Regresión lineal

Luis Fernando Delgado Muñoz - Ing. Agroindustrial, M.Sc

lfdelgadom@unal.edu.co

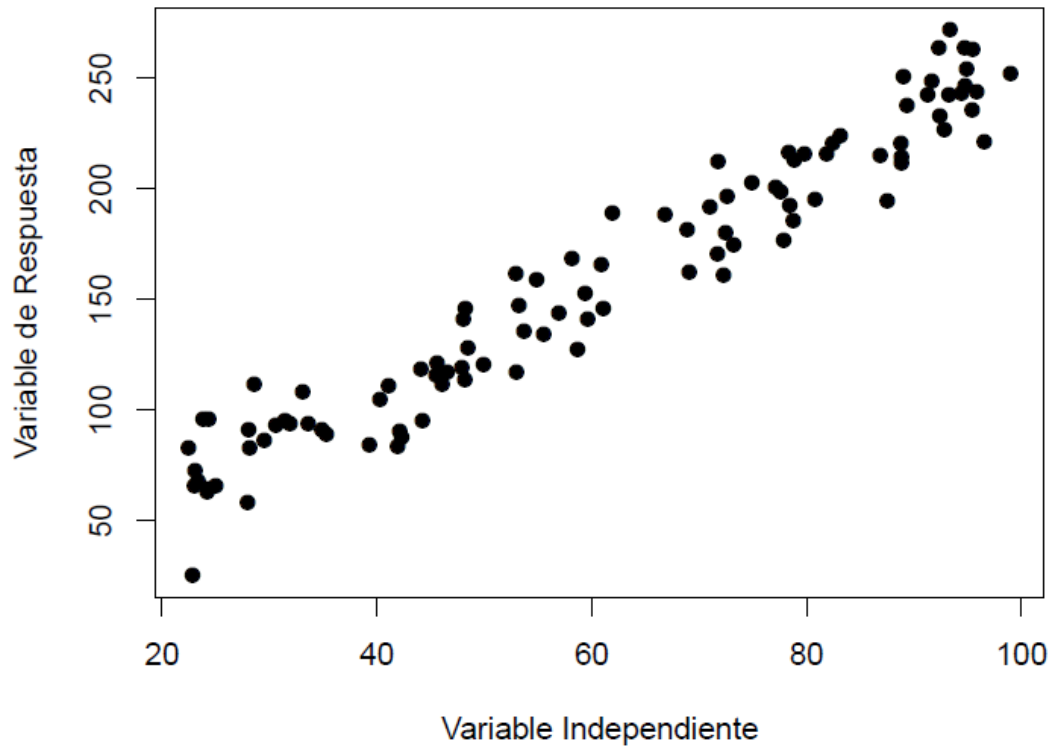
Universidad Nacional de Colombia Facultad de Ingeniería y Administración Departamento de Ingenieria

