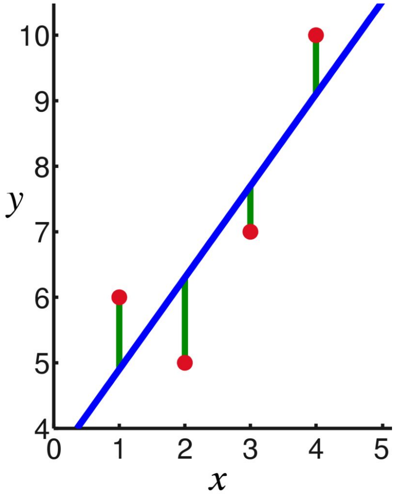
Linear Regression Model

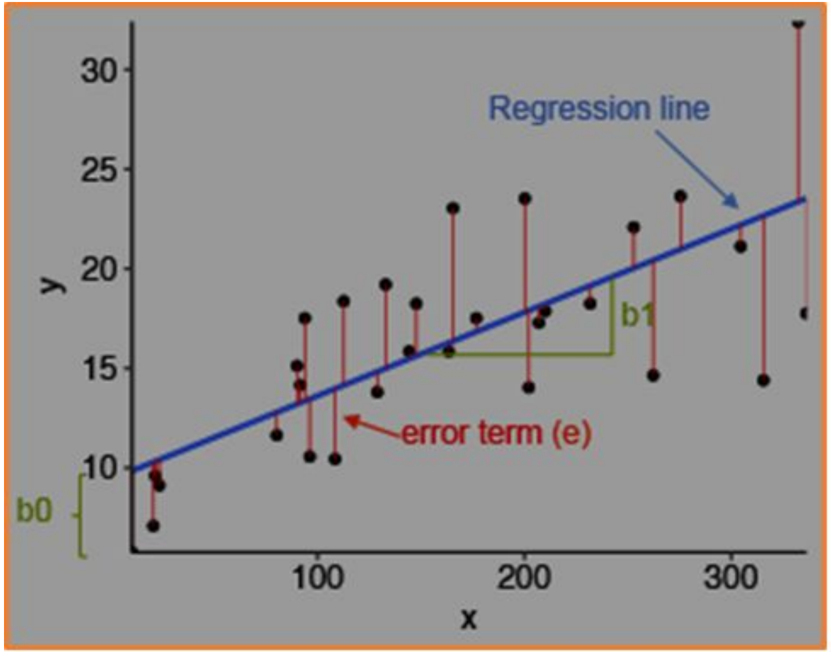
Linear Regression

- LinearRegression fits a linear model with coefficients w = (w1, …, wp) to minimize the residual sum of squares between the observed targets in the dataset, and the targets predicted by the linear approximation.



Y = w0 + w1.X

Linear Regression



Estimated (or predicted) y value

Estimate of the regression intercept

Estimate of the regression slope

Independent variable

$$y_i = b_0 + b_1 x + e$$

Error term

Linear Regression

$$\underbrace{\boxed{\hat{y}}}_{\text{target}} = \underbrace{\boxed{\beta_0 + \beta_1}}_{\text{coefficients}} \underbrace{\boxed{X}}_{\text{input}} + \underbrace{\boxed{\in}}_{\text{random error}}$$

$$y = \alpha + \beta x,$$

$$y_i = \alpha + \beta x_i + \varepsilon_i.$$

$$\hat{\varepsilon}_i = y_i - \alpha - \beta x_i.$$

In other words, $\widehat{\alpha}$ and $\widehat{\beta}$ solve the following minimization problem:

$$\text{Find } \min_{\alpha,\,\beta} Q(\alpha, \beta), \quad \text{for } Q(\alpha, \beta) = \sum_{i-1}^{n} \hat{\varepsilon}_i^{\,2} = \sum_{i-1}^{n} (y_i - \alpha - \beta x_i)^2 \ .$$

By expanding to get a quadratic expression in α and β, we can derive values of α and β that minimize the objective function Q (these minimizing values are denoted $\widehat{\alpha}$ and $\widehat{\beta}$):[

