# ANALISIS DE REGRESION SIMPLE

**Dr. Porfirio Gutiérrez González**

# Regresión Lineal

En la búsqueda de mejoras o en la solución de problemas es necesario, frecuentemente, investigar la relación entre factores (o variables). Para lo cual existen varias herramientas estadísticas, entre las que se encuentran el diagrama de dispersión, el análisis de correlación y el análisis de regresión.

El análisis de regresión puede usarse para explicar la relación de un factor con otro(s). Para ello, son necesarios los datos, y estos pueden obtenerse de experimentos planeados, de observaciones de fenómenos no controlados o de registros históricos.

# Regresión lineal simple

Sean dos variables **X** y **Y**. Supongamos que se quiere explicar el comportamiento de **Y** con el de **X**. Para esto, se mide el valor de **Y** sobre un conjunto de **n** valores de **X**, con lo que se obtienen **n** parejas de puntos $(X_1, Y_1)$, $(X_2, Y_2)$,...,$(X_n, Y_n)$.

A **Y** se le llama la variable dependiente o la variable de respuesta y a **X** se le conoce como variable independiente o variable regresora.

# Regresión lineal simple

Supongamos que las variables **X** y **Y** están relacionadas linealmente y que para cada valor de **X**, **Y** es una variable aleatoria. Es decir, supongamos que cada observación de **Y** puede ser descrita por el modelo:

$$Y=ß_0 +ß_1X+e$$

donde **e** es un error aleatorio con media cero y varianza $\sigma^2$ y es de suponerse que los errores no están correlacionados, lo que significa que el valor de un error no depende del valor de cualquier otro error.

# ✓ Estimación de los parámetros $ß_0$ y $ß_1$

Los parámetros $ß_0$ y $ß_1$ son desconocidos y se deben de estimar con los datos de la muestra.

Para estimar $ß_0$ y $ß_1$ se usa el método de mínimos cuadrados

$$y_i = \beta_0 + \beta_1 x_1 + \varepsilon_i, \quad \text{i=1,2….n}$$

Se puede considerar que la ecuación anterior es un **modelo de regresión,** escritos en términos de los $n$ pares de datos $(y_i, x_i)$ $(i = 1,2, \cdots, n)$. Así el criterio de mínimos cuadrados es

$$S(\beta_0, \beta_1) = \sum_{I=1}^{N} (y_i - \beta_0 - \beta_1 x_i)^2$$

Los estimadores, por mínimos cuadrados, de $\beta_0$ y $\beta_1$, que se designaran por $\hat{\beta}_0$y $\hat{\beta}_1$, deben satisfacer

$$\left.\frac{\partial S}{\partial \beta_0}\right|_{\beta_0,\beta_1} = -2\sum_{i=1}^{n}\left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_1\right) = 0$$

$$\left.\frac{\partial S}{\partial \beta_1}\right|_{\beta_0,\beta_1} = -2\sum_{i=1}^{n}\left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_1\right)x_1 = 0$$

Se simplifican estas dos ecuaciones y se obtiene

$$n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i$$

$$\hat{\beta}_0 \sum_{i=1}^{n} x_i + \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i$$

Estas ecuaciones son llamadas **ecuaciones normales de mínimos cuadrados.** Su solución es la siguiente:

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i - \frac{\left(\sum_{i=1}^{n} y_i\right)\left(\sum_{i=1}^{n} x_i\right)}{n}}{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}}$$

En donde

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \quad \text{y} \quad \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Además

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

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

De esta forma

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}$$

La diferencia entre el valor observado $y_i$ y el valor ajustado correspondiente $\hat{y}_i$ se llama **residual**. Matemáticamente, el i-ésimo residual es

$$e_i = y_i - \hat{y}_i = y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 x_1\right),$$

Los residuales tienen un papel importante para investigar la **adecuación** del modelo de regresión ajustado, y para detectar diferencias respecto a las hipótesis básicas.