Correlación lineal

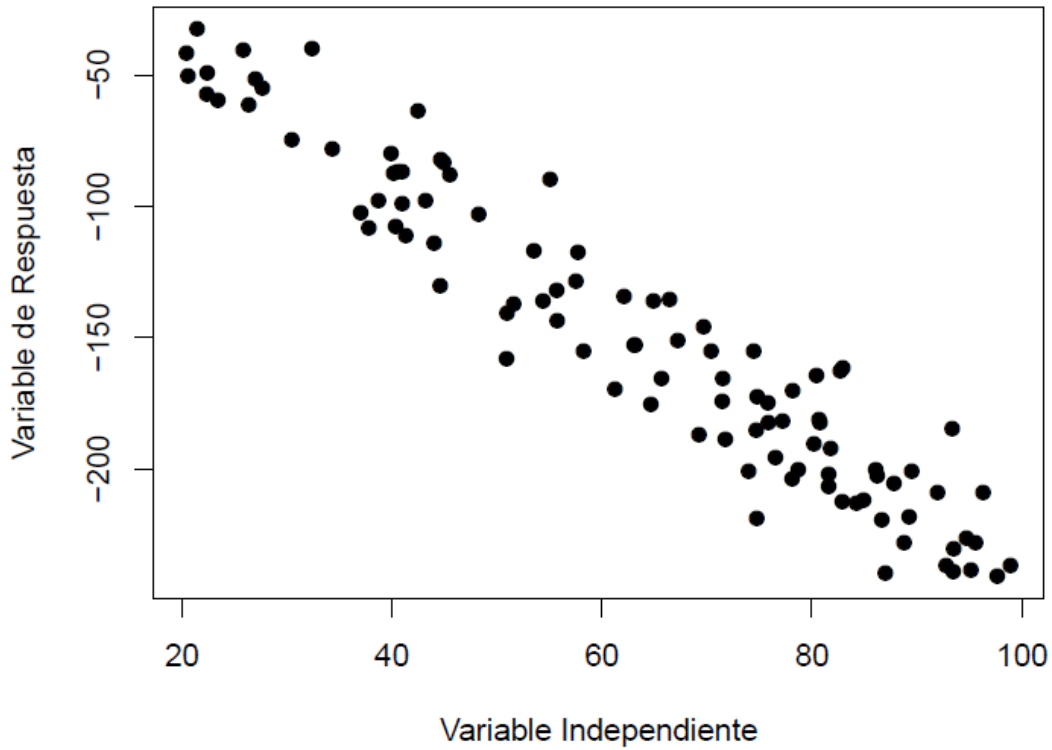
Nos enfocaremos en el estudio de la **Correlación Lineal**.

Esto quiere decir que el análisis de correlación que llevaremos a cabo solamente nos dará información sobre la existencia de una relación de tipo lineal.

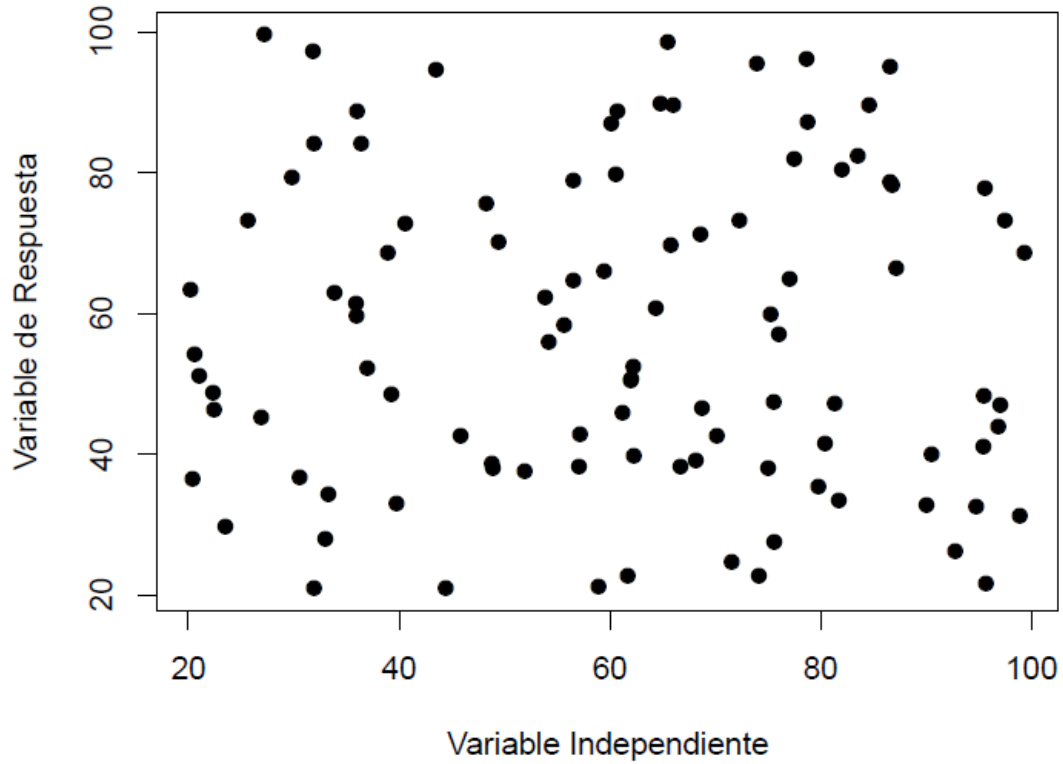
Tipos de Correlación Lineal



Correlación Positiva (Dependencia Positiva)



