NOVENA Y DÉCIMA CLASE

Analytics AoZ

1. Pruebas semiparamétricas
   - ❑ Modelo robusto - ROS (Regression Orden in Statistics)
2. Correlación en Rstudio
   - ❑ Definiciones y Conceptos.
   - ❑ Correlación de Pearson.
   - ❑ Correlación de Spearman.
   - ❑ Correlación de Kendall.
3. Regresión Lineal:
   - ❑ Definiciones y Conceptos.
   - ❑ Supuesto de la Regresión Lineal.
   - ❑ Gráficos Adicionales para ver normalidad.
4. Ejercicio de Regresión Lineal Múltiple
5. Regresiones no Lineales
   - ❑ Regresión Cuadrática
   - ❑ Regresión Polinómica
   - ❑ Otros tipos.
6. Aplicado en Rstudio a una Base de Datos Geológica (Hidrogeológica, Mina).

REPASO7 y REPASO8 : EJERCICIO PARA AFIANZAR LO APRENDIDO

GUIA PARA CALSIFICAR TEST DE HIPÓTESIS CON VARIABLES DE RESPUESTA CONTINUA

Analytics AoZ

Table 4.1. Guide to the classification of some hypothesis tests with continuous response variables.

[-, not applicable]

Parametric	Nonparametric	Permutation
Two independent data groups (chap. 5)		
Two-sample *t*-test	Rank-sum test (two-sample Wilcoxon; Mann-Whitney test)	Two-sample permutation test
Matched pairs of data (chap. 6)		
Paired *t*-test	Signed-rank test, sign test	Paired permutation test
Three or more independent data groups (chap. 7)		
Analysis of variance	Kruskal-Wallis test	One-way permutation test
Three or more dependent data groups (chap. 7)		
Analysis of variance without replication	Friedman test, aligned-rank test	-
Two-factor group comparisons (chap. 7)		
Two-factor analysis of variance	Brunner-Dette-Munk (BDM) test	Two-factor permutation test
Correlation between two continuous variables (chap. 8)		
Pearson's r (linear correlation)	Spearman's ρ or Kendall's τ (monotonic correlation)	Permutation test for Pearson's r
Model of relation between two continuous variables (chaps. 9 and 10)		
Linear regression	Theil-Sen line	Bootstrap of linear regression

Referencia: Statical Methods in Water Resource 2020 , USGS

ANALISIS BIVARIADO

Análisis Bivariado

dos Variables cualitativas

Cualitativa vs. Cuantitativa

Cuantitativa vs Cuantitativa

Gráficos

politómicas

Grafico de dispersión

Correlación de Pearson

Regresión lineal

Sobrevida

CORRELACIÓN

Mide la fuerza de asociación entre dos variables cuantitativas continuas, tales como dos concentraciones químicas, o entre cantidad de precitación y tiempo, otras.

La correlación no provee evidencia para una relación causal entre las dos variables.

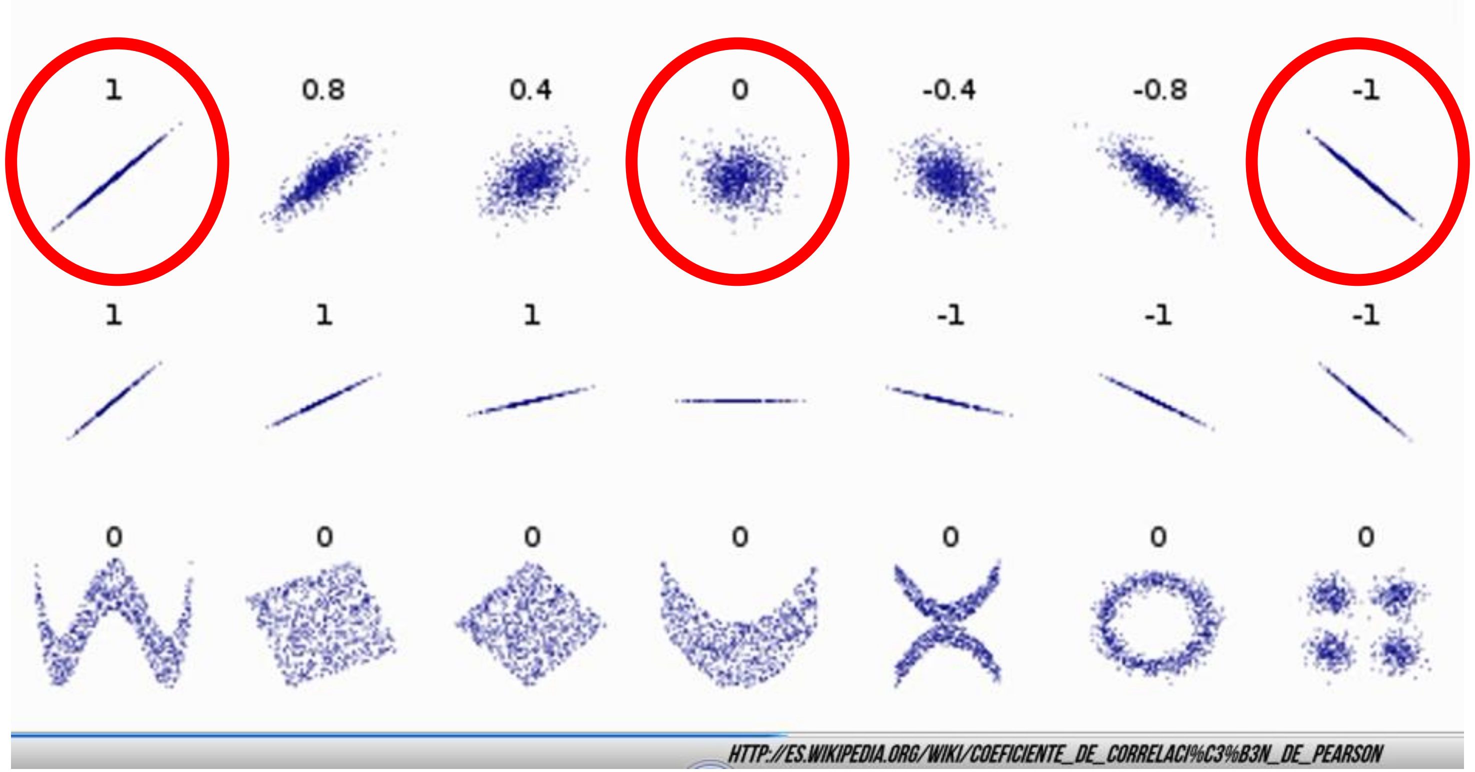
Para evidencias esta causalidad debemos recurrir a fuentes fuera de la estadística como el conocimiento del proceso envuelto.

Medidas de correlación (en general asignadas como ρ) no tienen unidad de medida (adimensional) y la escala tiene a estar en $-1 <= \rho <= 1$. Cuando no existe correlación entre las variables se indica que $\rho=0$, mientras que si es correlación negativa total ($\rho=-1$) y positiva total ($\rho=1$).

H0: $\rho = 0$ versus H1: $\rho \neq 0$

Cuando una variable es una medida de el tiempo o posición, correlación testea para tendencia temporal o tendencia espacial.

CORRELACIÓN VALORES Y DIAGRAMA DE DISPERSIÓN



Monotónica versus Correlación Lineal

Analytics AoZ

La fuerza de una medida lineal es disminuida o diluida por no linealidad, resultando en un bajo coeficiente de correlación y menos significante que la relación lineal teniendo el mismo diagrama de dispersión.

Tenemos tres medidas para medir la correlación. Ninguna de estas detectara relaciones no monotónicas, cuando el patrón es doble de regreso u como el mostrado en la siguiente diapositiva figura 8.3

No detecta correlaciones no monotónicas

Kendall´s tau = tau
Spearman´s rho = rho

Rangos (relaciones monotónicas) → *Resistentes a efectos de outliers.*

Pearson´s = r

Relación Lineal

Monotonica (no linear) versus Correlación Lineal

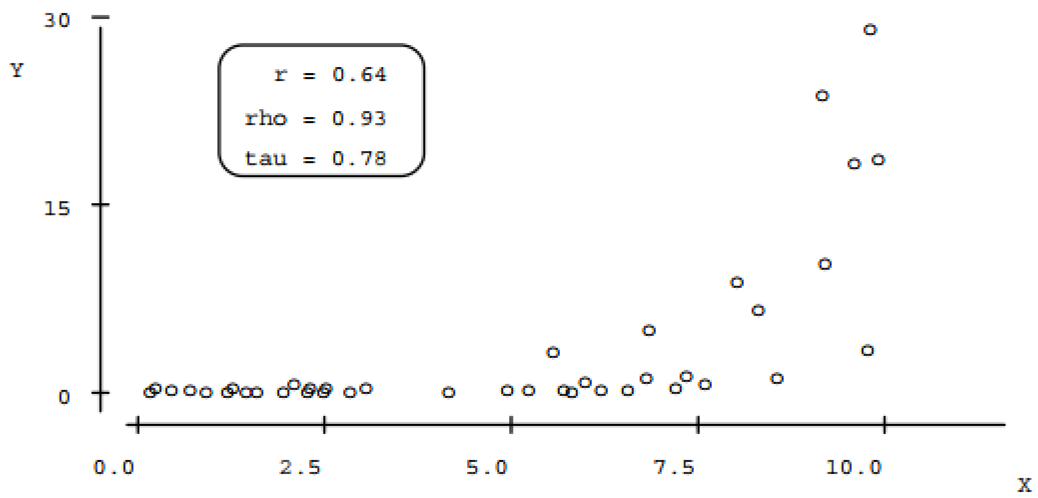


Figure 8.1 Monotonic (nonlinear) correlation between x and y.

