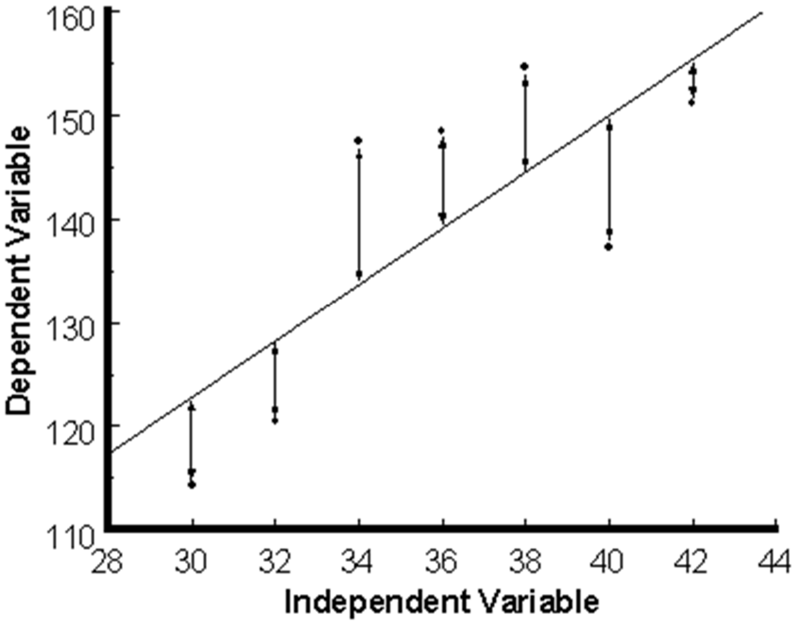
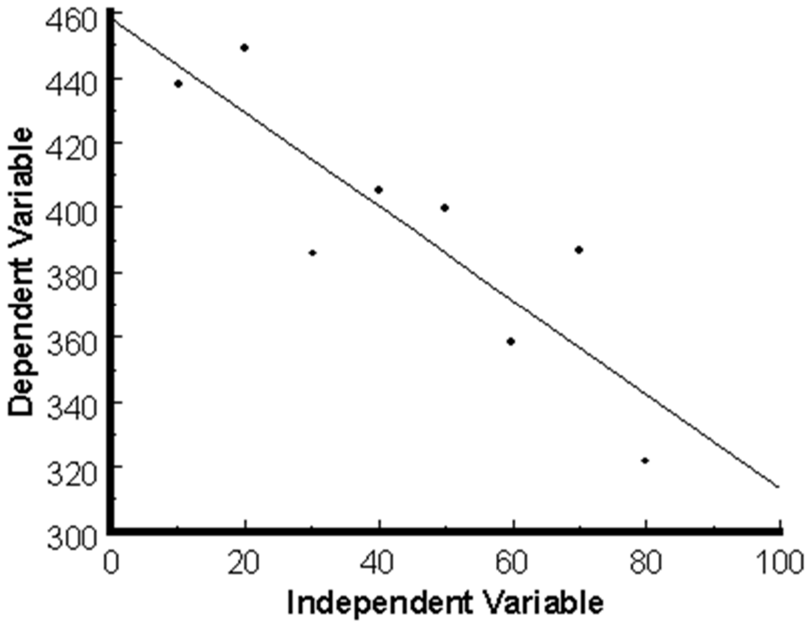
Regression

Regression is a method of fitting an equation to a data set. Simple regression involves one independent variable and one dependent variable. Multiple regression allows more than one independent variable and one dependent variable. Least squares method is the most common method of regression. With least squares regression a straight line is fit to the data by minimizing the sum of the squared distance of each point from the line. The arrows in the figure below display the vertical distance from the best-fit line. This best fit line is determined by minimizing the squared distances represented by the arrow.



The coefficient of determination r^2 is a measure of the amount of variability in the data explained by the regression model. For example, if $r^2 = 0.91$, then 91% of the variability in the model is explained by the regression model. The correlation coefficient, r, measures the correlation between the independent variable and the dependent variable. Perfect positive correlation is achieved when $r = 1$. Perfect negative correlation is achieved when $r = -1$. If there is no correlation between the variables (the variables are independent) then $r = 0$. The figure above displays positive correlation, if the independent variable increases, the dependent variable increases. The figure below demonstrates negative correlation, if the independent variable increases, the dependent variable decreases.