$$\widehat{\alpha} = \bar{y} - (\widehat{\beta}\,\bar{x}),$$

$$\widehat{\beta} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

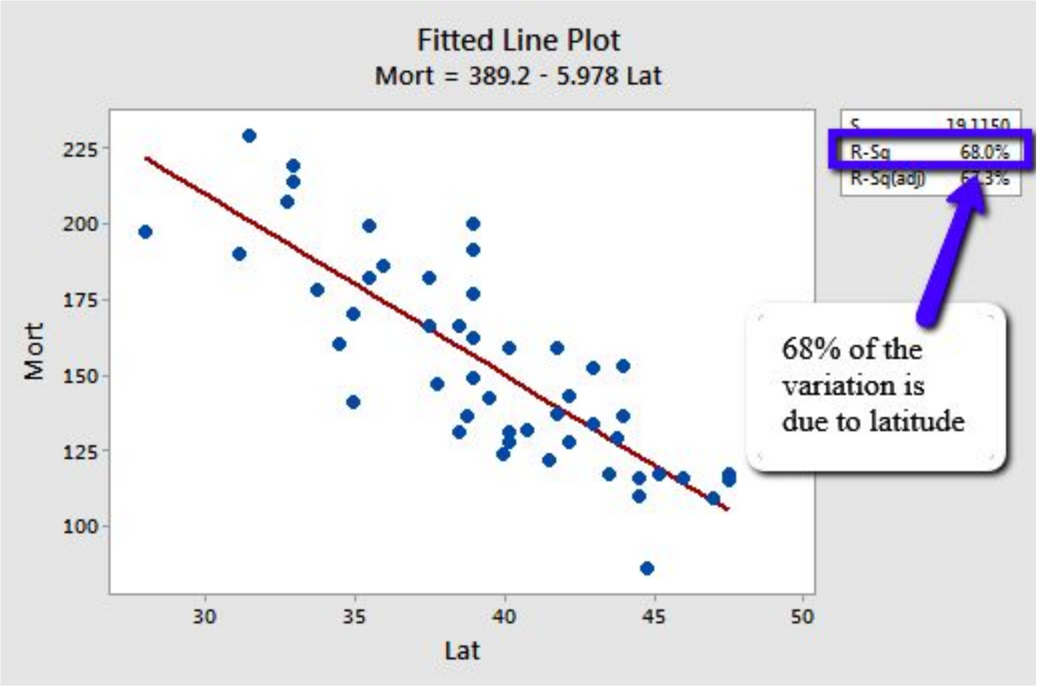
Coeficiente de Determinação

In regression, the R^2 coefficient of determination is a statistical measure of how well the regression predictions approximate the real data points. An R^2 of 1 indicates that the regression predictions perfectly fit the data.

$$R^2 = 1 - \frac{RSS}{TSS}$$

$$RSS = \sum_{i=1}^{n} (y_i - f(x_i))^2 \qquad \mathrm{TSS} = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

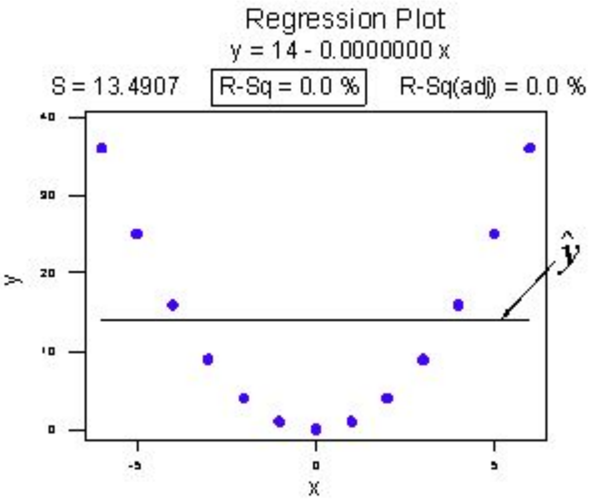
R-squared

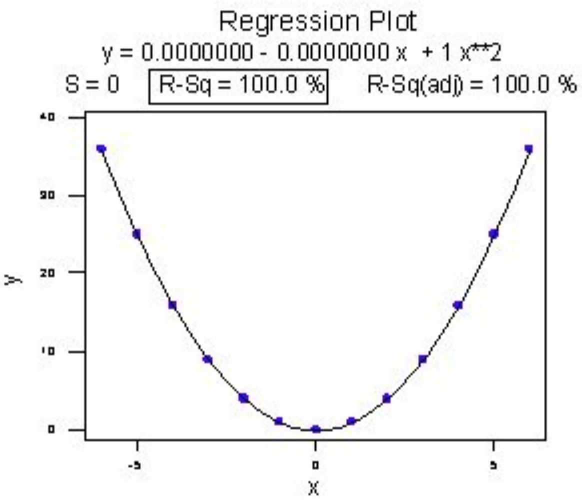

Fitted Line Plot
Mort = 389.2 - 5.978 Lat
S 19.1150
R-Sq 68.0%
R-Sq(adj) 67.3%
68% of the variation is due to latitude
Mort
Lat
225
200
175
150
125
100
30
35
40
45
50

Caution #1

The coefficient of determination r2 and the correlation coefficient *r* quantify the strength of a *linear* relationship. It is possible that r2= 0% and *r* = 0, suggesting there is no linear relation between *x* and *y*, and yet a perfect curved (or "curvilinear" relationship) exists.