**Ejemplo:** Datos del propelente

Un motor de cohete se forma pegando entre si un propelente de ignición y un propelente de sostenimiento dentro de una caja metálica. La resistencia al corte de la pegadura entre los dos propelentes es una característica importante de calidad. Se cree que la resistencia al corte se relaciona con la edad, en semanas, del lote del propelente de sostenimiento. Se hicieron 20 observaciones de resistencia al corte y la edad del lote correspondiente de propelente, y se ven en la siguiente tabla

| observacion | y | x |
|---|---|---|
| 1 | 2158.7 | 15.5 |
| 2 | 1678.15 | 23.75 |
| 3 | 2316 | 8 |
| 4 | 2061.3 | 17 |
| 5 | 2207.5 | 5.5 |
| 6 | 1708.3 | 19 |
| 7 | 1784.7 | 24 |
| 8 | 2575 | 2.5 |
| 9 | 2357.9 | 7.5 |
| 10 | 2256.7 | 11 |
| 11 | 2165.2 | 13 |
| 12 | 2399.55 | 3.75 |
| 13 | 1779.8 | 25 |
| 14 | 2336.75 | 9.75 |
| 15 | 1765.3 | 22 |
| 16 | 2053.5 | 18 |
| 17 | 2414.4 | 6 |
| 18 | 2200.5 | 12.5 |
| 19 | 2654.2 | 2 |
| 20 | 1753.7 | 21.5 |

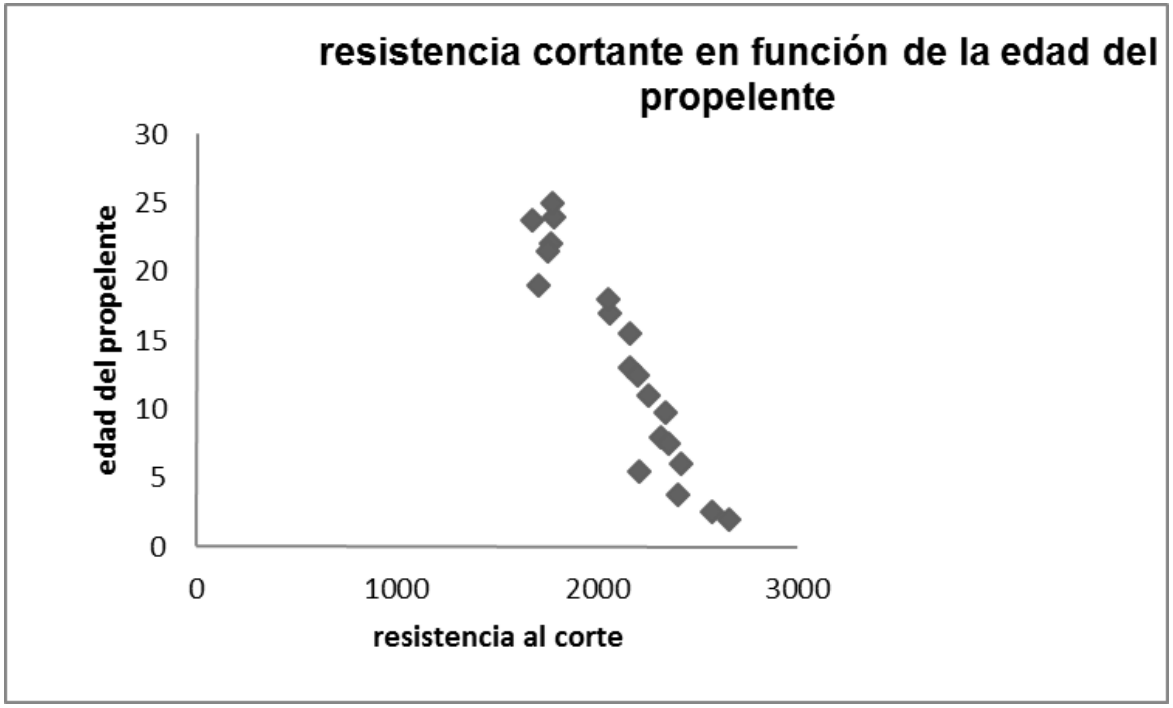
$$S_{xx} = \sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n} = 4677.69 - \frac{71422.56}{20} = 1106.56$$

$$S_{xy} = \sum_{i=1}^{n} y_i x_i - \frac{\left(\sum_{i=1}^{n} y_i\right)\left(\sum_{i=1}^{n} x_i\right)}{n} = 528492.64 - \frac{(267.25)(42627.15)}{20} = -41112.65$$