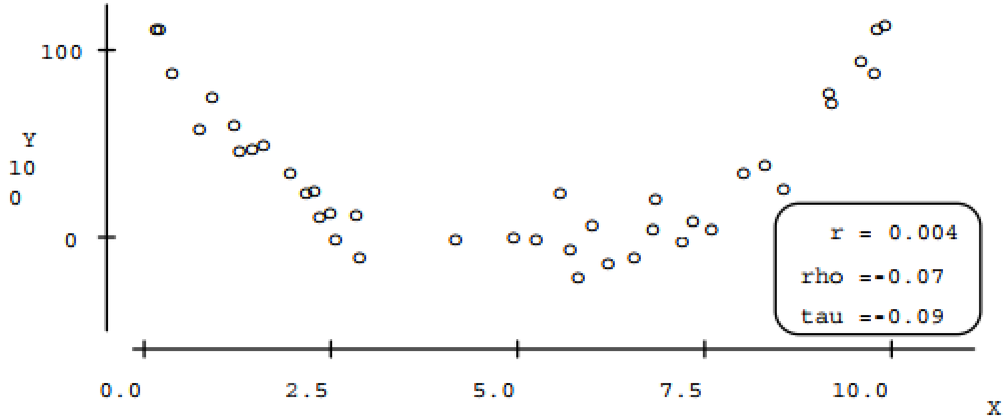


Figure 8.3 Non-monotonic relationship between X and Y.

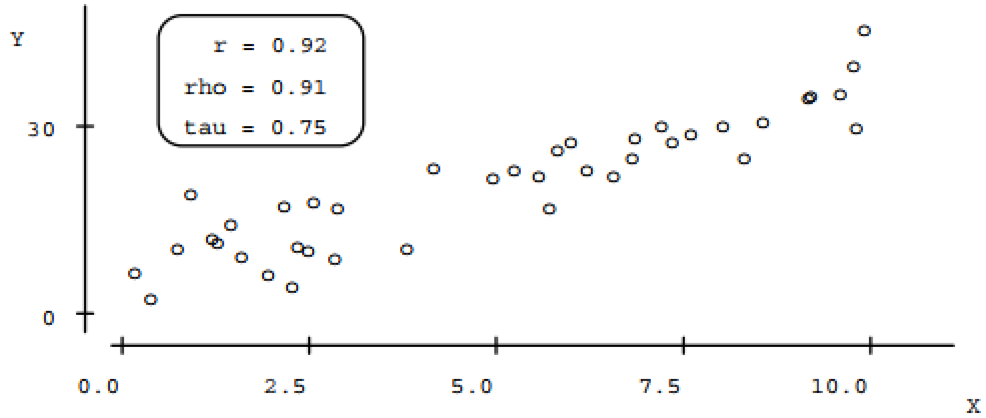


Figure 8.2 Linear correlation between X and Y.

PEARSON´S R

- La medida más común para medir la correlación es la de Pearson.
- r mide la relación lineal de asociación entre dos variables cuantitativas.
- Pearson´s r es no resistente a los outliers como tau o rho porque es calculado usando medida no resistente – media y desviación estándar. Este asume que la data sigue una distribución normal bivariante.
- Pearson´s r no es útil para describir la corrlación entre variables hidrológicas no transformadas.

$$r = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i-\bar{x}}{s_x}\right)\left(\frac{y_i-\bar{y}}{s_y}\right) \longleftrightarrow r = \frac{Cov_{xy}}{S_x S_y}$$

Donde:

$$Cov_{xy} = \frac{\sum_1^n (x_i-\bar{x})(y_i-\bar{y})}{n-1}$$

- La significancia de r puede ser testeada por la determinación r difiere de cero. El test estadístico ***tr*** es calculado y comparado con una tabla de la **distribución t** con **n-2 grados de libertad**.

$$t_r = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

- Ejemplo si tenemos lo siguiente:

Example 1: 10 pairs of x and y are given below, ordered by increasing x:

y	1.22	2.20	4.80	1.28	1.97	1.46	2.64	2.34	4.84	2.96
x	2	24	99	197	377	544	632	3452	6587	53,170

	x	y
mean	6508.6	2.57
s	16531.6	1.31

$$r = \frac{1}{9}\sum_{i=1}^{9}\left(\frac{x_i - 6508.6}{16531.6}\right)\left(\frac{y_i - 2.57}{1.31}\right) = 0.174$$

$$t_r = \frac{0.174\sqrt{8}}{\sqrt{1-(0.174)^2}} = 0.508$$

El test para ver que r es diferente significativamente de cero, y además y es linealmente dependiente de x

con un p-valor=0.63 desde la tabla de la distribución t.
Además H:r=0 no es rechazado, y ***y*** no es linealmente dependiente (o relacionado) para x como medida de r. Esto difiere de cálculos que son realizados con rho y tau, los cuales dan p-valores de 0.04 y 0.07 respectivamente que indican asociación entre x e y.

No depende de las unidades de medidas de las variables. No es adecuado cuando hay outliers, distribución asimétrica, menor a 30 datos, entre otras consideraciones.

SPEARMAN´S RHO

Analytics AoZ

- Alternativa al coeficiente de correlación de Pearson, no paramétrico.
- rho es el más sencillo para entender como coeficiente de correlación linear calculado desde rangos de un conjunto de data.
- Rho y tau están a diferentes escalas para medir la misma correlación.
- P-valores de Rho y tau deben ser significativamente iguales.
- Es importante mencionar que para muestras grandes y aproximaciones de rangos para rho no se ajusta bien la distribución para un buen test estadístico en muestras pequeñas (n<20), en contraste a Kendall´s tau. Esta es una razón porque tau es preferido a rho.

$$\mathrm{rho} = \frac{\sum_{i=1}^{n} (Rx_i Ry_i) - n\left(\frac{n+1}{2}\right)^2}{n(n^2-1)/12}$$

Otra forma:

$$r_s = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)}$$

Donde:

$d_i = RX_i - RY_i$

n: número de pares de observaciones.

PRUEBA DE HIPÓTESIS SPEARMAN

Analytics AoZ

1. <u>Planteamiento de las hipótesis.</u>

H_0: $\rho = 0$ (el coeficiente de correlación no es significativo)

H_1: $\rho \neq 0$ (el coeficiente de correlación es significativo)

2. <u>Estadístico de prueba</u>

$$t_c = \frac{r_s\sqrt{n-2}}{\sqrt{1-r_s^{\,2}}}$$

3. <u>Criterio de decisión</u>

A un nivel de significación de $\alpha = 0{,}05$, se obtiene valor crítico.

Se rechaza H_0 si $|t_c| > t_{(n-2,1-\alpha/2)}$ ó Se rechaza H_0 si p-valor < α.

4. <u>Criterio de decisión</u>

El coeficiente de correlación es significativo.

- Ejemplo si tenemos lo siguiente:

Example 1: 10 pairs of x and y are given below, ordered by increasing x:

y	1.22	2.20	4.80	1.28	1.97	1.46	2.64	2.34	4.84	2.96
x	2	24	99	197	377	544	632	3452	6587	53,170

Example 1, continued

For the example 1 data, the data ranks are

Ry	1	5	9	2	4	3	7	6	10	8
Rx	1	2	3	4	5	6	7	8	9	10

$(Rx_i \bullet Ry_i)$	1	10	27	8	20	18	49	48	90	80,	$\Sigma = 351$

$$\text{Rho} = \frac{351 - 10\,(5.5)^2}{1099/12} = \frac{48.5}{82.5} = 0.588$$

El p-valor = 0.04 desde la tabla de Bhattacharyya y Jhonson (1977).

La aproximación del nivel de significancia test para r de Pearson en los rangos de data tenia p-valor = 0.074, no cercano al valor exacto del p. Mientras usando el rho de Spearman para muestras menores a 20, el valor exacto de p deberá ser usado.