Simple Regression

Simple linear regression is used to estimate the coefficients of the model:

$y_\mathrm{i} = a + bx_\mathrm{i} + e_\mathrm{i}$

where y_i is the dependent variable, x_i is the independent variable, and e_i is the residual, the error in the fit of the model.

Using the method of least squares, the coefficients are estimated from the following expressions.

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

$$a = \frac{\sum_{i=1}^{n} y_i - b\sum_{i=1}^{n} x_i}{n}$$

where n is the sample size. The sum of squares of x is:

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

The sum of squares of y is:

$$S_{yy} = \sum_{i=1}^{n} y_i^2 - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n}$$

The sum of squares of *x* and *y* is:

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

The sample coefficient of determination is:

$$r^2 = \frac{S_{xy}^2}{S_{xx} S_{yy}}$$

The residual in the model is a random variable with a mean of zero, and each individual e_i has the same variance, σ^2. An estimate of this variance is:

$$s^2 = \frac{S_{yy} - bS_{xy}}{n-2}$$

The variance of the estimate of b is:

$$\mathrm{var}(b) = \frac{s^2}{S_{xx}}$$

The variance of the estimate of a is:

$$\mathrm{var}(a) = \frac{\sum_{i=1}^{n} x_i^2}{nS_{xx}}$$

The covariance of the estimates of a and b is:

$$\mathrm{cov}(a,b) = -\bar{x}\,\mathrm{var}(b)$$

An interval that contains *b* with a confidence of (1-α) is

$$b \pm t_{\alpha/2}\sqrt{\frac{s^2}{S_{xx}}}$$

$t_{\alpha/2}$ has n-2 degrees of freedom.

An interval that contains a with a confidence of (1-α) is

$$a \pm \frac{t_{\alpha/2}s\sqrt{\sum_{i=1}^{n} x_i^2}}{\sqrt{nS_{xx}}}$$

$t_{\alpha/2}$ has n-2 degrees of freedom.

An interval that contains y when $x = x_o$ with a confidence of (1-α) is

$$\hat{y}_0 \pm t_{\alpha/2}s\sqrt{\frac{1}{n}+\frac{(x_0-\bar{x})^2}{Sxx}}$$

$t_{\alpha/2}$ has n-2 degrees of freedom.

A prediction interval that contains a single response with a confidence of (1-α) is

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

$t_{\alpha/2}$ has n-2 degrees of freedom.

Example

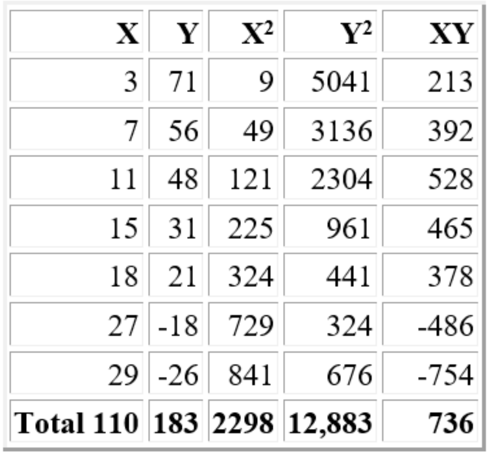
Using linear regression with the data given below:

a. estimate the *y*-intercept.
b. estimate the slope.
c. determine a 90% confidence interval for the *y*-intercept.
d. determine a 90% confidence interval for the slope.
e. determine a 90% prediction interval for the predicted *y* when *x* = 19.
f. determine a 90% confidence interval for the mean *y* when *x* = 19.

X	Y
3	71
7	56
11	48
15	31
18	21
27	-18
29	-26

Solution

Using the table below, the regression parameters are computed as follows.



	X	Y	X²	Y²	XY
	3	71	9	5041	213
	7	56	49	3136	392
	11	48	121	2304	528
	15	31	225	961	465
	18	21	324	441	378
	27	-18	729	324	-486
	29	-26	841	676	-754
Total	**110**	**183**	**2298**	**12,883**	**736**

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

$$a = \frac{\sum_{i=1}^{n} y_i - b\sum_{i=1}^{n} x_i}{n}$$

$$b(\text{slope}) = \frac{7(736) - 110(183)}{7(2298) - (110)^2} = -3.76$$

$$a(\text{y-intercept}) = \frac{183 - (-3.76)(110)}{7} = 85.2$$

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

$$S_{yy} = \sum_{i=1}^{n} y_i^2 - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n}$$