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{-41112.65}{1106.56} = -37.15$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 2131.3575 - (-37.15)13.3625 = 2627.82$$

$$\hat{y} = 2627.82 - 37.15x$$


resistencia cortante en función de la edad del propelente
edad del propelente
30
25
20
15
10
5
0
0
1000
2000
3000
resistencia al corte

Después de obtener el ajuste por mínimos cuadrados, surgen varias preguntas interesantes:

1. ¿Qué tan bien se ajusta esta ecuación a los datos?
2. ¿Es probable que el modelo sea útil como predictor?
3. ¿Se viola alguna de los supuestos básicos (como la de la varianza constante y la de errores no correlacionados)? Y en caso afirmativo ¿Qué tan grave es eso?

## Estimación de $\sigma^2$

Además de estimar $\beta_0$ y $\beta_1$ , se requiere un estimado de $\sigma^2$ y estimar un intervalo pertinente al modelo de regresión.

| Observado | Ajustado | | |
|---|---|---|---|
| 2158.7 | 2051.94 | 106.76 | 11397.6976 |
| 1678.15 | 1745.42 | -67.27 | 4525.2529 |
| 2316 | 2330.59 | -14.59 | 212.8681 |
| 2061.3 | 1996.21 | 65.09 | 4236.7081 |
| 2207.5 | 2423.48 | -215.98 | 46647.3604 |
| 1708.3 | 1921.9 | -213.6 | 45624.96 |
| 1784.7 | 1736.14 | 48.56 | 2358.0736 |
| 2575 | 2534.94 | 40.06 | 1604.8036 |
| 2357.9 | 2349.17 | 8.73 | 76.2129 |
| 2256.7 | 2219.13 | 37.57 | 1411.5049 |
| 2165.2 | 2144.83 | 20.37 | 414.9369 |
| 2399.55 | 2488.5 | -88.95 | 7912.1025 |
| 1779.8 | 1698.98 | 80.82 | 6531.8724 |
| 2336.75 | 2265.57 | 71.18 | 5066.5924 |
| 1765.3 | 1810.44 | -45.14 | 2037.6196 |
| 2053.5 | 1959.06 | 94.44 | 8918.9136 |
| 2414.4 | 2404.9 | 9.5 | 90.25 |
| 2200.5 | 2163.4 | 37.1 | 1376.41 |
| 2654.2 | 2553.52 | 100.68 | 10136.4624 |
| 1753.7 | 1829.02 | -75.32 | 5673.1024 |

$$SS_{Res} = \sum_{i=1}^{n} e_i^2 = \sum_{I=1}^{n} \left( y_i - \hat{y}_i \right)^2$$

$$SS_{Res} = \sum_{i=1}^{n} e_i^2 = 166253.7043$$

$$\hat{\sigma}^2 = \frac{SS_{Res}}{n-2} = 9236.31691$$

Error estándar de estimación de la regresión

$$\hat{\sigma} = 96.105759$$

# VARIANZA DE LOS ESTIMADORES DE LOS PARAMETROS DE REGRESION

$$Var\left(\hat{\beta}_1\right) = \frac{\hat{\sigma}^2}{S_{xx}} \qquad Var\left(\hat{\beta}_0\right) = \hat{\sigma}^2\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)$$

$$S_{xx} = 1106.56 \qquad \hat{\sigma}^2 = \frac{SS_{Res}}{n-2} = 9236.31691 \qquad \bar{x} = 13.3625$$

$$Var\left(\hat{\beta}_1\right) = \frac{9236.31691}{1106.56} = 8.3468 \qquad \sigma_{\hat{\beta}_1} = \sqrt{8.3468} = 2.88$$

$$Var\left(\hat{\beta}_0\right) = 9236.31691\left(\frac{1}{20} + \frac{178.55}{1106.56}\right)$$

$$Var\left(\hat{\beta}_0\right) = 1952.2036 \qquad \sigma_{\hat{\beta}_0} = \sqrt{1952.2036} = 44.18$$