Correlación Negativa (Dependencia Negativa)

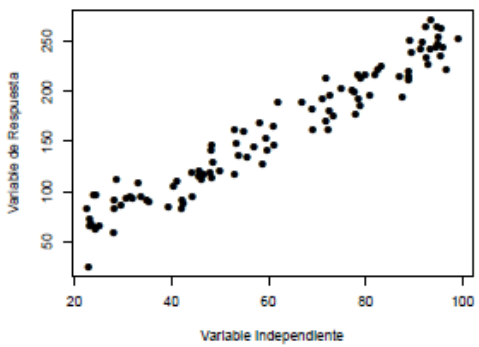


Independencia

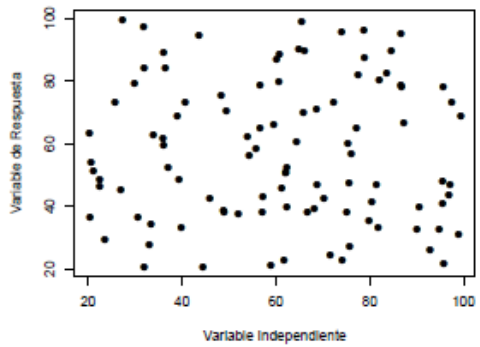
Consideremos la situación donde se quiere estudiar la relación que existe entre dos variables aleatorias llamadas X y Y .

Esta relación se puede examinar mediante el Coeficiente de Correlación el cual denotaremos con la letra (rho).ρ

