REGRESIÓN LINEAL SIMPLE Y CORRELACIÓN.

1. Análisis de regresión. Es una técnica estadística para modelar e investigar la relación entre dos o más variables. El análisis de regresión es una colección de herramientas estadísticas para encontrar las estimaciones de los parámetros del modelo de regresión.

2. Diagrama de dispersión. Se trata de una gráfica en la que cada par (x_i , y_i) está representado por un punto graficado en un sistema de coordenadas bidimensionales. Al inspeccionar el diagrama de dispersión se observa que, aun cuando ninguna curva simple pasará exactamente por todos los puntos, hay claros indicios de que los puntos se encuentran dispersos aleatoriamente alrededor de una línea recta.

Se debe completar la tabla de datos con lo siguiente:

x_i	y_i	x_i^2	$x_i y_i$	y_i^2
$\sum_{i=1}^{n} x_i$	$\sum_{i=1}^{n} y_i$	$\sum_{i=1}^{n} x_i^2$	$\sum_{i=1}^{n} x_i y_i$	$\sum_{i=1}^{n} y_i^2$

3. Modelo de regresión lineal simple: $Y = \beta_0 + \beta_1 x + \varepsilon$. β_0 es la ordenada al origen, β_1 es la pendiente y ε es el término del error aleatorio.

4. Coeficientes de regresión. β_0 es la ordenada al origen, β_1 es la pendiente.

5. Promedios: $\bar{x} = \dfrac{\sum_{i=1}^{n} x_i}{n}$ $\bar{y} = \dfrac{\sum_{i=1}^{n} y_i}{n}$

6. Suma de cuadrados de *x*:

$$S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2 \qquad S_{xx} = \sum_{i=1}^{n} x_i^2 - \frac{(\sum_{i=1}^{n} x_i)^2}{n} \qquad S_{xx} = \sum_{i=1}^{n} x_i^2 - n(\bar{x})^2$$

7. Suma de cuadrados de *y* (Suma de cuadrados total):

$$S_{yy} = \sum_{i=1}^{n}(y_i - \bar{y})^2 \qquad S_{yy} = \sum_{i=1}^{n} y_i^2 - \frac{(\sum_{i=1}^{n} y_i)^2}{n} \qquad S_{yy} = \sum_{i=1}^{n} y_i^2 - n(\bar{y})^2$$

8. Suma de cuadrados de los productos de *x* y:

$$S_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) \qquad S_{xy} = \sum_{i=1}^{n} x_i y_i - \frac{\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n} \qquad S_{xy} = \sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}$$

9. Varianza de las variables: $S_x^2 = \dfrac{S_{xx}}{n-1}$ $S_y^2 = \dfrac{S_{yy}}{n-1}$

10. Covarianza: $Cov(x,y) = \dfrac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n}$ $Cov(x,y) = \dfrac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n^2}$

$$Cov(x,y) = \frac{\sum_{i=1}^{n} x_i y_i}{n} - \bar{x}\bar{y} \qquad Cov(x,y) = \frac{S_{xy}}{n}$$

11. Suma de cuadrados.

Regresión: $SC_R = \sum_{i=1}^{n}(\hat{y} - \bar{y})^2$ $SC_R = \dfrac{(S_{xy})^2}{S_{xx}}$

Error: $SC_E = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ $SC_E = S_{yy} - \dfrac{(S_{xy})^2}{S_{xx}}$ $SC_E = (1 - r^2) S_{yy}$

$$SC_E = \sum_{i=1}^{n} y_i^2 - b_0 \sum_{i=1}^{n} y_i - b_1 \sum_{i=1}^{n} x_i y_i$$

$$SC_E = \frac{\left[n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2 \right] \times \left[n\sum_{i=1}^{n} y_i^2 - (\sum_{i=1}^{n} y_i)^2 \right] - (n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i)^2}{n\left[n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2 \right]}$$

Total: $SC_T = \sum_{i=1}^{n}(y_i - \bar{y})^2$ $SC_T = S_{yy}$

12. Cuadrado medio.

Regresión: $CM_R = \dfrac{(S_{xy})^2}{S_{xx}}$ $CM_R = SC_R$

Error: $CM_E = \dfrac{SC_E}{n-2}$ $CM_E = \dfrac{S_{yy} - \frac{(S_{xy})^2}{S_{xx}}}{n-2}$