## PRUEBA DE HIPOTESIS DE LOS PARAMETROS DE REGRESION

**Uso de la prueba** *t-student para probar las hipótesis de los parámetros de regresión.*

Supongamos que se desea probar la hipótesis que la pendiente es igual a una constante, por ejemplo, a $\beta_{10}$. Las hipótesis correspondientes son

$$H_0: \beta_1 = \beta_{10}$$

$$H_a: \beta_1 \neq \beta_{10}$$

Estadístico de prueba

$$t_0 = \frac{\hat{\beta}_1 - \beta_{10}}{\sigma_{\hat{\beta}_1}}$$

Se rechaza Ho, si $\left| t_0 \right| > t_{\frac{\alpha}{2}, n-2}$

## PRUEBA DE HIPOTESIS DE LOS PARAMETROS DE REGRESION

**Uso de la prueba** *t-student para probar las hipótesis de los parámetros de regresión.*

Supongamos que se desea probar la hipótesis que el intercepto es igual a una constante, por ejemplo, a $\beta_{00}$. Las hipótesis correspondientes son

$H_0: \beta_0 = \beta_{00}$

$H_a: \beta_0 \neq \beta_{00}$

Estadístico de prueba

$$t_0 = \frac{\hat{\beta}_0 - \beta_{00}}{\sigma_{\hat{\beta}_0}}$$

Se rechaza Ho, si $\left| t_0 \right| > t_{\frac{\alpha}{2}, n-2}$

Ejemplo para

$$H_0: \beta_1 = 0$$

$$H_a: \beta_1 \neq 0$$

$$t_0 = \frac{\hat{\beta}_1 - \beta_{10}}{\sigma_{\hat{\beta}_1}}$$

$$t_0 = \frac{-37.15 - 0}{2.88911} = -12.85$$

$$t_{\frac{\alpha}{2,n-2}} = t_{0.025,18} = 2.445$$

$$\left|-37.15\right| > 2.445$$

Se rechaza la hipótesis Ho, por lo que podemos concluir que la pendiente no es cero.

Ejemplo para

$$H_0: \beta_0 = 0$$

$$H_a: \beta_0 \neq 0$$

$$t_0 = \frac{\hat{\beta}_0 - \beta_{00}}{\sigma_{\hat{\beta}_0}}$$

$$t_0 = \frac{2627.82 - 0}{44.18} = 59.47$$

$$t_{\frac{\alpha}{2,n-2}} = t_{0.025,18} = 2.445$$

$$\left|59.47\right| > 2.445$$

Se rechaza la hipótesis Ho, por lo que podemos concluir que el intercepto no es cero.

# Significancia del modelo de regresión

| Fuente de variación | Suma de cuadrados | Grados de libertad | Cuadrado medio | Fo |
|---|---|---|---|---|
| Regresión | 1527334,94 | 1 | 1527334,94 | 165,21 |
| Residual | 166402,65 | 18 | 9244,5918 | |
| Total | 1693,73 | 19 | | |

$$SS_T = \sum_{i=1}^{n} y_i^2 - n\bar{y}^2 = 1693{,}737\,.60 \qquad \hat{\beta}_1 = -37.15$$

$$S_{xy} = \sum_{i=1}^{n} y_i x_i - \frac{\left(\sum_{i=1}^{n} y_i\right)\left(\sum_{i=1}^{n} x_i\right)}{n} = 528492.64 - \frac{(267.25)(42627.15)}{20} = -41112.65$$