KENDALL´S TAU

- Medida la fuerza de relación entre dos variables continuas monotónicas.
- Es muy útil para variables que exhiben asimetría alrededor de la relación general.
- Puede ser implementado en caso donde la data esta censurada, tales como concentraciones debajo del limite de reporte (valores de variables geológicas debajo del límite de detección de laboratorio).
- Si r=0.9 entonces tau=0.7 o superior.
- Si la muestra es grande la aproximación de los p-valor es muy cercano al valor exacto, incluso en muestras medianamente pequeñas.
- Tau es fácil de calcular a mano, resistente a outliers, y medida de todas las correlaciones monotónicas (linear y no linear).
- Como es un método de correlación por rangos, tau es invariante para poder de transformaciones monotónicas de una o ambas variables, es decir para la correlación de **log(y)** versus **log(x)** sería idéntica a **y** versus **log(x),** y de **y** versus **x**.

A two-sided test for correlation will evaluate the following equivalent statements for the null hypothesis H_0, as compared to the alternate hypothesis H_1:

H_0:
- a) no correlation exists between x and y ($\tau = 0$), or
- b) x and y are independent, or
- c) the distribution of y does not depend on x, or
- d) Prob ($y_i < y_j$ for $i < j$) = 1/2.

H_1:
- a) x and y are correlated ($\tau \neq 0$), or
- b) x and y are dependent, or
- c) the distribution of y (percentiles, etc.) depends on x, or
- d) Prob ($y_i < y_j$ for $i < j$) $\neq$ 1/2.

8.2.2 Large Sample Approximation

For n > 10 the test statistic can be modified to be closely approximated by a normal distribution. This large sample approximation Z_S is the same form of approximation as used in Chapter 5 for the rank-sum test, where now

$d = 2$ (S can vary only in jumps of 2),

$\mu_S = 0$, and

$\sigma_S = \sqrt{(n/18)\bullet(n-1)\bullet(2n+5)}$.

$$Z_S = \begin{cases} \dfrac{S-1}{\sigma_s} & \text{if } S>0 \\ 0 & \text{if } S=0 \\ \dfrac{S+1}{\sigma_s} & \text{if } S<0 \end{cases} \qquad [8.3]$$

The null hypothesis is rejected at significance level α if $|Z_S| > Z_{crit}$ where Z_{crit} is the value of the standard normal distribution with a probability of exceedance of $\alpha/2$. In the case where some of the x and/or y values are tied the formula for σ_S must be modified, as discussed in the next section.

Analytics AoZ

The test statistic S measures the monotonic dependence of y on x. Kendall's S is calculated by subtracting the number of "discordant pairs" M, the number of (x,y) pairs where y decreases as x increases, from the number of "concordant pairs" P, the number of (x,y) pairs where y increases with increasing x:

$$\boxed{S = P - M} \qquad [8.1]$$

where P = "number of pluses", the number of times the y's increase as the x's increase, or the number of $y_i < y_j$ for all $i < j$,
M = "number of minuses," the number of times the y's decrease as the x's increase, or the number of $y_i > y_j$ for $i < j$.
for all $i = 1,....(n-1)$ and $j = (i+1),.....n$.

Note that there are n(n–1)/2 possible comparisons to be made among the n data pairs. If all y values increased along with the x values, S = n(n–1)/2. In this situation, the correlation coefficient τ should equal +1. When all y values decrease with increasing x, S = –n(n–1)/2 and τ should equal –1. Therefore dividing S by n(n–1)/2 will give a value always falling between –1 and +1. This then is the definition of τ, measuring the strength of the monotonic association between two variables:

Kendall's tau correlation coefficient

$$\boxed{\tau = \frac{S}{n(n-1)/2}} \qquad [8.2]$$

To test for significance of τ, S is compared to what would be expected when the null hypothesis is true. If it is further from 0 than expected, H_0 is rejected. For $n \leq 10$ an exact test should be computed. The table of exact critical values is found in table B8 of the Appendix.

<u>Example 1:</u> 10 pairs of x and y are given below, ordered by increasing x:

y	1.22	2.20	4.80	1.28	1.97	1.46	2.64	2.34	4.84	2.96
x	2	24	99	197	377	544	632	3452	6587	53,170

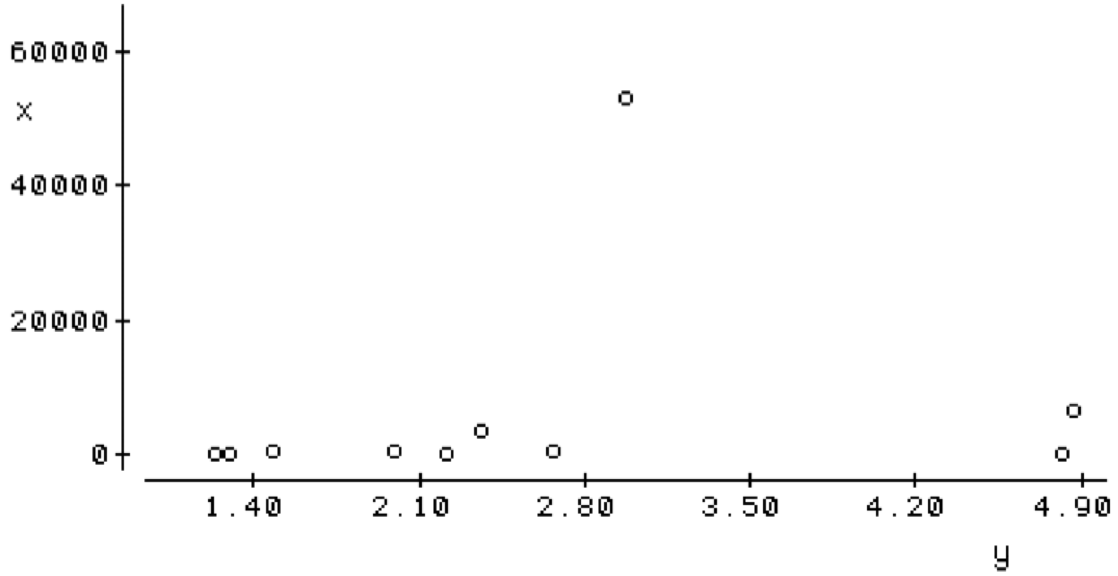


Figure 8.4 Example 1 data showing one outlier present.

To compute S, first compare $y_1 = 1.22$ with all subsequent y's (y_j, j>1).

2.20 > 1.22, so score a +
4.80 > 1.22, score a +
1.28 > 1.22, score a +
1.97 > 1.22, score a + etc.

All subsequent y's are larger, so there are 9 +'s for i=1.

Move on to i=2, and compare $y_2 = 2.20$ to all subsequent y's.

4.80 > 2.20, so score a +
1.28 < 2.20, score a –
1.97 < 2.20, score a –
1.46 < 2.20, score a – etc.

There are 5 +'s and 3 –'s for i=2. Continue in this way, until the final comparison of y_{n-1} = 4.84 to y_n. It is convenient to write all +'s and –'s below their respective y_i, as below:

y_i	1.22	2.20	4.80	1.28	1.97	1.46	2.64	2.34	4.84	2.96
	+	+	–	+	–	+	–	+	–	
	+	–	–	+	+	+	+	+		
	+	–	–	+	+	+	+			
	+	–	–	+	+	+				
	+	+	–	+	+					
	+	+	+	+						
	+	+	–							
	+	+								
	+									

In total there are 33 +'s (P = 33) and 12 –'s (M = 12). Therefore S = 33 – 12 = 21.
There are 10•9/2 = 45 possible comparisons, so $\tau = 21/45 = 0.47$.
Turning to table B8, for n=10 and S=21, the exact p-value is 2•0.036 = 0.072.

The large sample approximation is

$$Z_S = (21-1) / \sqrt{(10/18)\cdot(10-1)\cdot(20+5)}$$
$$= 20/(11.18) = 1.79.$$

From a table of the normal distribution, the 1-sided quantile for 1.79 = 0.963
so that $p \cong 2 \cdot (1-.963) = 0.074$

8.2.3 Correction for Ties

To compute τ when ties are present, tied values of **either x or y** produce a 0 rather than + or – . Ties do not contribute to either P or M. S and τ are computed exactly as before. An adjustment is required for the large sample approximation Z_S , however, by correcting the σ_S formula.

In order to compute σ_S in the presence of ties, both the number of ties and the number of values involved in each tie must be counted. Consider for example a water quality data set (in units of µg/L) of 17 values (n=17) shown here in ascending order.

<1, <1, <1, <1, <1, 2, 2, 2, 3, 5, 5, 7, 9, 10, 10, 14, 18.