13. Recta de regresión ajustada o estimada: $\hat{y} = b_0 + b_1 x$

14. Ecuaciones normales de mínimos cuadrados.

$$b_0 \sum_{i=1}^{n} x_i + b_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i \qquad b_0 n + b_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i$$

15. Coeficientes de la regresión:

15-1. Pendiente: $b_1 = \dfrac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2}$ $b_1 = \dfrac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n(\bar{x})^2}$ $b_1 = \dfrac{S_{xy}}{S_{xx}}$

$$b_1 = \frac{SC_R}{S_{xx}} \qquad b_1 = \frac{CM_R}{S_{xx}} \qquad b_1 = r\left(\frac{S_y}{S_x}\right) \qquad b_1 = r\sqrt{\frac{S_{yy}}{S_{xx}}}$$

15-2. Ordenada al origen. $b_0 = \dfrac{\sum_{i=1}^{n} x_i^2 \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} x_i y_i \sum_{i=1}^{n} x_i}{n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2}$ $b_0 = \bar{y} - b_1\bar{x}$

Estos cálculos son extremadamente sensibles a la aproximación. Esto es especialmente cierto para el cálculo del coeficiente de determinación. Por tanto, se aconseja en aras de la exactitud, efectuar los cálculos hasta con cinco o seis cifras decimales.

16. Error estándar de estimación.

$$Se = \sqrt{CME} \qquad Se = \sqrt{\frac{SC_E}{n-2}} \qquad Se = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-2}} \qquad Se = \sqrt{\frac{S_{yy} - \frac{(S_{xy})^2}{S_{xx}}}{n-2}}$$

$$Se = \sqrt{\frac{S_{yy} - b_1 S_{xy}}{n-2}} \qquad Se = \sqrt{\frac{\sum_{i=1}^{n} y_i^2 - b_0 \sum_{i=1}^{n} y_i - b_1 \sum_{i=1}^{n} x_i y_i}{n-2}}$$

$$Se=\sqrt{\frac{n\sum_{i=1}^{n}y_i^2-(\sum_{i=1}^{n}y_i)^2}{n(n-2)}-\frac{\left[n\sum_{i=1}^{n}x_iy_i-\sum_{i=1}^{n}x_i\sum_{i=1}^{n}y_i\right]^2}{n(n-2)\left[n\sum_{i=1}^{n}x_i^2-(\sum_{i=1}^{n}x_i)^2\right]}}$$

17. Coeficiente de correlación.

$$r=\frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} \qquad r=\sqrt{\frac{SC_R}{S_{yy}}} \qquad r=\sqrt{\frac{CM_R}{S_{yy}}}$$

$$r=\sqrt{\frac{SC_R}{SC_T}} \qquad r=\sqrt{\frac{CM_R}{SC_T}} \qquad r=\frac{S_{xy}}{(n-1)S_xS_y} \qquad r=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{(n-1)S_xS_y}$$

$$r=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2\times\sum_{i=1}^{n}(y_i-\bar{y})^2}} \qquad r=\frac{n\sum_{i=1}^{n}x_iy_i-\sum_{i=1}^{n}x_i\sum_{i=1}^{n}y_i}{\sqrt{n\sum_{i=1}^{n}x_i^2-(\sum_{i=1}^{n}x_i)^2}\cdot\sqrt{n\sum_{i=1}^{n}y_i^2-(\sum_{i=1}^{n}y_i)^2}}$$

18. Coeficiente de determinación.

$$r^2=\frac{(S_{xy})^2}{S_{xx}S_{yy}} \qquad r^2=\frac{SC_R}{S_{yy}} \qquad r^2=\frac{CM_R}{S_{yy}}$$

$$r^2=\frac{SC_R}{SC_T} \qquad r^2=\frac{CM_R}{SC_T} \qquad r=\frac{(S_{xy})^2}{(n-1)^2S_x^2S_y^2} \qquad r=\frac{\left[\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})\right]^2}{(n-1)^2S_x^2S_y^2}$$

$$r^2=\frac{b_0\sum_{i=1}^{n}y_i+b_1\sum_{i=1}^{n}x_iy_i-n(\bar{y})^2}{\sum_{i=1}^{n}y_i^2-n(\bar{y})^2} \qquad r^2=\frac{\left(n\sum_{i=1}^{n}x_iy_i-\sum_{i=1}^{n}x_i\sum_{i=1}^{n}y_i\right)^2}{\left[n\sum_{i=1}^{n}x_i^2-(\sum_{i=1}^{n}x_i)^2\right]\cdot\left[n\sum_{i=1}^{n}y_i^2-(\sum_{i=1}^{n}y_i)^2\right]}$$

