente independientes el coeficiente **r** es cero. No obstante, si **r** es cero no implica que las variables sean independientes.

Medidas: Correlación de Pearson

•**r** calcula la dependencia lineal solo entre pares de variables y no proporciona información sobre la asociación simultanea de mas de dos variables.

•**r** es independiente del origen y la escala. Es adimensional. No es una medida porcentual.

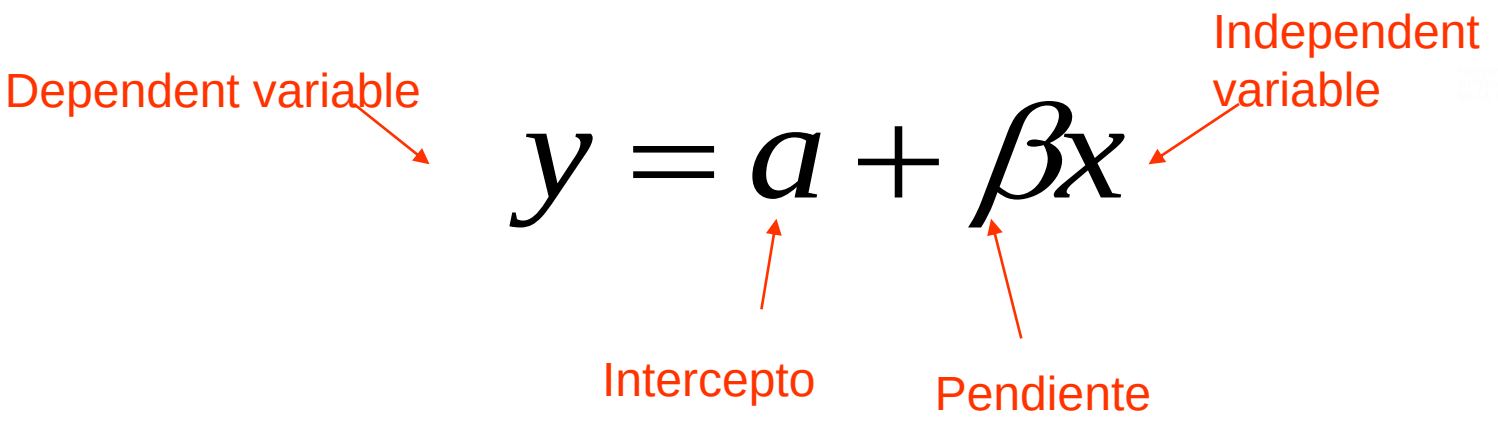
•Una de las condiciones necesarias para extrapolar lo interpretado en el coeficiente de correlación, demanda que las variables sean continuas y **normalmente distribuidas**.

•Existen alternativas de correlación como: correlación de Spearman , Tau de kendall, entre otras**.**

Forma explicita de la relación:

La forma funcional más sencilla:

Ecuación de una recta



Regresión lineal Simple: Modelo

A nivel estadístico:

Simple. Una sola variable independiente.

$$y_i = a + bx_i + e_i$$

Elementos :
Variables y término de error.
Relación funcional.
Parámetros (a, b).

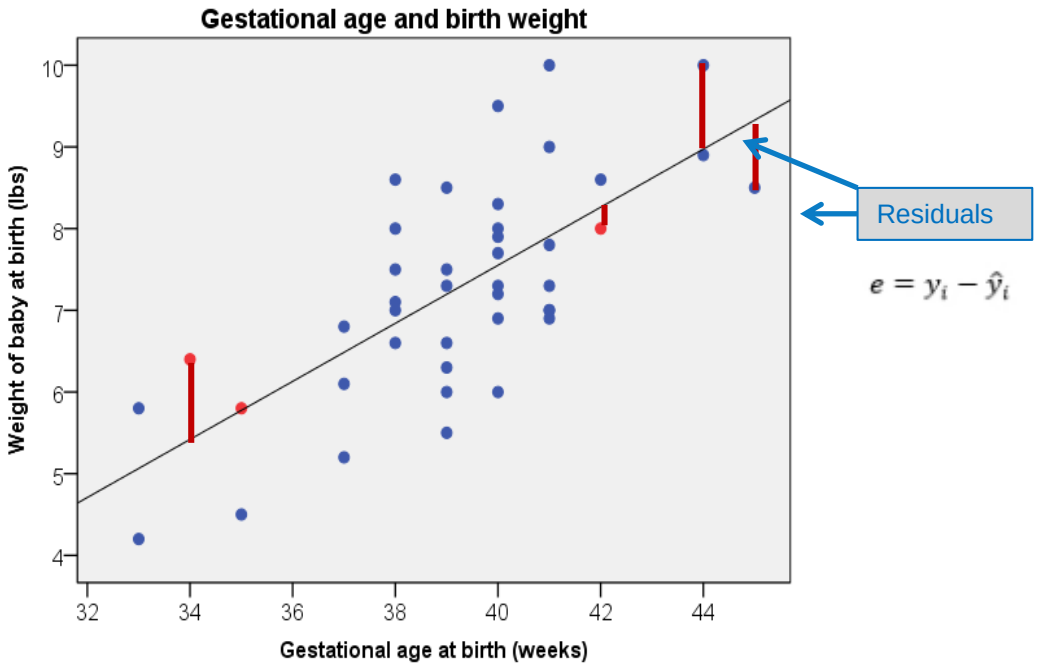
b: La variable dependiente cambia en β2 unidades en promedio cuando la variable independiente (Xi) cambia en una unidad