There are a total of 4 tied groups in the data set. The largest tied group in the data set is of 5 values (tied at <1 µg/L), there are no tied groups of 4, there is 1 tied group of 3 (at 2 µg/L), and there are 2 tied groups of 2 (at 5 and 10 µg/L). For completeness note that there are 5 "ties" of extent 1 (untied values at 3, 7, 9, 14, and 18 µg/L). These appropriately never add to the correction because (i–1) always equals zero. Kendall (1975) defined the variable t_i as the number of ties of extent i. For this data set $t_5 = 1$ (1 tie of extent 5), $t_4 = 0$ (no ties of extent 4), $t_3 = 1$ (1 tie of extent 3), $t_2 = 2$ (2 ties of extent 2) and $t_1 = 5$ (5 "ties" of extent 1). For i>5, $t_i = 0$. Kendall's correction to σ_S in the presence of ties is:

$$\sigma_S = \sqrt{\frac{[n\,(n-1)\,(2n+5) - \sum_{i=1}^{n} t_i\,(i)\,(i-1)\,(2i+5)\,]}{18}} \qquad [8.4]$$

So for the example water quality data:

$$\sigma_S = \sqrt{[17\bullet16\bullet39 - 5\bullet1\bullet0\bullet7 - 2\bullet2\bullet1\bullet9 - 1\bullet3\bullet2\bullet11 - 1\bullet5\bullet4\bullet15] / 18}$$

or $\sigma_S = \sqrt{567} = 23.81$. Notice that if the data set could have been measured with sufficient precision (including a lower detection limit) so that no ties existed, then $\sigma_S = \sqrt{589.333} = 24.28$. Thus the ties here represent a rather small loss of information.

Example 2:

The example 1 data are modified to include ties, as follows:

y	1.22	2.20	4.80	1.28	1.97	1.97	2.64	2.34	4.84	2.96
x	2	24	99	99	377	544	632	3452	6587	53,170

Using a 0 to denote a tie, the comparisons used to compute P, M, and S are:

+	+	0_x	+	0_y	+	−	+	−
+	−	−	+	+	+	+	+	
+	−	−	+	+	+	+		
+	−	−	+	+	+			
+	+	−	+	+				
+	+	+	+					
+	+	−						
+	+							
+								

0_x : tie in x
0_y : tie in y

In total there are 33 +'s (P=33) and 10 −'s (M=10). Therefore S = 33−10 = 23, and $\tau = 23/45 = 0.51$. The exact two-sided p-value from table B8 is $2 \cdot 0.023 = 0.046$. For the large sample approximation, there are 2 ties of extent 2, so that

$$\sigma_S = \sqrt{[10 \cdot 9 \cdot 25 - 2 \cdot 2 \cdot 1 \cdot 9] / 18} = \sqrt{123} = 11.09$$

whereas without the tie σ_S was 11.18. Computing Zs,

$$Z_S = (23-1) / \sqrt{123}$$
$$= 22/(11.09) = 1.98.$$

From a table of the normal distribution, the 1-sided quantile for 1.98 = 0.976 so that $p \cong 2 \cdot (1-.976) = 0.048$.

R function to compute Correlation test

Analytics AoZ

To perform Correlation, the R function **cor**() can be used as follow:

```
cor(x, y = NULL, use = "everything",method = c("pearson", "kendall", "spearman"))
```

Arguments

x: numeric vector, matrix or data frame.

y: NULL (default) or a vector, matrix or data frame with compatible dimensions to x. The default is equivalent to y = x (but more efficient).

na.rm: logical. Should missing values be removed?

use: an of) one of the strioptional character string giving a method for computing covariances in the presence of missing values. This must be (an abbreviation ngs "everything", "all.obs", "complete.obs", "na.or.complete", or "pairwise.complete.obs".

method: a character string indicating which correlation coefficient (or covariance) is to be computed. One of "pearson" (default), "kendall", or "spearman": can be abbreviated.

V: symmetric numeric matrix, usually positive definite such as a covariance matrix.

"LO QUE ESCUCHO LO OLVIDO. LO QUE VEO LO RECUERDO. PERO LO QUE HAGO, LO ENTIENDO."

VAMOS A RESOLVER UN EJERCICIO

NOVENA Y DÉCIMA CLASE

Analytics AoZ

1. Pruebas semiparamétricas
   - Modelo robusto - ROS (Regression Orden in Statistics)
2. Correlación en Rstudio
   - Definiciones y Conceptos.
   - Correlación de Kendall.
   - Correlación de Spearman.
   - Correlación de Pearson.
3. Regresión Lineal Simple:
   - Definiciones y Conceptos.
   - Supuesto de la Regresión Lineal.
   - Gráficos Adicionales para ver supuestos.
4. Ejercicio de Regresión Lineal Múltiple
5. Regresiones no Lineales
   - Regresión Cuadrática
   - Regresión Polinómica
   - Regresion de Poisson, Logistica, Cox.
6. Aplicado en Rstudio a una Base de Datos Geológica (Hidrogeológica, Mina).

REPASO7 y REPASO8 : EJERCICIO PARA AFIANZAR LO APRENDIDO

REGRESIÓN LINEAL SIMPLE

Analytics AoZ

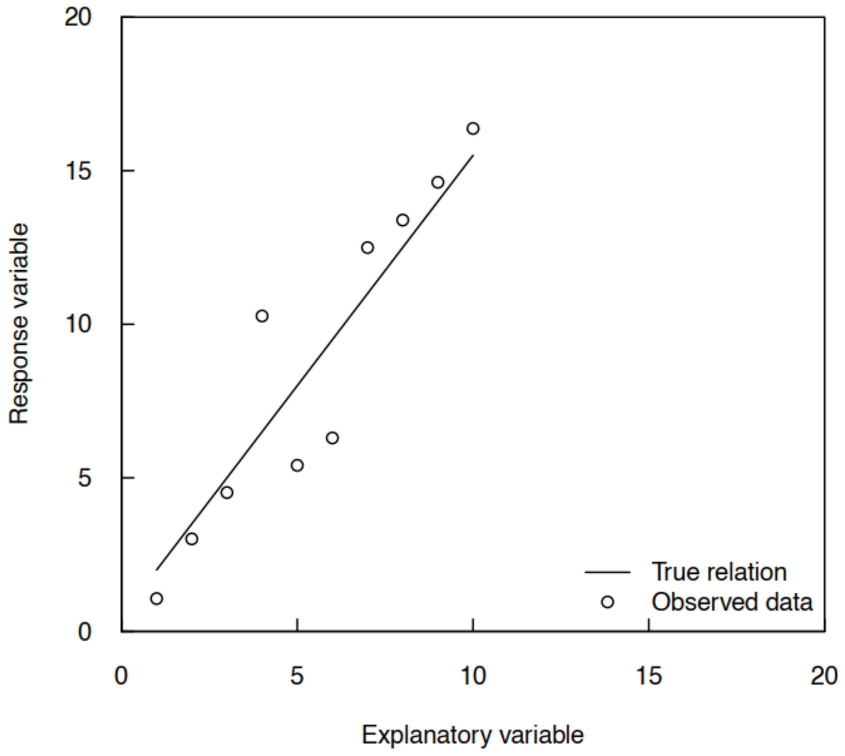
El modelo para la regresión lineal simple es:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