$$\boldsymbol{SS_R} = (-37.15)(-41112.65) = 1527334.947$$

$\boldsymbol{MS_R} = 1527334.947/1 = 1527334.947$ **Fo=1527334.947/9244.5918=165.21**

$$\boldsymbol{SS_{Res}} = 1693737.60 - 1527334.947 = 166402.6530$$

$\boldsymbol{MS_{Res}}$=166402.6530/18=9244.5918

| Fuente de variación | Suma de cuadrados | Grados de libertad | Cuadrado medio | Fo |
|---|---|---|---|---|
| Regresión | 1527334,94 | 1 | 1527334,94 | 165,21 |
| Residual | 166402,65 | 18 | 9244,5918 | |
| Total | 1693,73 | 19 | | |

**si** $F_0 > F_{\alpha,\, n-2}$ **El modelo es significativo**

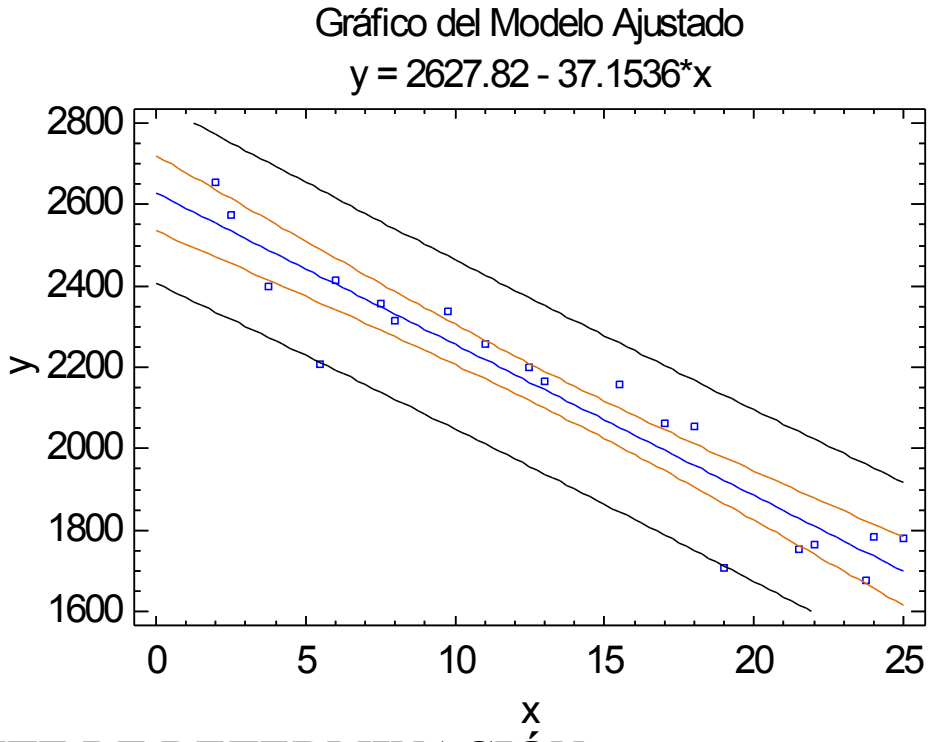
$$F_{0.05,\, 18} = 4.414$$

$$F_0 = 165.21 > F_{0.05,\, 18} = 4.414$$

**Por lo tanto el modelo es significativo.**

| | Mínimos Cuadrados | Estándar | Estadístico | |
|---|---|---|---|---|
| Parámetro | Estimado | Error | T | Valor-P |
| Intercepto | 2627.82 | 44.1839 | 59.4746 | 0.0000 |
| Pendiente | -37.1536 | 2.88911 | -12.8599 | 0.0000 |

| Fuente | Suma de Cuadrados | Gl | Cuadrado Medio | Razón-F | Valor-P |
|---|---|---|---|---|---|
| Modelo | 1.52748E6 | 1 | 1.52748E6 | 165.38 | 0.0000 |
| Residuo | 166255. | 18 | 9236.38 | | |
| Total (Corr.) | 1.69374E6 | 19 | | | |