Variación explicada: $VE=r^2S_{yy}$ **Variación sin explicar:** $VE=(1-r^2)S_{yy}$

19. Coeficiente de determinación corregido.

$$\bar{r}^2=1-(1-r^2)\times\frac{n-1}{n-2}$$

20. Error estándar estimado de la ordenada al origen:

$$s_{b_0}=Se\sqrt{\frac{1}{n}+\frac{\bar{x}^2}{S_{xx}}} \qquad s_{b_0}=Se\sqrt{\frac{\sum x_i^2}{nS_{xx}}} \qquad s_{b_0}=Se\sqrt{\frac{\sum x_i^2}{n\sum_{i=1}^{n}x_i^2-(\sum_{i=1}^{n}x_i)^2}}$$

21. Error estándar estimado de la pendiente:

$$s_{b_1}=\frac{Se}{\sqrt{S_{xx}}} \qquad s_{b_1}=Se\sqrt{\frac{n}{n\sum_{i=1}^{n}x_i^2-(\sum_{i=1}^{n}x_i)^2}}$$

22. Error estándar del coeficiente de correlación.

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

23. Intervalos de confianza en el análisis de regresión.

23-1. Intervalo de confianza de $(1-\alpha)$% para la **ordenada al origen** β_0.

$$b_0-t_{\alpha/2,n-2}\,Se\sqrt{\frac{1}{n}+\frac{\bar{x}^2}{S_{xx}}}<\beta_0<b_0+t_{\alpha/2,n-2}\,Se\sqrt{\frac{1}{n}+\frac{\bar{x}^2}{S_{xx}}}$$

23-2. Intervalo de confianza de $(1-\alpha)$% para la **pendiente** β_1.

$$b_1-t_{\alpha/2,n-2}\frac{Se}{\sqrt{S_{xx}}}<\beta_1<b_1+t_{\alpha/2,n-2}\frac{Se}{\sqrt{S_{xx}}}$$

23-3. Intervalo de confianza de $(1-\alpha)$% alrededor de la **respuesta media** en el valor de $x=x_0$.

$$\hat{\mu}_{Y|_{x_0}}-t_{\alpha/2,n-2}\,Se\sqrt{\frac{1}{n}+\frac{(x_0-\bar{x})^2}{S_{xx}}}<y_0<\hat{\mu}_{Y|_{x_0}}+t_{\alpha/2,n-2}\,Se\sqrt{\frac{1}{n}+\frac{(x_0-\bar{x})^2}{S_{xx}}}$$

Donde $\hat{\mu}_{Y|_{x_0}}=\bar{y}+b_1(x-\bar{x})=\hat{y}$

23-4. Intervalo de confianza de $(1-\alpha)$% para una **observación futura** y_0 (Intervalo de predicción).

$$\hat{y}_0-t_{\alpha/2,n-2}\,Se\sqrt{1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{S_{xx}}}<y_0<\hat{y}_0+t_{\alpha/2,n-2}\,Se\sqrt{1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{S_{xx}}}$$

El valor $\hat{y}_0$ se calcula a partir del modelo de regresión $\hat{y}=b_0+b_1x_0$.

Intervalo de confianza de $(1-\alpha)$% para el **coeficiente de correlación poblacional** ρ.

$$\tanh\left(\tanh^{-1}r-\frac{z_{\alpha/2}}{\sqrt{n-3}}\right)<\rho<\tanh\left(\tanh^{-1}r+\frac{z_{\alpha/2}}{\sqrt{n-3}}\right)$$

24. Pruebas para los parámetros poblacionales.

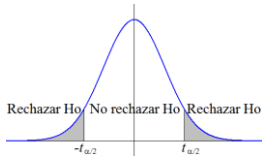
24-1. Prueba para β_0.

Hipótesis nula. Hipótesis alternativa.

H_0: $\beta_0=\beta_{0,0}$ H_1: $\beta_0\neq\beta_{0,0}$ (Bilateral)

Nivel de significancia: α.