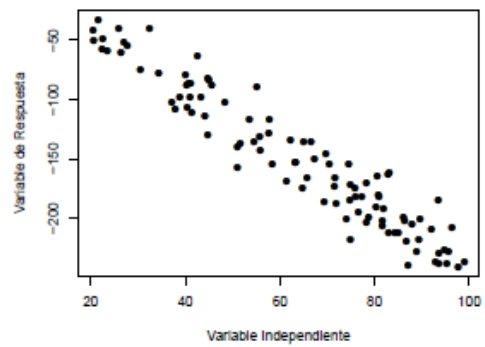
$\rho > 0$



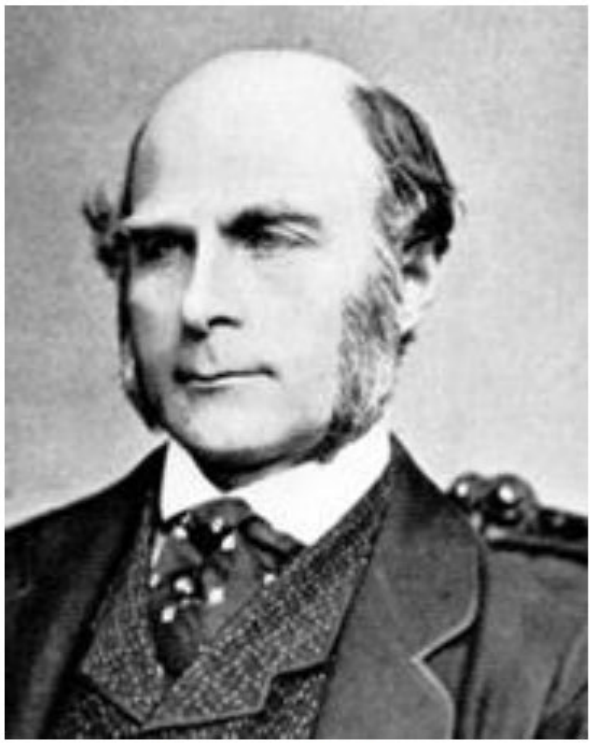
$\rho \approx 0$



$\rho < 0$



La regresión



Francis Galton

El término regresión fue introducido por Francis Galton en su libro Natural inheritance (1889) y fue confirmado por su amigo Karl Pearson. Su trabajo se centró en la descripción de los rasgos físicos de los descendientes (variable Y) a partir de los de sus padres (variable X). Estudiando la altura de padres e hijos a partir de más de mil registros de grupos familiares, se llegó a la conclusión de que los padres muy altos tenían una tendencia a tener hijos que heredaban parte de esta altura, pero que revelaban también una tendencia a regresar a la media.

Modelo Lineal Simple

Nos preocuparemos entonces por encontrar una recta que represente de la mejor forma posible la relación que existe entre dos variables de tal forma que se cometa el menor "error" posible.

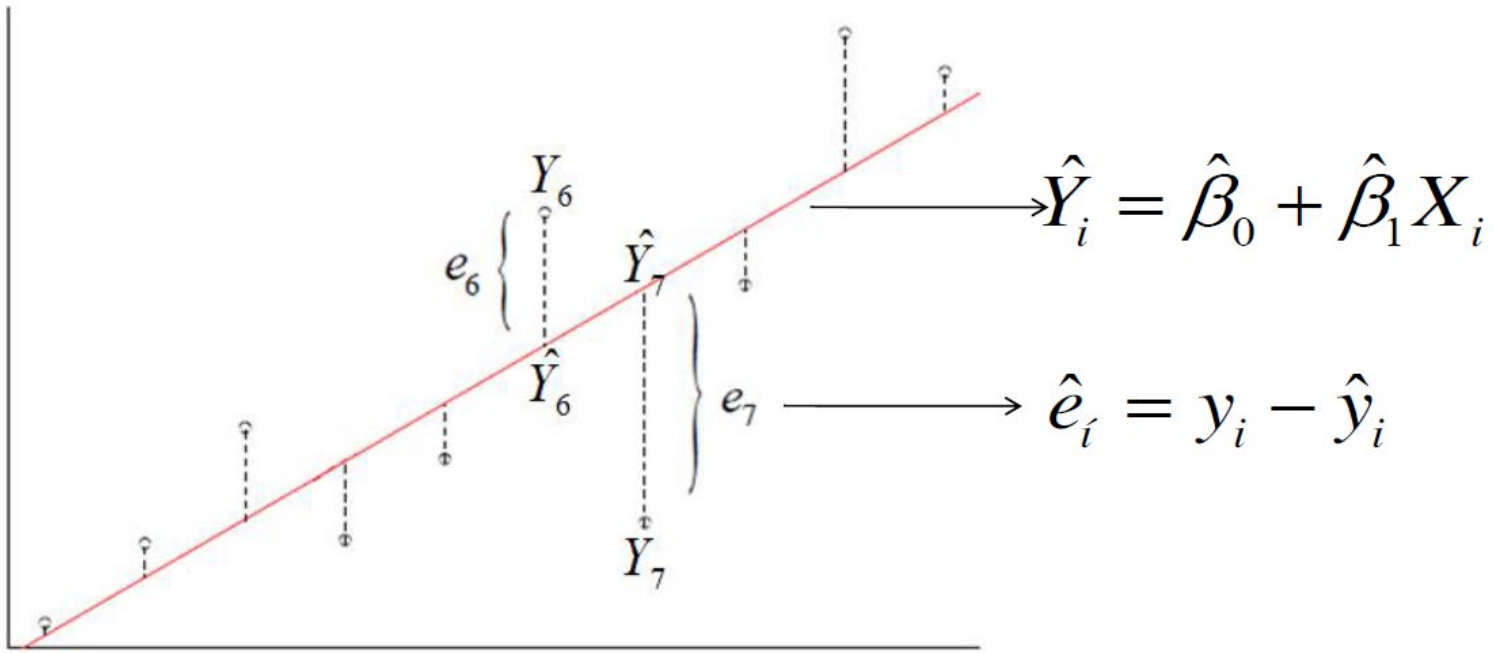
Para esto, plantearemos un modelo de la forma