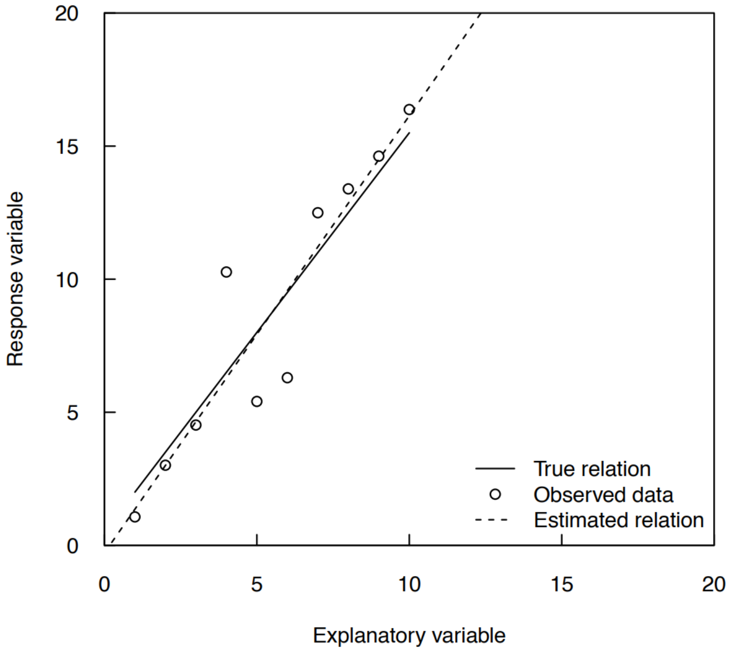
$$\text{for } i = 1, 2, \ldots, n \ ,$$

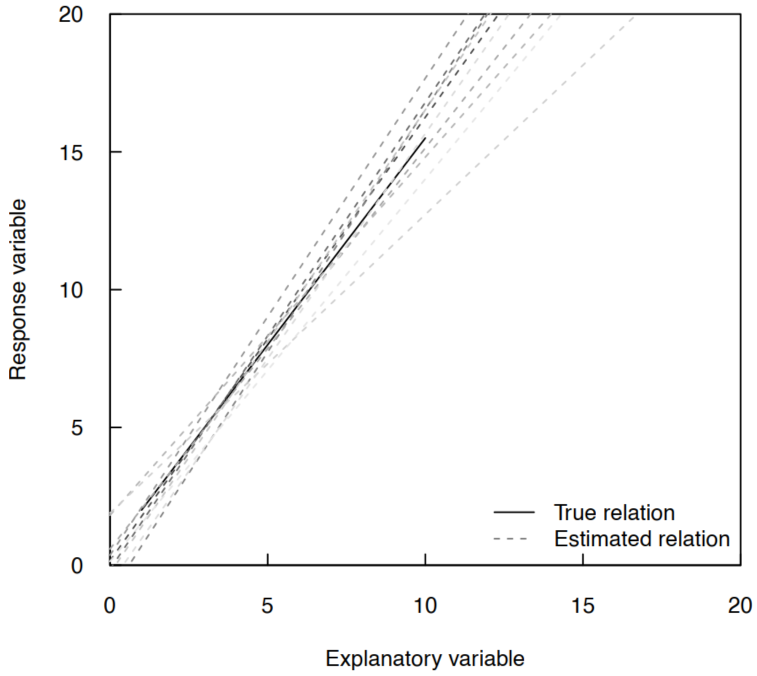
yi: es la ith observación de la variable respuesta
xi: es la ith observación de la variable explicadora,
β0: es el intercepto,
β1: es la pendiente (el cambio en y con respecto a x).
εi: es el error aleatorio residual para la ith observación,
n: es el tamaño de muestra.

El error alrededor del modelo lineal **εi**, es una **variable aleatoria**. Esto es, su *magnitud es la no explicada variabilidad de la data*. Los valores de **εi** son asumidos tener **media cero**, y **una varianza constante**, σ2, que **no depende de x**. Los valores de εi son asumidos a ser independientes de xi.



El modelo para la regresión lineal simple no es nada más que un problema de minimización; es, la regresión lineal es un proceso de estimar la línea que minimiza algunas medidas de distancia entre la línea observada y los puntos de data. En una regresión ordinaria de ***mínimos cuadrados***, la línea estimada minimiza la suma de las distancias verticales cuadradas entre la data observada y la línea.

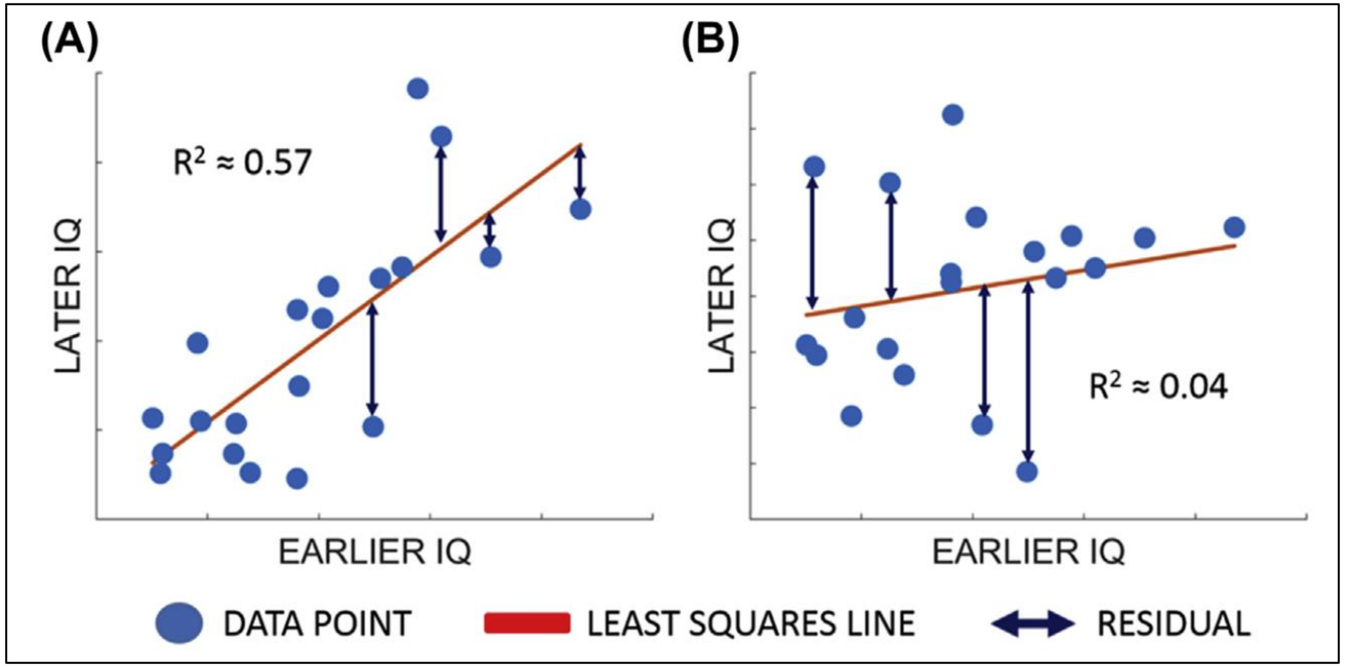


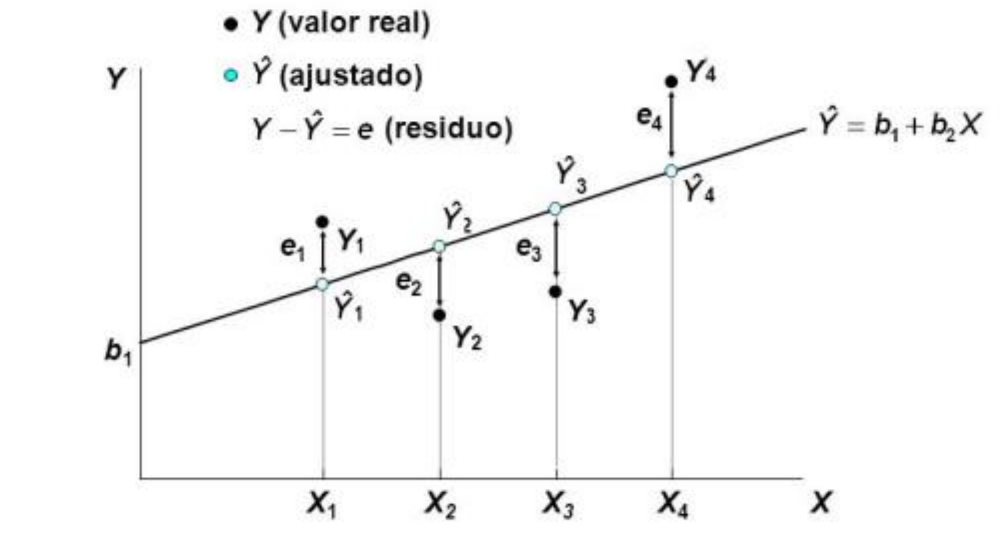


La solución de los mínimos cuadrados : encuentra dos estimados, b0 y b1, tales que la suma de las diferencias cuadrados entre el estimado y el observado se minimiza. En términos matemáticos:

$$\sum_{i=1}^{n} (\hat{y}_i - y_i)^2 \text{ is minimized, where } \hat{y}_i \text{ is the OLS estimate of } y: \qquad \hat{y}_i = b_0 + b_1 x_i$$

GRÁFICOS





Las diferencias entre valores reales y ajustados de Y se denominan residuos.

Supuestos Necesarios para aplicar mínimos cuadrados a la regresión.