a: (Es la parte de la variable dependiente que no depende de la variable independiente). El valor que toma la variable dependiente, en promedio, cuando la variable independiente (Xi) es cero

El objetivo fundamental del análisis de regresión es el estudio de la dependencia de una variable, llamada explicada en función de una o más variables llamadas explicativas.

Regresión lineal Simple: coeficientes

Estimaciones muestrales (medidas descriptivas).



Método de los mínimos cuadrados (MMC): Minimiza la suma de los errores de predicción al cuadrado, para encontrar a y b.

$$\hat{y} = a + bx$$

$$b = r\frac{s_y}{s_x} = \frac{S_{xy}}{s_x^2}$$

$$a = \bar{y} - b\bar{x}$$

Regresión lineal Simple: Medida de bondad de ajuste

Coeficiente de determinación.

Una medida muy común de la bondad de ajuste es el coeficiente de determinación R^2 *o* r^2, *la cual proporciona información* respecto a que tan bien la línea de regresión se ajusta a los datos.

Descomposición de la variabilidad de la variable dependiente:

Variabilidad de las observaciones = variabilidad de la predicciones + variabilidad asociada al error

$$r^2 = \frac{s^2_{\hat{y}}}{s^2_y} = \frac{Varianza\ de\ las\ predicciones}{Varianza\ de\ las\ observaciones}$$

$$0 \leq r^2 \leq 1$$

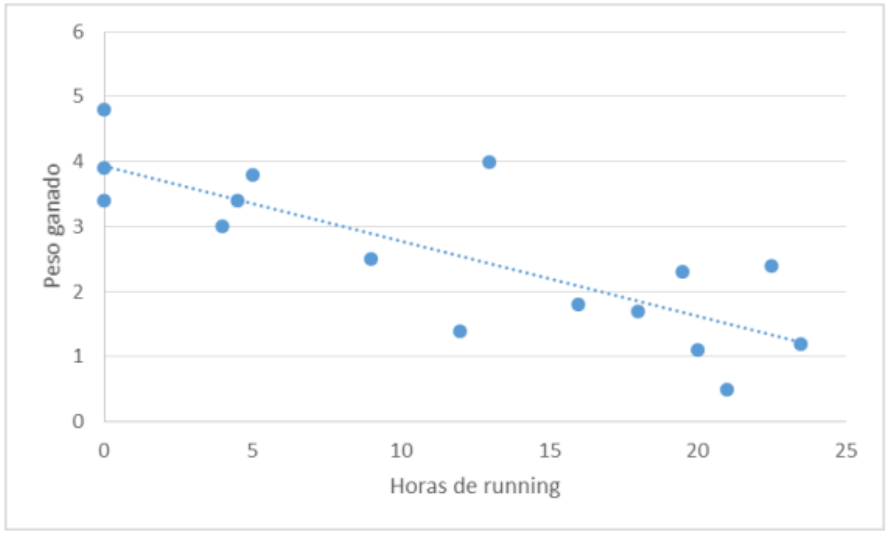
Solo en regresión lineal simple es el cuadrado de la correlación de Pearson.

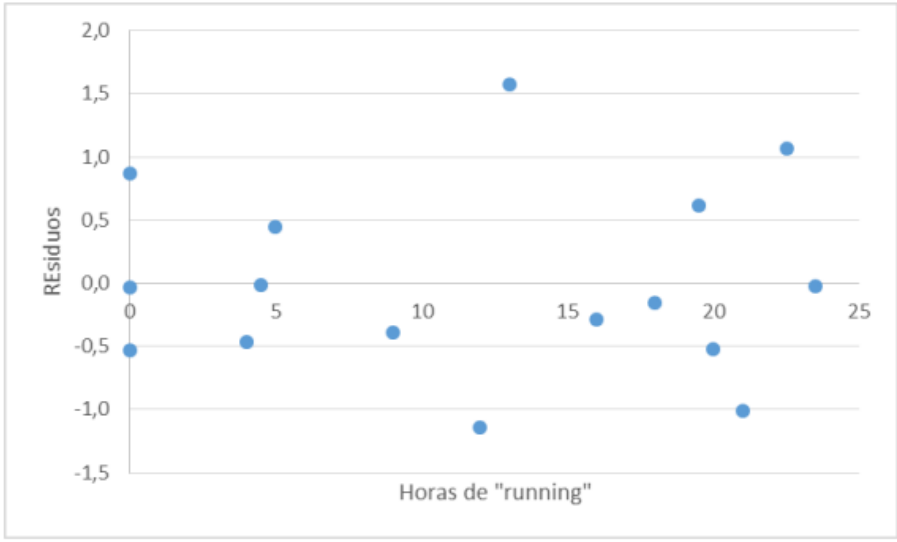
Se puede leer en porcentaje o en proporción.