Regla de decisión.



Rechazar H_0 si $t<-t_{\alpha/2,\,n-2}$ ó $t>t_{\alpha/2,\,n-2}$ No rechazar H_0 si $-t_{\alpha/2,\,n-2}<t<t_{\alpha/2,\,n-2}$

Estadístico de prueba: $t=\frac{b_0-\beta_{0,0}}{s_{b_0}}$, $s_{b_0}=Se\sqrt{\frac{1}{n}+\frac{\bar{x}^2}{S_{xx}}}$ y tiene $n-2$ grados de libertad.

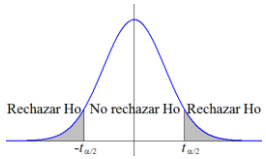
24-2. Prueba para β_1 (Prueba de significancia del modelo).

Hipótesis nula. Hipótesis alternativa.

H_0: $\beta_1=\beta_{1,0}$ H_1: $\beta_1\neq\beta_{1,0}$ (Bilateral)

Nivel de significancia: α.

Regla de decisión.



Rechazar H_0 si $t<-t_{\alpha/2,\,n-2}$ ó $t>t_{\alpha/2,\,n-2}$ No rechazar H_0 si $-t_{\alpha/2,\,n-2}<t<t_{\alpha/2,\,n-2}$

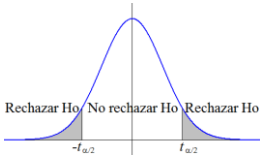
Estadístico de prueba: $t=\frac{b_1-\beta_{1,0}}{s_{b_1}}$, $s_{b_1}=\frac{Se}{\sqrt{S_{xx}}}$ y tiene $n-2$ grados de libertad.

24-3. Prueba para el coeficiente de correlación poblacional $\rho\neq 0$.

Hipótesis nula. Hipótesis alternativa.

H_0: $\rho=0$ H_1: $\rho\neq 0$ (Bilateral)

Regla de decisión.

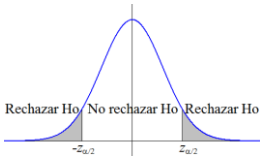


Rechazar H_0 si $t < -t_{\alpha/2,\, n-2}$ ó $t > t_{\alpha/2,\, n-2}$ No rechazar H_0 si $-t_{\alpha/2,\, n-2} < t < t_{\alpha/2,\, n-2}$

Estadístico de prueba: $t = \dfrac{r}{s_r}$, $s_r = \sqrt{\dfrac{1-r^2}{n-2}}$ y tiene $n-2$ grados de libertad.

24-4. Prueba para el coeficiente de correlación poblacional $\rho \neq \rho_0$.

Hipótesis nula. Hipótesis alternativa.
H_0: $\rho = \rho_0$ H_1: $\rho \neq \rho_0$ (Bilateral)
Nivel de significancia: α.
Regla de decisión.



Rechazar H_0 si $z < -z_{\alpha/2}$ ó $z > z_{\alpha/2}$ No rechazar H_0 si $-z_{\alpha/2} < z < z_{\alpha/2}$

Estadístico de prueba: $z = (\tanh^{-1} r - \tanh^{-1} \rho_0)\sqrt{n-3}$

25. Enfoque del Análisis de Varianza (ANOVA) para probar la significación de una regresión.

25-1.- Planteamiento de las hipótesis.

Hipótesis nula. Hipótesis alternativa.
H_0: $\beta_1 = 0$ H_1: $\beta_1 \neq 0$

25-2.- Nivel de significancia.

El nivel de significancia para la prueba es α.

25-3.- Regla de decisión.

No rechazar H_0 si $F < F_{\alpha,\, 1,\, n-2}$ Rechazar H_0 si $F > F_{\alpha,\, 1,\, n-2}$

25-4.- Estadístico de prueba.

$$F = \frac{CM_R}{CM_E}$$

CM_R = Cuadrado medio de la regresión.
CM_E = Cuadrado medio del error.
SC_T = Suma total corregida de los cuadrados o Suma de cuadrados total.
SC_R = Suma de los cuadrados de la regresión.
SC_E = Suma de los cuadrados de errores.