**COEFICIENTE DE DETERMINACIÓN**

$$R^2 = \frac{SS_R}{SS_T} = 1 - \frac{SS_{Res}}{SS_T}$$

$$R^2 = \frac{SS_R}{SS_T} = \frac{1527334.95}{1693737.60} = 0.9018$$

**Por tanto, el 90.18% de la variabilidad de la resistencia queda explicada por el modelo de regresión.**

## COEFICIENTE DE CORRELACION

$$r = \frac{\sum_{i=0}^{n} y_i\left(x_i - \bar{x}\right)}{\left[\sum_{i=0}^{n}\left(x_i - \bar{x}\right)^2 \sum_{i=0}^{n}\left(y_i - \bar{y}\right)^2\right]^{1/2}} = \frac{S_{xy}}{\left[S_{xx} S S_T\right]^{1/2}}$$

Con frecuencia es útil probar la hipótesis que el coeficiente de correlación es cero, esto es

$$H_0:\ \rho = 0$$

$$H_1:\ \rho \neq 0$$

$$t_0 = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

Que sigue la distribución $t$ con $n-2$ grados de libertad si $H_0:\ \rho = 0$ es cierta. Así, se rechazaría la hipótesis nula si $\left|t_0\right| > t_{\frac{\alpha}{2},n-2}$.

$$r = \frac{S_{xy}}{\left[S_{xx}SS_T\right]^{1/2}} = \frac{-41112.65}{\left[\left(1106.56\ \right)(1693737.60)\right]^{1/2}}$$

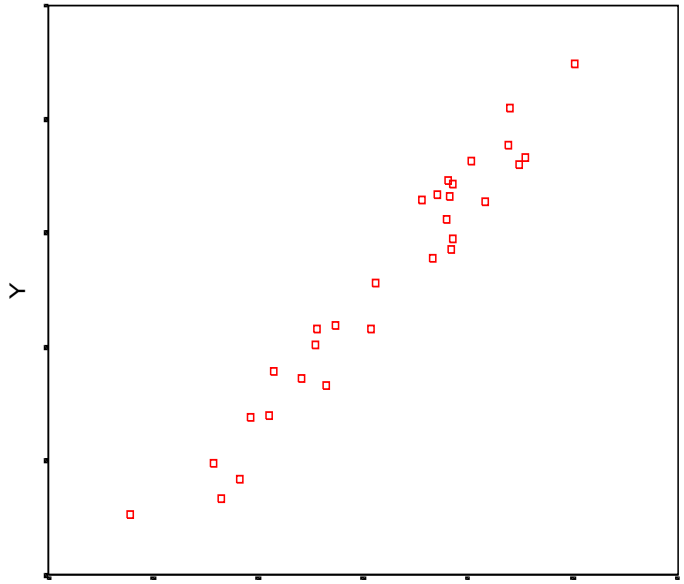
$$r = \frac{-41112.65}{\left[1873883531.136\right]^{1/2}}$$

$$\boldsymbol{r = \frac{-41112.65}{43.288.37} = -0.9497}$$

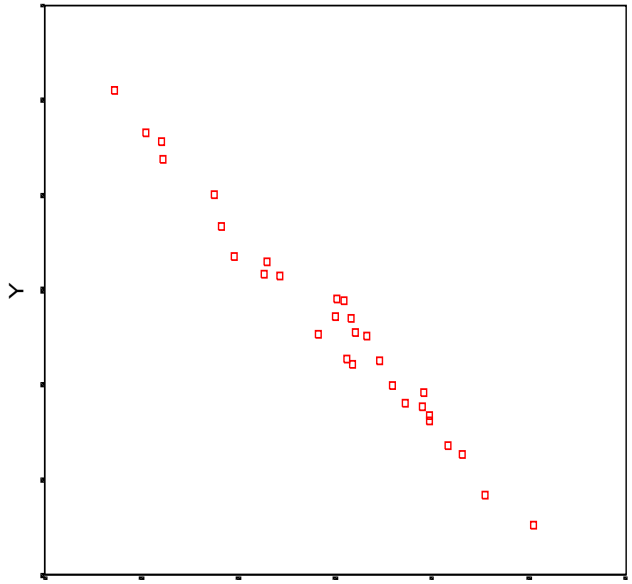
$$\boldsymbol{t_0 = \frac{-0.9497\sqrt{20-2}}{\sqrt{1-(-0.9497)^2}} = -41.11} \qquad \boldsymbol{t_{0.025,18} = 2.445}$$