Analytics AoZ

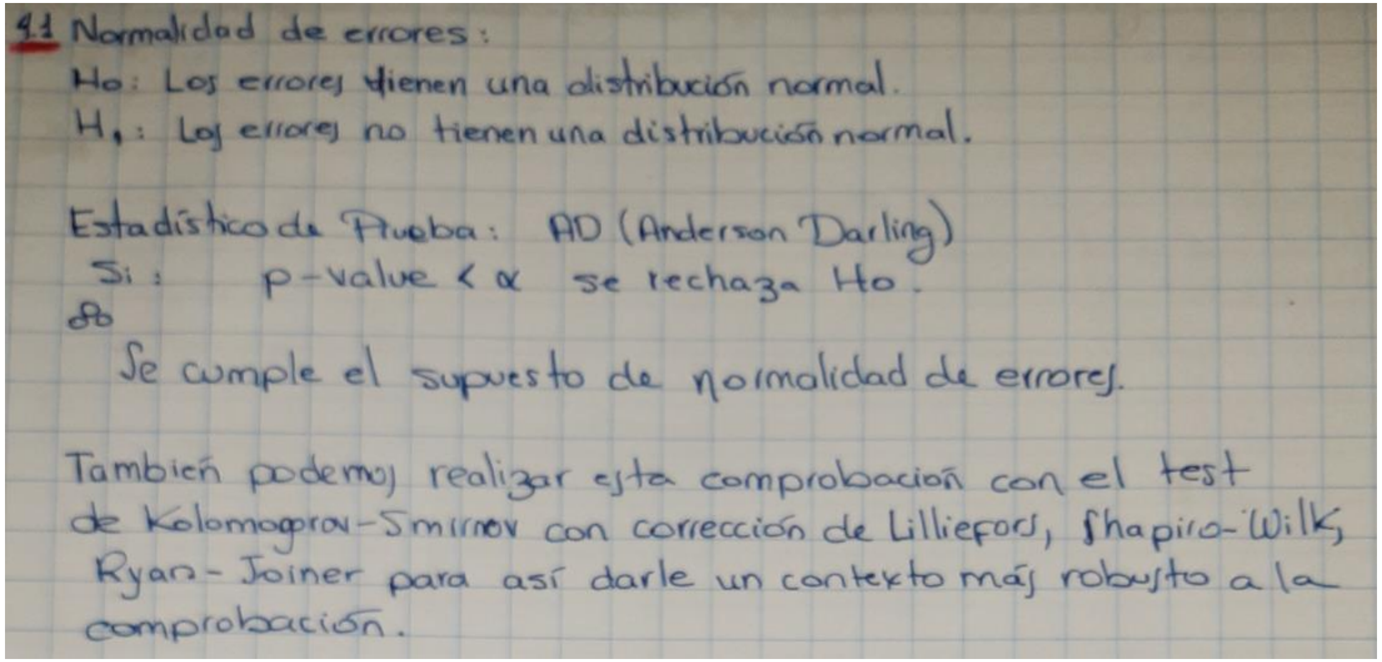
Table 9.2. Assumptions necessary for the purposes to which ordinary least squares (OLS) regression is applied.

[X, the assumption is required for that purpose; -, assumption is not required]

Assumption	Purpose			
	Predict *y* given *x*	Predict *y* and a variance for the prediction	Obtain best linear unbiased estimator of *y*	Test hypotheses, estimate confidence or prediction intervals
Model form is correct: *y* is linearly related to *x*.	X	X	X	X
Data used to fit the model are representative of data of interest.	X	X	X	X
Variance of the residuals is constant (homoscedastic). It does not depend on *x* or on anything else such as time.	-	X	X	X
The residuals are independent of *x*.	-	-	X	X
The residuals are normally distributed.	-	-	-	X

SUPUESTOS DEL MODELO DE REGRESION LINEAL

1. **Suposición de la normalidad de los errores:** Para comprobar este supuesto se usa la prueba de Anderson Darling, o el test de Kolomologorov-Smirnov con corrección de Lilliefors, Shapiro-Wilk, Ryan – Joiner para darle un contexto sería probar todos y comparar.


1.1 Normalidad de errores:

Ho: Los errores tienen una distribución normal.

H1: Los errores no tienen una distribución normal.

Estadístico de Prueba: AD (Anderson Darling)

Si: p-value < α se rechaza Ho.

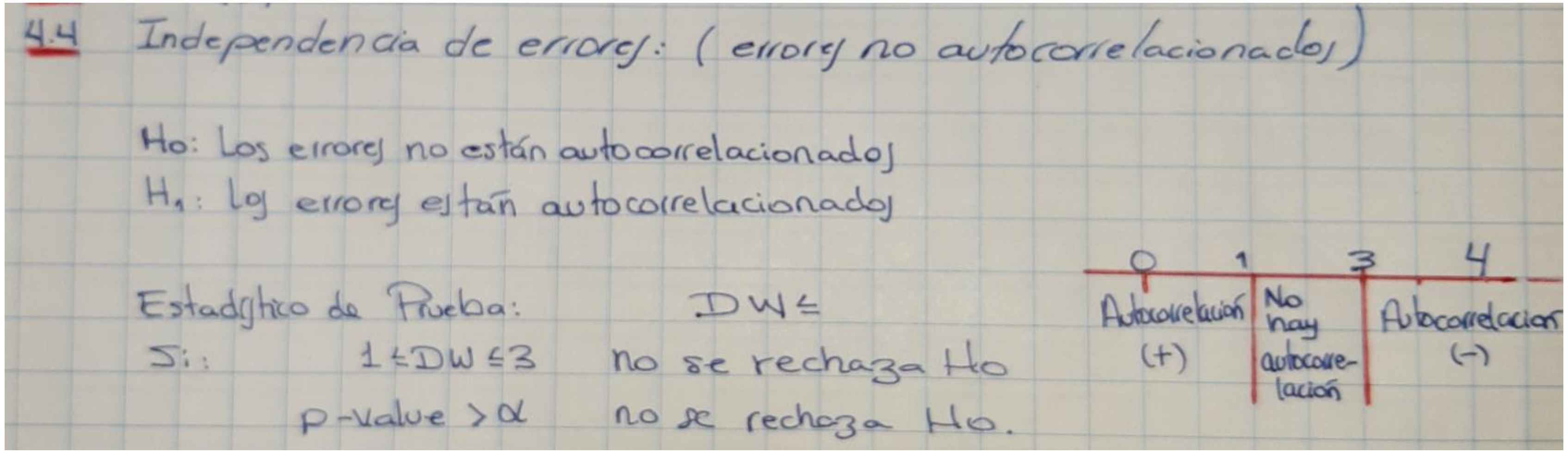
Se cumple el supuesto de normalidad de errores.

También podemos realizar esta comprobación con el test de Kolomogrov-Smirnov con corrección de Lilliefors, Shapiro-Wilk, Ryan-Joiner para así darle un contexto más robusto a la comprobación.

2. **Suposición de la media de los errores es cero:** $E(\varepsilon i) = 0$
3. **Suposición de la varianza de los errores es cero o cte:** $Var(\varepsilon i) = cte = \sigma 2$

4. Independencia de los errores (errores auto correlacionados):

Test de Durbin Watson (DW)



4.4 Independencia de errores: (errores no autocorrelacionados)

Ho: Los errores no están autocorrelacionados
H_1: Los errores están autocorrelacionados

Estadístico de Prueba: DW ≤

Si: $1 \leq DW \leq 3$ no se rechaza Ho

p-value $> \alpha$ no se rechaza Ho.

FORMA MATEMÁTICA DEL CÁLCULO

(5) Determinación del ecuación de regresión o modelo ajustado:

$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ $\qquad$ $\hat{\beta}_0, \hat{\beta}_1$: Aplicando el método de mínimos cuadrados.

El método mínimos cuadrados busca la recta más cerca a los puntos; es decir busca los valores en la cual $\underset{\text{(real)}}{y_i} - \underset{\text{(ajustado)}}{\hat{y}_i}$ sea la más pequeña y así la suma de todas las distancia se simboliza:

$$\text{Suma de cuadrados del error} = \underline{SCE} = \sum_{i=1}^{n} \big(\underbrace{y_i - \hat{y}_i}_{\varepsilon_i = y_i - \hat{y} \text{ (residuales)}}\big)^2$$

sea la más pequeña.

Se escoje $\hat{\beta}_0$ y $\hat{\beta}_1$ que minimice SCE.

FORMA MATEMÁTICA DEL CÁLCULO

"Background" matemático B_0 y B_1:

Si (x_i, Y_i) observados el valor observado Y está dado por Y_i, mientras que el estimado de Y está dado por: $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x$ ....(i)

Luego, la desviación entre el valor observado y estimado de Y está dado por:

$$Y_i - \hat{Y}_i = Y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i) \quad \forall i = 1,2,\dots,n \quad \dots (ii)$$

Después suma de cuadrados de desviaciones (sum square errors)

$$SSE = \sum_{i=1}^{n} [Y_i - \hat{Y}_i]^2 = \sum_{i=1}^{n} [Y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)]^2 \quad \forall i = 1,2,\dots n \quad \dots (iii)$$

Para minimizar se debe derivar las parciales e igualarlas a cero.
$\hat{\beta}_0, \hat{\beta}_1$ tal que minimice SSE

$$\frac{d\,SSE}{d\hat{\beta}_0} = -2 \sum_{i=1}^{n} [Y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)] = 0 \quad \dots (iv)$$

$$\frac{d\,SSE}{d\hat{\beta}_1} = -2 \sum_{i=1}^{n} [Y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)] \cdot X_i = 0 \quad \dots (v)$$

Al resolver las ecuaciones simultáneas se obtiene:

$$n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} X_i = \sum_{i=1}^{n} Y_i \dots (vi) \qquad \hat{\beta}_0 \sum_{i=1}^{n} X_i + \hat{\beta}_1 \sum_{i=1}^{n} X_i^2 = \sum_{i=1}^{n} X_i Y_i \dots (vii)$$

De estas dos ecuaciones se obtiene los parámetros estimados del modelo de regresión lineal simple:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} X_i Y_i - n\bar{X}\bar{Y}}{\sum_{i=1}^{n} X_i^2 - n\bar{X}^2} = \frac{SS_{xy}}{SS_{xx}} \dots (viii) \qquad \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{x} \dots (ix)$$

Luego podemos aproximar la varianza del error mediante el "error estándar de estimación".

$$S_e = \hat{\sigma} = \sqrt{\frac{\sum Y^2 - \hat{\beta}_0 \sum Y - \beta_1 \sum xY}{n-2}}$$

si S_e es pequeño más cerca a la recta de regresión.