$$Y = b_0 + b_1 x + e$$

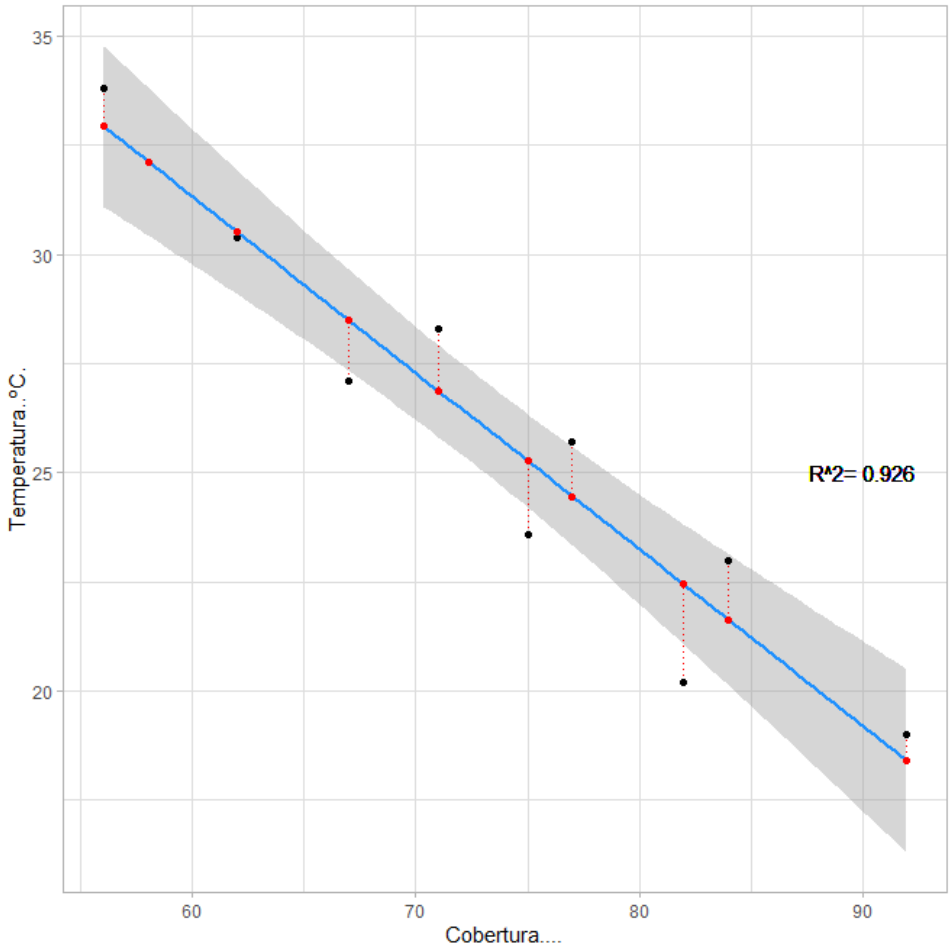
- Y: es la variable de respuesta o variable dependiente.
- X: es una covariable, variable independiente o variable explicativa.
- b_0: es el intercepto de la recta con el eje Y .
- b_1: es la pendiente de la recta.
- e: es una componente de error aleatorio.