$$\boldsymbol{\left|-41.11\right| > 2.445}$$

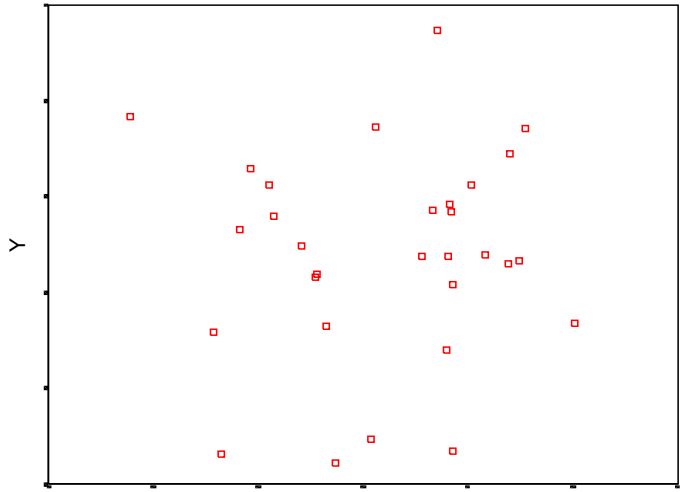
**Se rechaza la hipótesis Ho, por lo que podemos concluir que el $\rho \neq 0$**