* Otra forma de cálculo de la pendiente por Intervalo de confianza es: B_1 de la recta al $(1-\alpha)100\%$.

$$P(\hat{\beta}_1 - t_{(n-2;1-\alpha/2)} \hat{\sigma}_{\hat{\beta}_1} \le \beta_1 \le \hat{\beta}_1 + t_{(n-2;1-\alpha/2)} \hat{\sigma}_{\hat{\beta}_1}) = 1-\alpha$$

VALIDACIÓN DEL MODELO LINEAL

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

for $i = 1, 2, \ldots, n$,

1.Evaluación de la pendiente de la recta $\longrightarrow \beta_1 x_i$

Analytics AoZ

Validación del modelo Lineal

① Hipótesis:

Ho: $B_1 = 0$ (no existe relación lineal entre X e Y) ó (El parámetro B_1 no es significativo)

H1: $B_1 \neq 0$ (existe relación lineal entre X e Y) ó (" " " es significativo)

② Estadístico de prueba:

$$t_c = \frac{\hat{\beta}_1}{\dfrac{\hat{\sigma}_{Se}}{\sqrt{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}}} \qquad \frac{\hat{\sigma}_{Se}}{\sqrt{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}} = \hat{\sigma}_{\hat{\beta}_1}$$

③ Valor crítico: $t_{(n-2;1-\alpha/2)}$

④ Se rechaza Ho si:

$|t_c| > t_{(n-2;1-\alpha/2)}$ o p-valor $< \alpha$

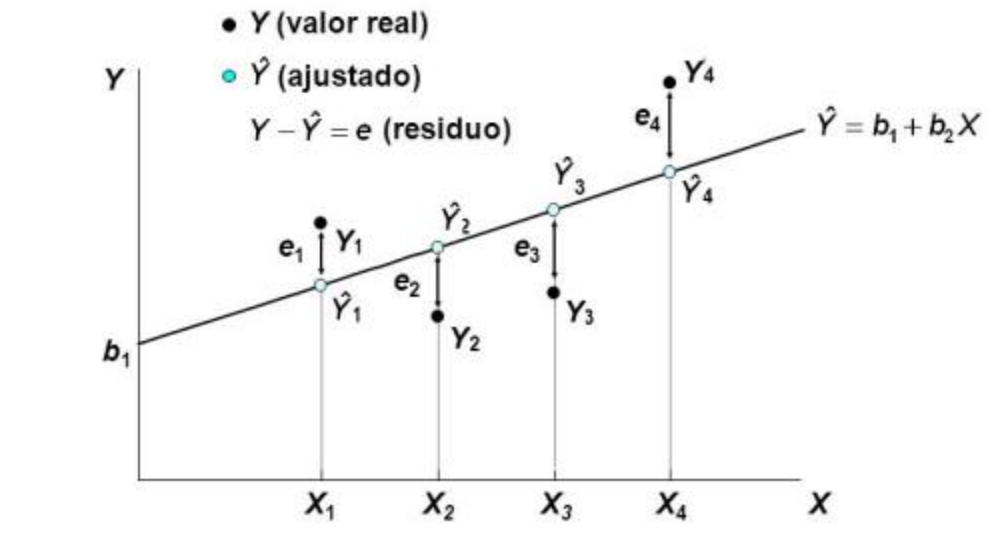
Obs: La misma prueba se puede llevar a cabo con el estadístico F de la tabla ANOVA, en este caso se rechaza Ho si:

$$F_c > F_{(1, n-2, 1-\alpha)}$$

Comparamos la varianza del modelo entre la varianza del residuo

(se puede aplicar regresión lineal)

2.**Prueba F, ANOVA** de una vía donde comparamos la varianza del modelo entre la varianza de los residuos.



Las diferencias entre valores reales y ajustados de *Y* se denominan residuos.

$e_i = y_i - \hat{y}_i$

donde : y_i es el valor obtenido

$\hat{y}_i$ es el valor predicho por la ecuación

La prueba F, será

$$F = \frac{SC\left(\text{Regresión}/b_0\right)/(p-1)}{SC\left(\text{Residual}\right)/(n-p)}$$

3.**Coeficiente de Determinación (r^2):** Indicador de la bondad de ajusta que proporciona la recta de regresión:

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

Indicador de la bondad

r^2 ε [0;1]

$r^2 \rightarrow 0$ la variabilidad observada en Y no explica la relación con x.

$r^2 \rightarrow 1$ la variabilidad observada en Y es explicada en gran parte con x.

Plots Diagnóstico para Análisis de Regresión Lineal

Comprobación de los supuestos del error y gráficos adicionales para Validación del modelo de regresión lineal.

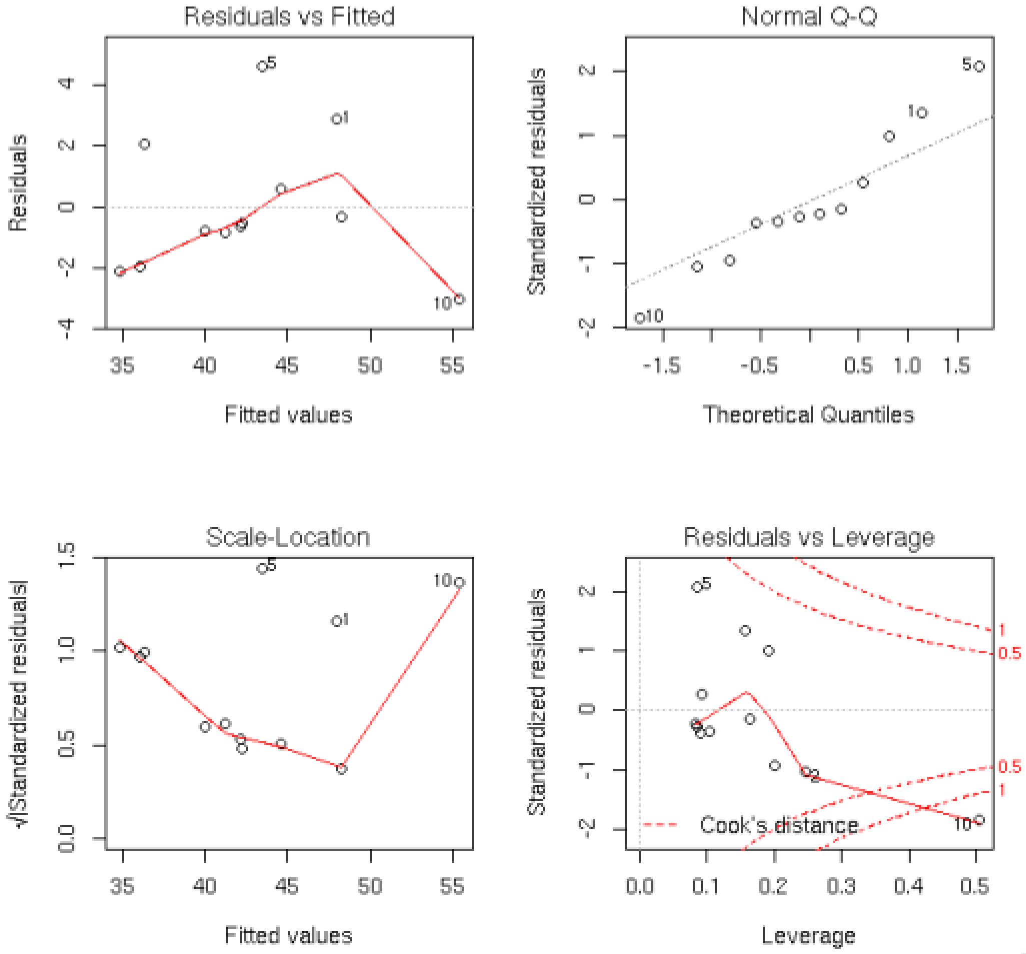
Dentro de los gráficos están:

- Valores predichos frente a los residuos (c)
- Gráfico Q-Q de normalidad (d)
- Valores predichos frente a la raíz cuadrada de los residuos estandarizados. (en valor absoluto). (b)
- Residuos estandarizados frente laverange. (a)

(a) El gráfico se utiliza para detectar puntos con una influencia importante en el cálculo de estimaciones de los parámetros. En caso de detectarse algún punto fuera de los límites que establecen las líneas discontinuas debe estudiarse este punto de forma aislada para detectar, por ejemplo, si esa importancia se debe a un error nuestro.