Fuente de Variación	Suma de cuadrados	Grados de libertad	Cuadrado medio	F
Regresión	SC_R	1	CM_R	CM_R/CM_E
Error	$SC_E = SC_T - SC_R$	$n-2$	CM_E	
Total	SC_T	$n-1$		

Fuente de Variación	Suma de cuadrados	Grados de libertad	Cuadrado medio	F
Regresión	$\sum_{i=1}^{n}(\hat{y}-\bar{y})^2$	1	$\sum_{i=1}^{n}(\hat{y}-\bar{y})^2$	CM_R/CM_E
Error	$\sum_{i=1}^{n}(y_i-\hat{y}_i)^2$	$n-2$	$\dfrac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{n-2}$	
Total	$\sum_{i=1}^{n}(y_i-\bar{y})^2$	$n-1$		

Fuente de Variación	Suma de cuadrados	Grados de libertad	Cuadrado medio	F
Regresión	$\dfrac{(S_{xy})^2}{S_{xx}}$	1	$\dfrac{(S_{xy})^2}{S_{xx}}$	CM_R/CM_E
Error	$S_{yy} - \dfrac{(S_{xy})^2}{S_{xx}}$	$n-2$	$\dfrac{S_{yy} - \dfrac{(S_{xy})^2}{S_{xx}}}{n-2}$	
Total	S_{yy}	$n-1$		

El procedimiento de análisis de varianza para probar la significación de la regresión es equivalente a la prueba t de la sección 23-2. Es decir, cualquiera de los dos procedimientos llevará a las mismas conclusiones. Se cumple, en general, que el cuadrado de una variable aleatoria t con v grados de libertad es una variable aleatoria F, con 1 y v grados de libertad en el numerador y el denominador, respectivamente. Por lo tanto, la prueba utilizando t es equivalente a la prueba basada en F. Sin embargo, cabe observar que la prueba t es un tanto más flexible por cuanto permitiría hacer la prueba contra una hipótesis alternativa de una cola, en tanto que la prueba F está restringida a una hipótesis alternativa de dos colas.

26.- Suma de cuadrados.

Regresión.

$$SC_R = \sum_{i=1}^{n}(\hat{y}-\bar{y})^2 \qquad SC_R = \frac{(S_{xy})^2}{S_{xx}} \qquad SC_R = b_1 S_{xy}$$

Error.

$$SC_E = \sum_{i=1}^{n}(y_i-\hat{y}_i)^2 \qquad SC_E = S_{yy} - \frac{(S_{xy})^2}{S_{xx}} \qquad SC_E = (1-r^2)S_{yy}$$

$$SC_E = \sum_{i=1}^{n} y_i^2 - b_0\sum_{i=1}^{n} y_i - b_1\sum_{i=1}^{n} x_i y_i$$

$$SC_E = \frac{\left[n\sum_{i=1}^{n}x_i^2 - (\sum_{i=1}^{n}x_i)^2\right]\times\left[n\sum_{i=1}^{n}y_i^2 - (\sum_{i=1}^{n}y_i)^2\right] - (n\sum_{i=1}^{n}x_i y_i - \sum_{i=1}^{n}x_i\sum_{i=1}^{n}y_i)^2}{n\left[n\sum_{i=1}^{n}x_i^2 - (\sum_{i=1}^{n}x_i)^2\right]}$$

$$SC_E = SC_T - SC_R$$

Total.

$$SC_T = \sum_{i=1}^{n}(y_i-\bar{y})^2 \qquad SC_T = \sum_{i=1}^{n}y_i^2 - \frac{(\sum_{i=1}^{n}y_i)^2}{n} \qquad SC_T = \sum_{i=1}^{n}y_i^2 - n(\bar{y})^2 \qquad SC_T = S_{yy}$$

26.1. Grados de libertad.

Regresión: 1 Error: $n-2$ Total: $n-1$

26.2 Cuadrado medio.

Regresión: $CM_R = SC_R$ Error: $CM_E = \dfrac{SC_E}{n-2}$

26.3 Coeficiente de determinación.

$$R^2 = \frac{SC_R}{SC_T}$$

Autor: **MSc. Ing. Willians Medina**.
Teléfono / Whatsapp: **+58-424-9744352**
e-mail: **medinawj@gmail.com**
Twitter: **@medinawj**

El presente formulario está disponible en formato digital en la siguiente dirección:
https://www.tutoruniversitario.com/
Puerto La Cruz, abril de 2025.