Estimación de Parámetros por MCO

Mínimos Cuadrados Ordinarios (MCO): El objetivo de este procedimiento es estimar los parámetros tal que la suma de cuadrados (*SC*) de las diferencias entre las observaciones y la línea recta estimada sea mínima (*Min SCError)*



Estimación de Parámetros por MCO



Validación de Supuestos

Cuando se plantea el modelo de regresión:

$$Y = b_0 + b_1 x + e$$

Se asumen algunos supuestos sobre el termino error aleatorio:

1. Es una variable aleatoria con media cero.

2. Es una variable aleatoria con varianza constante.

3. Son independientes entre si.

4. Siguen una distribución normal.

"Estos supuestos deben cumplirse para abonarle eficiencia al modelo obtenido, por tanto deben ser validados"

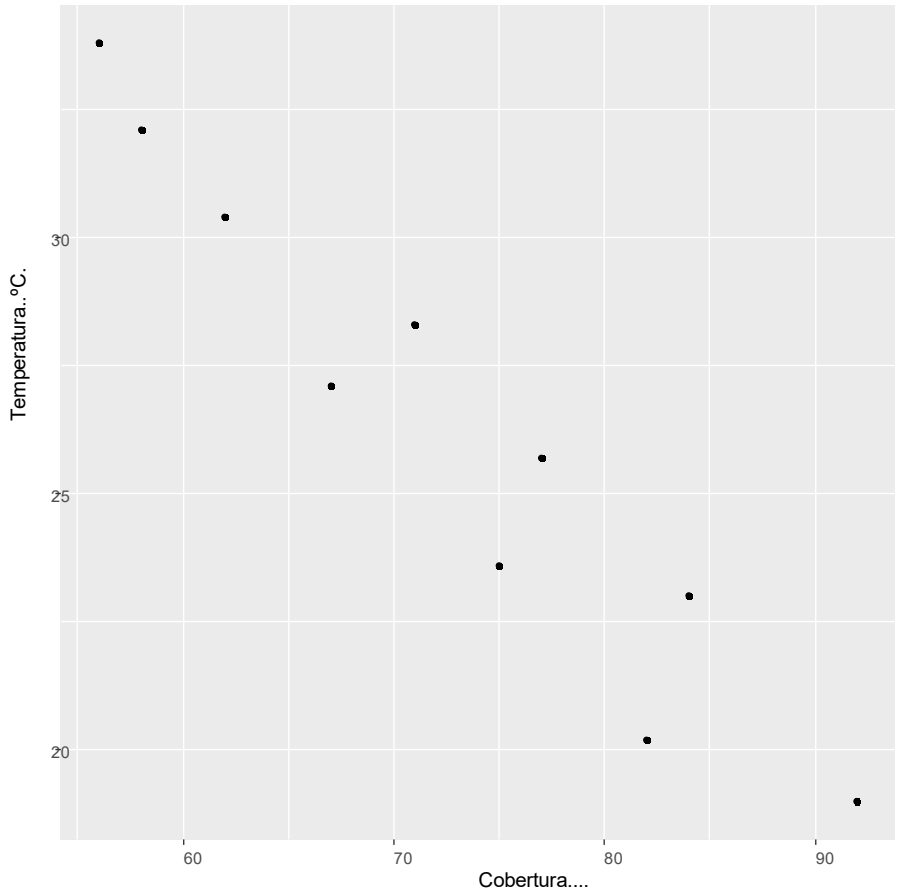
Ejemplo

La cobertura arbórea tiene un efecto marcado sobre la temperatura del suelo. Esta relación es importante en el diseño de sistemas combinados de producción de cultivos leñosos y herbáceos (sistemas agroforestales). Se realizaron mediciones de la temperatura del suelo al mediodía, en verano, a 10 cm de profundidad en sistemas agroforestales con diferente porcentaje de cobertura arbórea. Se analizó la relación entre esas variables mediante regresión lineal.