Los gráficos (b) y (c) se utilizan para contrastar gráficamente independencia, la homocedasticidad y la linealidad de residuos. Idealmente, los residuos deben estar aleatoriamente a lo largo del gráfico, sin formar ningún tipo de patrón.

(d) Busca comparar en un plot los valores de cuántiles muestrales (Ne ejemplo) con los cuantiles esperados bajo la hipótesis de normalidad, por tanto, si la distribución de errores es normal dichos diagramas tendrán a ser rectas que pasan por el origen, en otras palabras, cuanto más se desvíen de esa recta menos normales serán los errores.



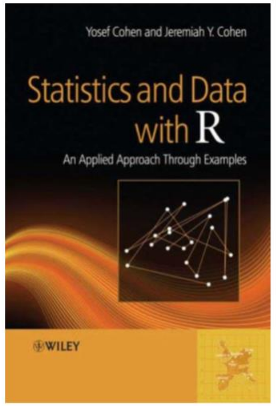
Analytics AoZ

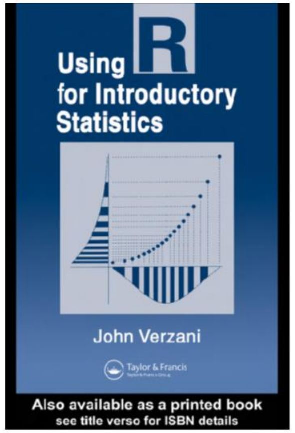
Analytics AoZ

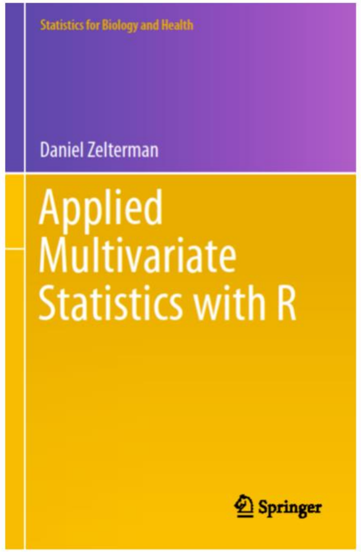
"LO QUE ESCUCHO LO OLVIDO. LO QUE VEO LO RECUERDO. PERO LO QUE HAGO, LO ENTIENDO."

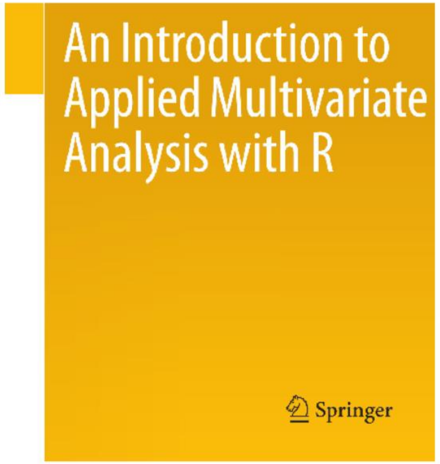
VAMOS A RESOLVER UN EJERCICIO

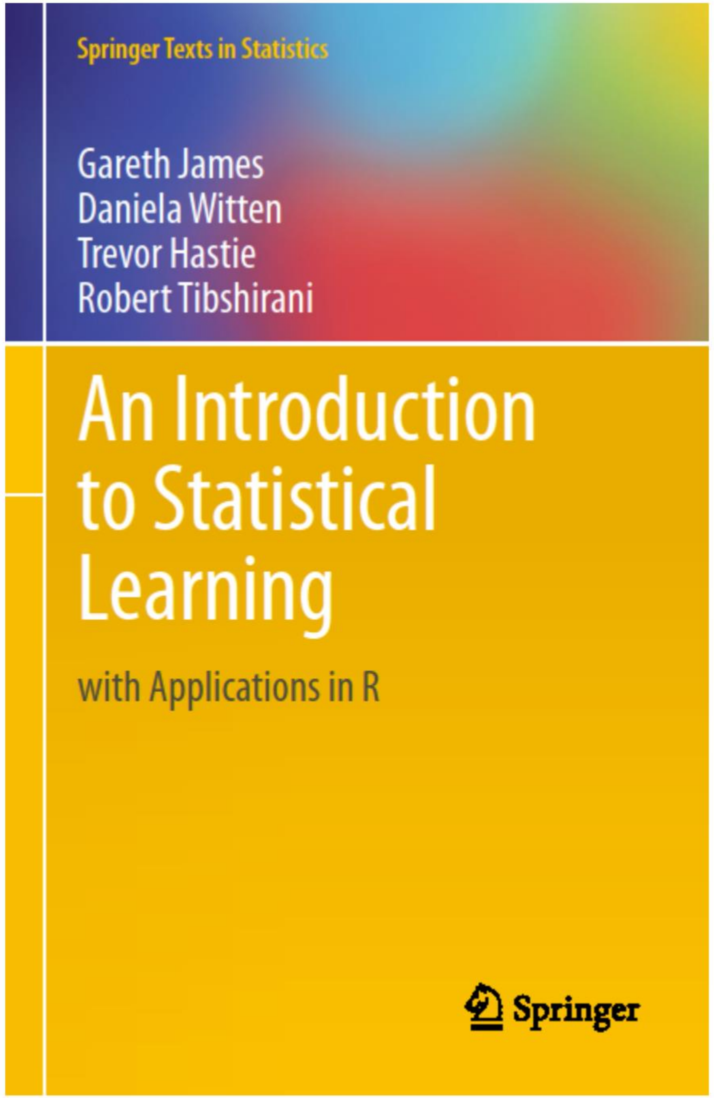
BIBLIOGRAFÍA:

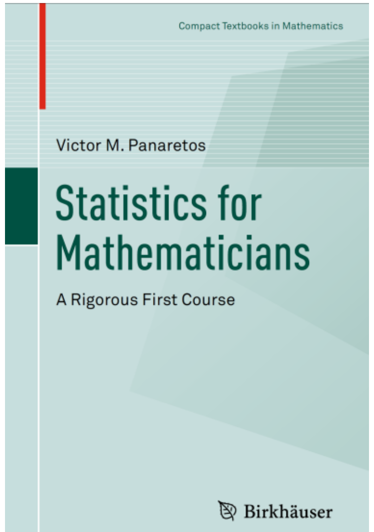


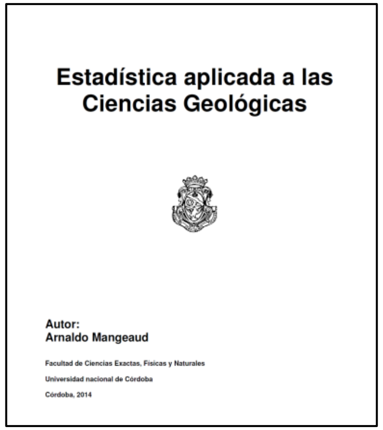


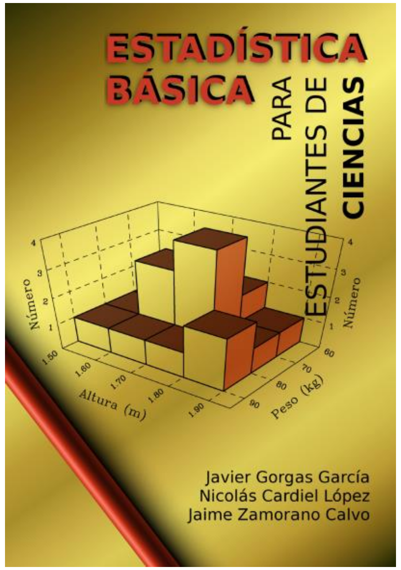












BIBLIOGRAFÍA ESTADÍSTICA APLICADA A GEOLOGÍA EXTRA:

Practical Methods for Data Analysis (US EPA QA/G-9, 2000)

Helsel, D. R., & Hirsch, R. M. (2002). *Statistical methods in water resources* (Vol. 323). Reston, VA: US Geological Survey.

Salvador Figu
eras, M y Gargallo, P. (2003): "Análisis Exploratorio de Datos, 5campus.com, Estadística <http://www.5campus.com/leccion/aed>

Ramalle-Gómara, E., & De Llano, J. A. (2003). Utilización de métodos robustos en la estadística inferencial. *Atención Primaria*, *32*(3), 177-182.

Verzani, J. (2005). *Using R for introductory statistics*. CRC press.

Cohen, Y., & Cohen, J. Y. (2008). *Statistics and Data with R: An applied approach through examples*. John Wiley & Sons.

Arnaldo Mangeaud (2014). *Estadística aplicada a las Ciencias Geológicas*. Universidad nacional de Córdova.

Helsel, D.R., Hirsch, R.M., Ryberg, K.R., Archfield, S.A., and Gilroy, E.J., 2020, *Statistical methods in water resources: U.S. Geological Survey Techniques and Methods*, book 4, chapter A3, 458 p.