S_{xx}, S_{yy} and S_{xy} are computed as follows.

$$S_{xx} = 2298 - \frac{110^2}{7} = 569.4$$

$$S_{yy} = 12{,}883 - \frac{183^2}{7} = 8098.9$$

$$S_{xy} = 736 - \frac{110(183)}{7} = -2139.7$$

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

For the confidence interval computations, the critical value of the *t*-statistic is required. For a 90% confidence interval 5% of the area will be in each tail of the distribution, thus α = 0.05, and since there are 7 data points, the degrees-of-freedom is *n*-2 = 5. The critical value of the *t*-statistic with α = 0.05 and 5 degrees-of-freedom is 2.015. (Refer this web link https://www.tdistributiontable.com/)

The sample standard deviation of the residuals in the regression model is

$$s = \sqrt{\frac{8098.9 - (-3.76)(-2139.7)}{7-2}} = 3.42*$$

$$s^2 = \frac{S_{yy} - bS_{xy}}{n-2}$$

*Note that the above equation yields a value of 3.27. The difference is due to rounding errors. If all decimal points for the above calculations were carried throughout the resulting value for *s* would be 3.42.

A 90% confidence interval for the *y*-intercept is

$$85.2 - \frac{2.015(3.42)\left(\sqrt{2298}\right)}{\sqrt{7(569.4)}} < a < 85.2 + \frac{2.015(3.42)\left(\sqrt{2298}\right)}{\sqrt{7(569.4)}}$$

$$79.97 < a < 90.43$$

$$a \pm \frac{t_{\alpha/2}\, s \sqrt{\sum_{i=1}^{n} x_i^2}}{\sqrt{nS_{xx}}}$$

A 90% confidence interval for the slope is

$$-3.76 - 2.015\sqrt{\frac{3.42^2}{569.4}} < b < -3.76 + 2.015\sqrt{\frac{3.42^2}{569.4}}$$

$$-4.05 < b < -3.47$$

$$b \pm t_{\alpha/2}\sqrt{\frac{s^2}{S_{xx}}}$$

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

A 90% prediction interval for the predicted *y* when *x* = 19 is

$$y_0 = 85.2 + (-3.76)(19) = 13.76$$

$$13.76 - 2.015(3.42)\sqrt{1 + \frac{1}{7} + \frac{(19-15.7)^2}{569.4}} < y_0 < 13.76 + 2.015(3.42)\sqrt{1 + \frac{1}{7} + \frac{(19-15.7)^2}{569.4}}$$

$$6.3 < y_0 < 21.2$$

A 90% confidence interval for the predicted y when x = 19 is $\hat{y}_0 \pm t_{\alpha/2} s \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{Sxx}}$

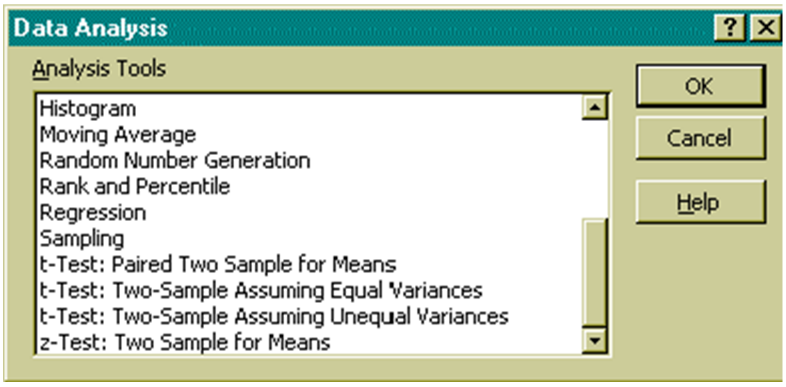
$$13.76 - 2.015(3.42)\sqrt{\frac{1}{7} + \frac{(19-15.7)^2}{569.4}} < y_0 < 13.76 + 2.015(3.42)\sqrt{\frac{1}{7} + \frac{(19-15.7)^2}{569.4}}$$

$$11.0 < y_0 < 16.6$$

Detailed calculations are available in the Microsoft Excel file "**Chapt-8 Regression.xlsx**".

Solution Using Excel's Analysis Tool Pack

This problem is solved much easier with the assistance of Microsoft Excel. Microsoft Excel contains an *Analysis Tool Pack* that can be used to perform this test. To access this package, click the *Tools* from the menu then click *Data Analysis*. If *Data Analysis* is not on the menu the *Analysis Tool Pack* has not been activated. To activate the *Analysis Tool Pack,* click the *Tools* menu and select *Add-Ins*, then check the *Analysis Tool Pack* box and click *OK*. Click the "Data" menu, select "Data Analysis" and the screen shown below will appear.


