



# CORRELATION AND REGRESSION

## Correlation

Assessing a linear relationship is called correlation analysis. We can test if a relationship exists.

Pearson correlation coefficient is the most common measure of correlation.

Notation:  $r$  : sample correlation coefficient;  $\rho$ : population correlation coefficient

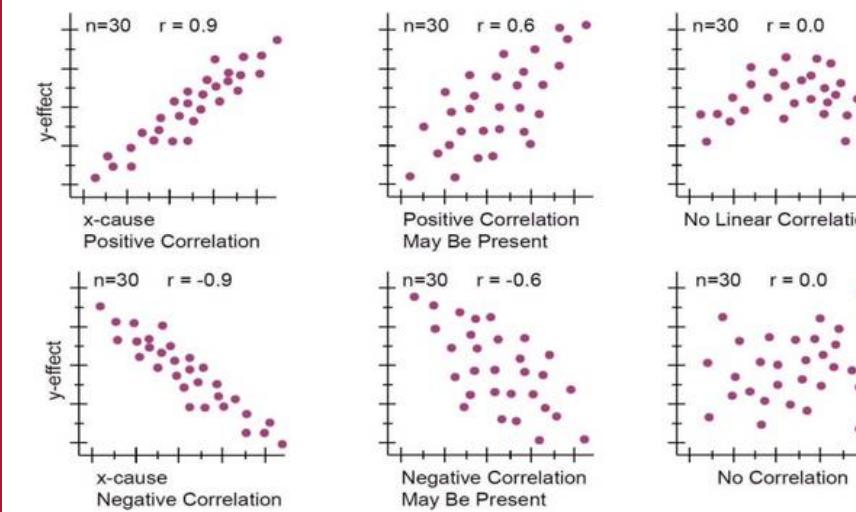
$$r = \frac{\sum xy - n\bar{x}\bar{y}}{(n-1)s_x s_y}$$

Excel Function: **CORREL(x,y)**

Correlation coefficient is always between -1 and 1.

If the points in a scatterplot go in an upward positive direction, correlation coefficient will be positive.

A correlation coefficient close to zero indicates the two variables are not linearly correlated.



## Test for Significance

Consider a random sample of size  $n$  of paired data for two variables  $X$  and  $Y$ .

We can test whether the correlation between  $X$  and  $Y$  is zero:

$$H_0: \rho = 0 \\ H_a: \rho \neq 0$$

We use t-test with test statistic:

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

where  $t$  follows a student distribution with  $n - 2$  degree of freedom.

Look up the critical value  $t^{*\alpha,n-2}$

If  $|t| > t^{*\alpha,n-2}$ , reject  $H_0$ , otherwise, do not reject  $H_0$ .

## Test for Significance Excel

Critical value  $t^{*\alpha,n-2}$  can be calculated in Excel as:

**T.INV(1 - α/2, n - 2)**

P-value can also be computed to be compared with  $\alpha$ :

**p-value = 2 × T.DIST(-ABS(t), n - 2, TRUE)**

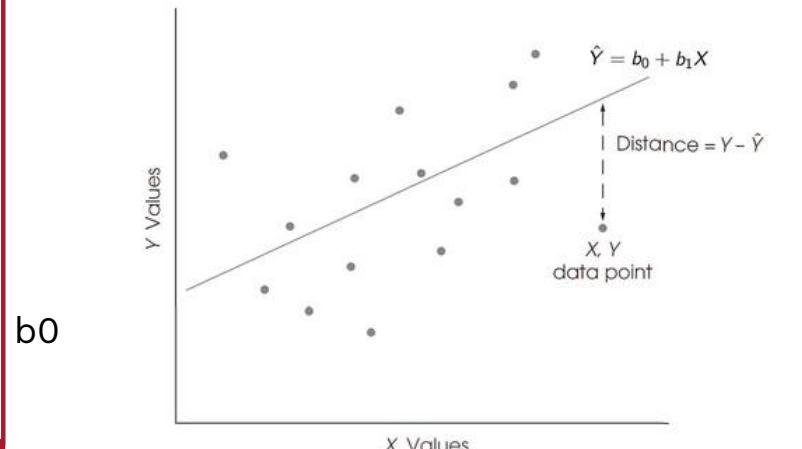
If p-value  $< \alpha$ , reject  $H_0$ , conclude that the correlation is significant.

## Simple Least Squared Linear Regression

If two variables  $X$  and  $Y$  are related to each other, we can predict the value of one variable using the value of the other.

We fit a curve to the data to find the best relationship, the simplest of these being a straight line.

This straight line is called the least squares regression line of  $Y$  on  $X$ .



$$Y = \beta_0 + \beta_1 X + c$$

Simple linear regression model is:

**Y** = dependent variable (response variable)

**X** = independent variable (predictor or explanatory variable)

**β0** = intercept (value of  $Y$  when  $X = 0$ )

**β1** = slope of regression line

**c** = random error

## Simple Least Squared Linear Regression Cont.

The true intercept and slope are unknown. They are estimated using sample data

Regression equation based on sample data:

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

**ŷ**

**b0**

**b1**

**e**

= predicted value of  $Y$

= estimate of  $\beta_0$

= estimate of  $\beta_1$

=  $Y - \hat{Y}$ : regression residuals

## Multiple Linear Regression

Multiple linear regression model:  $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k + c$

R<sup>2</sup> indicates how well the model fits the data

R<sup>2</sup> close to 1 indicates good fit, close to zero indicates poor fit

For simple regression, R<sup>2</sup> is the squared of correlation coefficient, R<sup>2</sup> =  $r^2$

R<sup>2</sup> increases as more variables are added to the model.

We should use the adjusted R<sup>2</sup> that penalizes the adding of non-sense variables.

**t-test:** indicates significance of a variable in the model.

In a good model, all variables are significant. If the p-value of a variable is greater than 0.05, omit it.

**F-test:**

- p-value < 0.05: at least one of  $X$  variables is useful in the model at 5% significance level.
- p-value > 0.05: none of  $X$  variables are useful.