Datos

Cobertura (%)	Temperatura (ºC)
56	33.8
58	32.1
62	30.4
67	27.1
71	28.3
75	23.6
77	25.7
82	20.2
84	23
92	19

Diagrama de dispersión de los datos



Calculo de variables necesarias para el calculo:

Cobertura (%)	Temperatura (ºC)	xy	x^2	y^2
56	33.8	1892.8	3136	1142.44
58	32.1	1861.8	3364	1030.41
62	30.4	1884.8	3844	924.16
67	27.1	1815.7	4489	734.41
71	28.3	2009.3	5041	800.89
75	23.6	1770	5625	556.96
77	25.7	1978.9	5929	660.49
82	20.2	1656.4	6724	408.04
84	23	1932	7056	529
92	19	1748	8464	361
724	**263.2**	**18549.7**	**53672**	**7147.8**

Calculo de coeficiente de correlación

$$\rho = \frac{n\left(\sum xy\right) - \left(\sum x\right)\left(\sum y\right)}{\sqrt{\left[n\left(\sum x^2\right) - \left(\sum x\right)^2\right]\left[n\left(\sum y^2\right) - \left(\sum y\right)^2\right]}}$$

$$\rho = \frac{10(18549.7) - (724)(263.2)}{\sqrt{[10(53672) - (524176)][10(7147.8) - (69274.24)]}}$$

$$\rho = \frac{-5059.8}{\sqrt{(12544)(2203.76)}}$$

$$\rho = \frac{-5059.8}{5257.75}$$

$$\rho = -0.9623$$

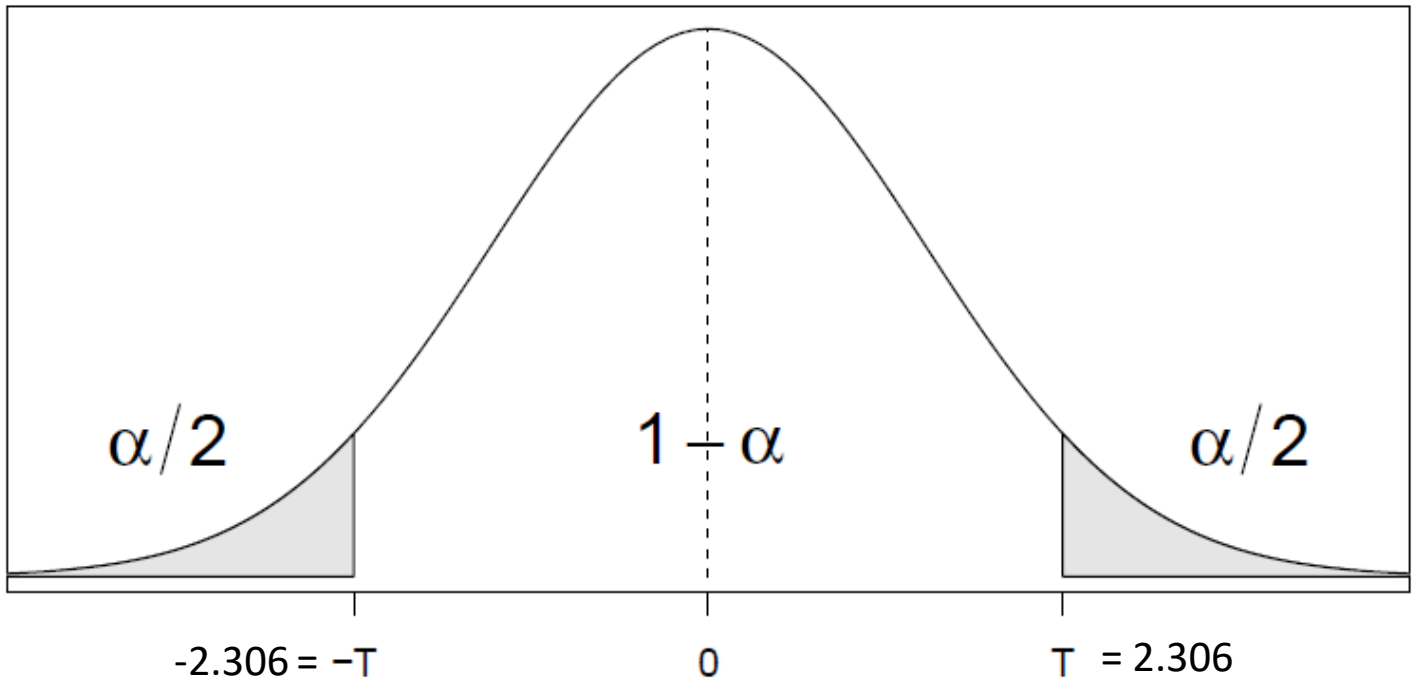
Prueba de significancia del coeficiente de correlación ρ