Mide la proporción de la variabilidad de y explicada por el modelo, o asociada a movimientos de la variable x.

Regresión lineal Simple: gráfico de residuos

Una de los criterios para evaluar una regresión lineal simple es el grafico de residuos.



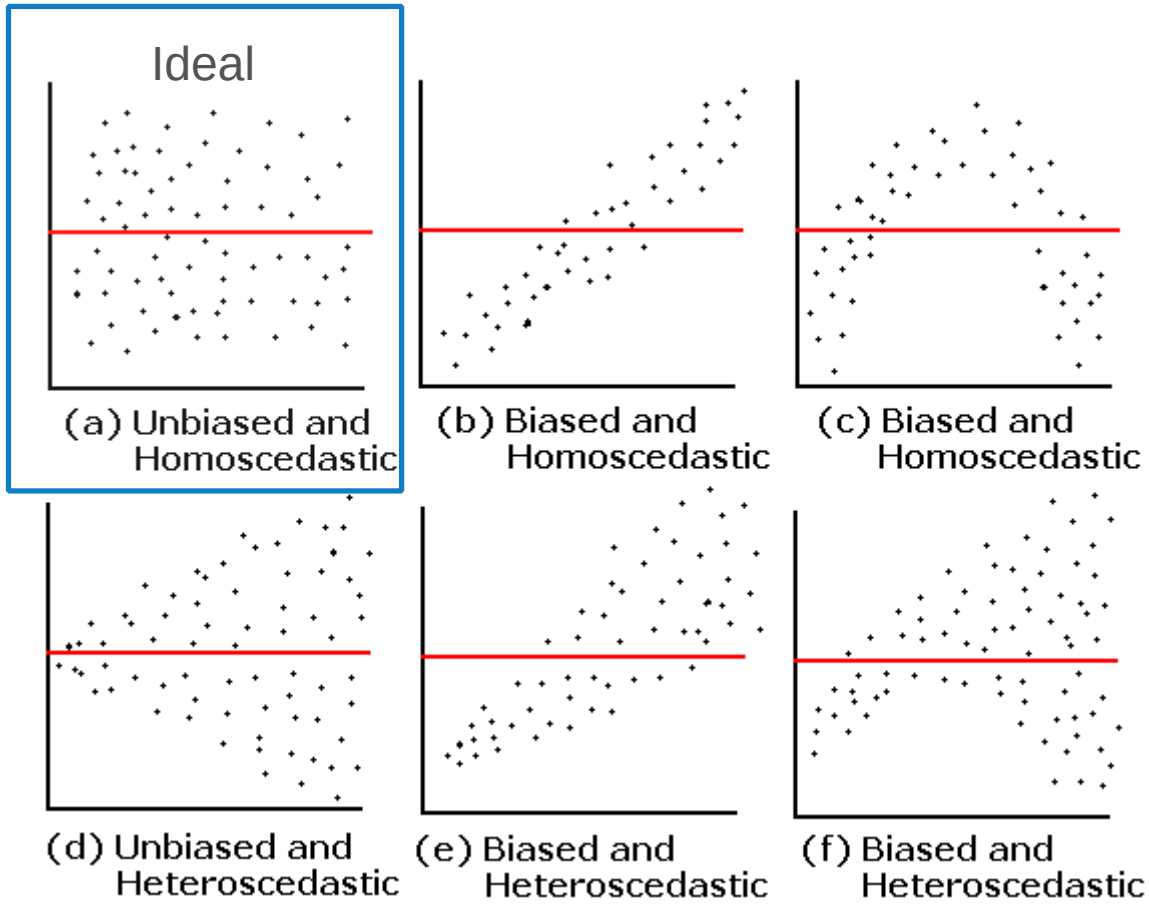


$$\widehat{peso\ ganado}_i = 3.93 - 0.12 horas\ running_i$$

Cuando la recta de regresión está captando y ajustándose bien a las observaciones entonces no deberíamos apreciar **ningún patrón en los residuos.**

Regresión lineal Simple: gráfico de residuos

Problema identificado con el grafico de residuos.



Próxima sesión

- Tema 4: Regresión y correlación.

  - Regresión no lineal.
  - Regresión lineal multivariante.
  - LTS (Least Trimmed Squares)

Inf. Importante

Fecha de entrega de actividad 1:

16/12/2022

www.unir.net