Caution #1


Regression Plot
y = 14 - 0.0000000 x
S = 13.4907
R-Sq = 0.0 %
R-Sq(adj) = 0.0 %
ŷ
y
X


Regression Plot
y = 0.0000000 - 0.0000000 x + 1 x**2
S = 0
R-Sq = 100.0 %
R-Sq(adj) = 100.0 %
y
X

Caution #2

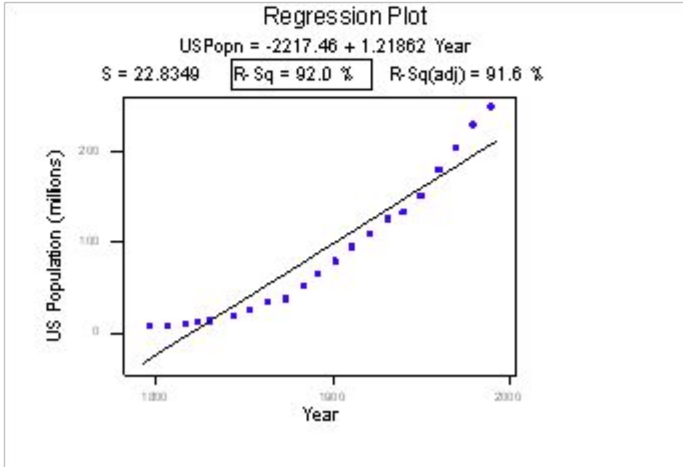
A large r2 value should not be interpreted as meaning that the estimated regression line fits the data well. Another function might better describe the trend in the data.

r2 ×100 percent of the variation in y is reduced by taking into account predictor x"

or:

r2 ×100 percent of the variation in y is 'explained by' the variation in predictor x."

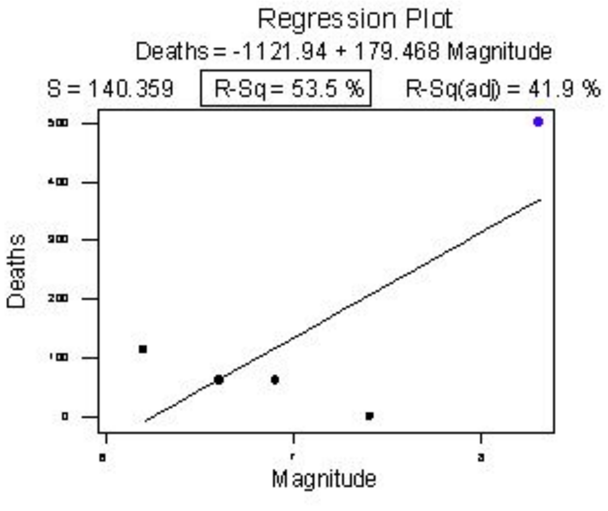
Caution #2

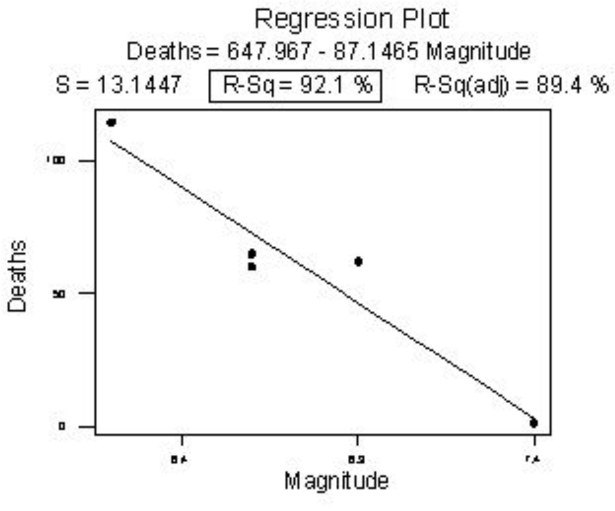


Regression Plot
USPopn = -2217.46 + 1.21862 Year
S = 22.8349
R-Sq = 92.0 %
R-Sq(adj) = 91.6 %
US Population (millions)
Year

Caution #3

The coefficient of determination r2 and the correlation coefficient r can both be greatly affected by just one data point (or a few data points).





Caution #4

Correlation (or association) does not imply causation.