Planteamiento de hipótesis

$$H_0 : \rho = 0$$

$$H_a : \rho \neq 0$$

Grafica para la hipótesis nula



Gl error: n-2
Gl error: 10-2
Gl error: 8 → Buscamos este valor en la tabla de distribución de t-Student

Regla de decisión

$$Si\ t_{exp} > 2.306 \text{ ó } t_{exp} < -2.306; \text{ no acepto la hipotesis nula}$$

Calculo de t - experimental

$$t_{exp} = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}} \Rightarrow t_{exp} = \frac{-0.9623}{\sqrt{\frac{1-(-0.9623)^2}{10-2}}} \Rightarrow t_{exp} = \frac{-0.9623}{0.09616}$$

$$\Downarrow$$

$$t_{exp} = -9.9107$$

Decisión

$Como$ la $t_{exp} = -9.9107 < -2.306;$

entonces no aceptamos la hipotesis nula : H_0

Calculo de coeficiente de determinación

$$R^2 = \mathrm{r}^2$$

$$\mathrm{R}^2 = (-0.9623)^2$$

$$R^2 = 0.9260$$

El modelo me esta explicando el 92.60% de la variabilidad total

El 92.60% de la variable independiente me esta explicando la variable dependiente.

$$R^2 = \frac{SCreg}{SCtotal}$$

Calculo de coeficientes de regresión b_0 y b_1:

$$b_1 = \frac{n\left(\sum xy\right) - \left(\sum x\right)\left(\sum y\right)}{n\left(\sum x^2\right) - \left(\sum x\right)^2}$$

$$b_1 = \frac{10(18549.7) - (724)(263.2)}{10(53672) - (524176)}$$

$$b_1 = \frac{-5059.8}{12544}$$

$$b_1 = -0.4033$$

Calculo de coeficientes de regresión b_0 y b_1:

$$b_0 = \frac{\left(\sum y\right) - b_1\left(\sum x\right)}{n}$$

$$b_0 = \frac{(263.2) - (-0.4033)(724)}{10}$$

$$b_0 = 55.5189$$

Ecuación de regresión estimada

$$\hat{Y} = b_0 + b_1 x$$

$$\hat{Y} = 55.52 - 0.4033x$$

Análisis de varianza de la regresión

Calculo de sumas de cuadrados de regresión

$$SCreg = \frac{1}{n}\left\{\frac{\left[n\left(\sum xy\right)-\left(\sum x\right)\left(\sum y\right)\right]^2}{n\left(\sum x^2\right)-\left(\sum x\right)^2}\right\}$$

$$SCreg = \frac{1}{10}\left\{\frac{\left[10(18549.7)-(724)(263.2)\right]^2}{10(53672)-(524176)}\right\}$$

$$SCreg = \frac{1}{10}\left\{\frac{25601576.04}{12544}\right\}$$

$$SCreg = 204.09$$

Calculo de sumas de cuadrados totales

$$SCTotal = \sum y^2 - \frac{1}{n}\left(\sum y\right)^2$$

$$SCT = 7147.8 - \frac{1}{10}(69274.24)$$

$$SCT = 220.376$$

Análisis de varianza

Fuentes de variación	Grados de libertad	Suma de cuadrados	Cuadrados medios	F calculada	F tabulada
Modelo	1	204.09	204.09	100.044	5.32
Error	8	16.29	2.04		
Total	9	220.37			

MUCHAS GRACIAS