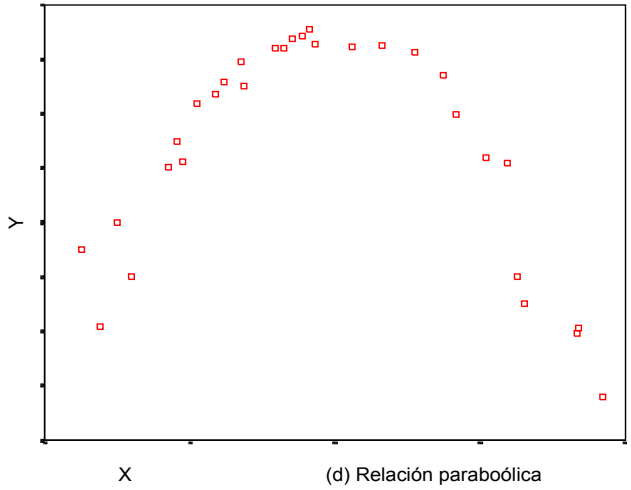

(a) Correlación lineal positiva




(b) Correlación lineal negativa

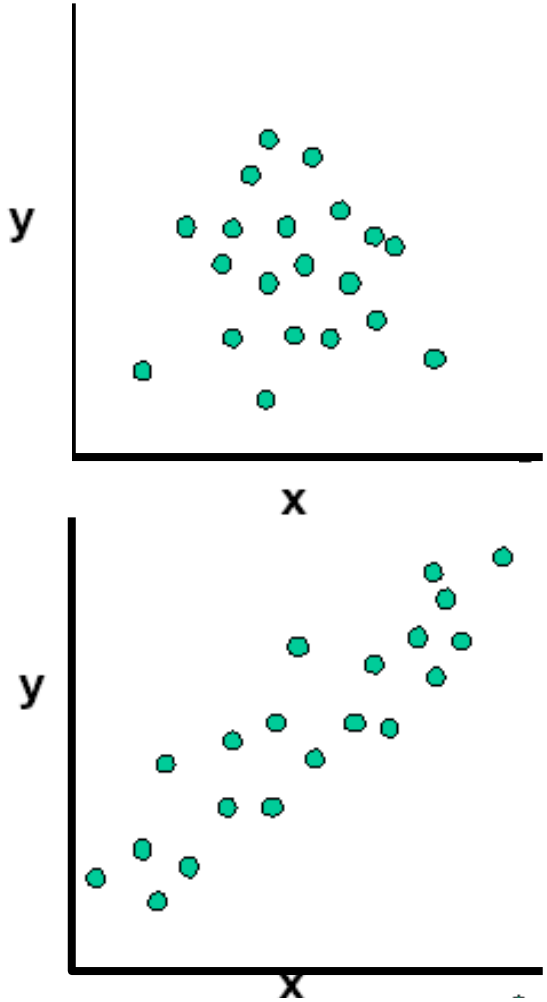



(c) No correlación




(d) Relación paraboólica

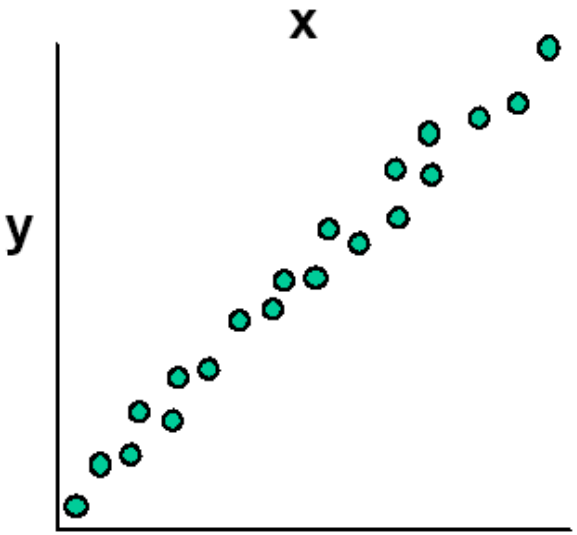
# ✓ Patrones de diagramas de dispersión



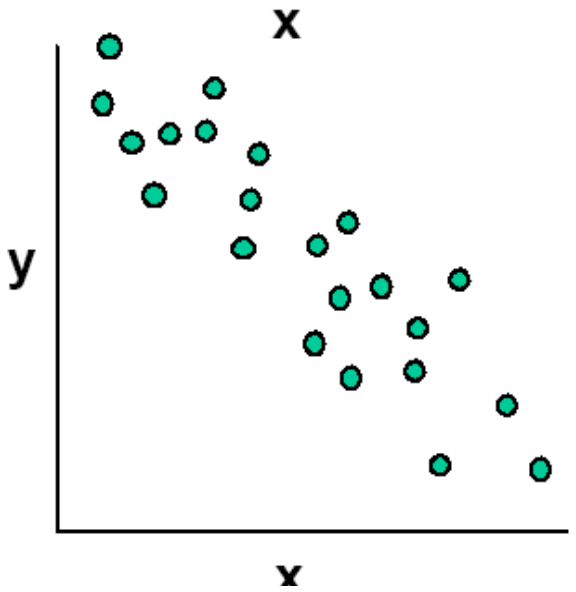

No existe tendencia hacia arriba ni hacia abajo. Las dos variables no se encuentran relacionadas.

Existe tendencia lineal de las variables. Las dos variables se encuentran relacionadas de manera positiva.

# ✓ Patrones de diagramas de dispersión




Existe una tendencia fuertemente positiva ya que los puntos dibujados forman una línea casi recta, por lo cual se argumenta en este caso que las dos variables están positiva y fuertemente relacionadas.




Existe una tendencia negativa ya que los puntos dibujados se encuentran en sentido opuesto, por lo cual se argumenta en este caso que las dos variables están negativamente relacionadas.