Data Analysis
Analysis Tools
Histogram
Moving Average
Random Number Generation
Rank and Percentile
Regression
Sampling
t-Test: Paired Two Sample for Means
t-Test: Two-Sample Assuming Equal Variances
t-Test: Two-Sample Assuming Unequal Variances
z-Test: Two Sample for Means
OK
Cancel
Help

Scroll down, select "Regression", then click "OK". The configuration screen will appear. Simply follow instructions.

Multiple Regression

In many cases estimation models require more than one independent variable. For example, a model to estimate a person's weight may include height, sex, and age. For multiple regression, the model being estimated is

$$y_i = b_0 + b_1 x_{1i} + b_2 x_{2i} + \ldots + b_k x_{ki} + e_i$$

where y_i is the dependent variable, x_{1i}, x_{2i},..., x_{ki}, are the independent variables, and e_i is the residual; the error in the fit of the model.

Using least squares regression, the parameters of the model are found by solving the matrix equation

$Y = XB + e$

where

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

$$X = \begin{bmatrix} 1 & (x_{11} - \bar{x}_1) & (x_{21} - \bar{x}_2) & \cdots & (x_{k1} - \bar{x}_k) \\ 1 & (x_{12} - \bar{x}_1) & (x_{22} - \bar{x}_2) & \cdots & (x_{k2} - \bar{x}_k) \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & (x_{1n} - \bar{x}_1) & (x_{2n} - \bar{x}_2) & \cdots & (x_{kn} - \bar{x}_k) \end{bmatrix}$$

$$B = \begin{bmatrix} b_0 \\ b_1 \\ b_2 \\ \vdots \\ b_k \end{bmatrix}$$

and

$$b_0 = b_0^{'} + \sum_{i=1}^{k} b_i \bar{x}_i$$

The matrix equation is solved using matrix algebra

$$B = (X'X)^{-1} X'Y$$

The transpose of the matrix *X* is written as *X*', and the inverse of matrix *X* is written as X^{-1}.

The amount of variability in the data explained by the regression model is measured by the coefficient of multiple correlation:

$$R^2 = \frac{\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}$$

Example

The tensile strength of a seal is dependent on the casting temperature, the casting pressure, and the volume of a chemical agent. Given the experimental data below, develop a model to predict the tensile strength.

Strength	Temperature	Pressure	Chemical Agent
39	192	1.2	12.2
29	199	1.4	13.8
39	196	1.5	12.8
39	208	1.3	12.6
35	202	1.2	13.1
32	186	1.4	13.6
33	183	1.2	13.1
30	180	1.3	13.5

Solution

The tensile strength prediction model is

$S = b\text{o} + b_1T + b_2P + b_3C$

where S is the tensile strength, T is the casting temperature, P is the casting pressure, and C is the volume of the chemical agent. The average of the 8 tensile strength readings is 34.5. The average of the 8 casting temperature readings is 193.25. The average of the 8 casting pressure readings is 1.3125. The average of the chemical agent volumes is 13.0875. The matrix equation used to determine the constants of the prediction model is

$$\begin{bmatrix} 39 \\ 29 \\ 39 \\ 39 \\ 35 \\ 32 \\ 33 \\ 30 \end{bmatrix} = \begin{bmatrix} 1 & -1.25 & -0.1125 & -0.8875 \\ 1 & 5.75 & 0.0875 & 0.7125 \\ 1 & 2.75 & 0.1875 & -0.2875 \\ 1 & 14.75 & -0.0125 & -0.4875 \\ 1 & 8.75 & -0.1125 & 0.0125 \\ 1 & -7.25 & 0.0875 & 0.5125 \\ 1 & -10.25 & -0.1125 & 0.0125 \\ 1 & -13.25 & -0.0125 & 0.4125 \end{bmatrix} \begin{bmatrix} b_0' \\ b_1 \\ b_2 \\ b_3 \end{bmatrix}$$

After completing the appropriate matrix algebra, the regression model is

S = 106.47 - 0.0658T + 9.3P - 7.4C + e

The computations required for the example are extremely tedious if done by hand. Fortunately, there are several computer routines that can be used. Microsoft Excel and Lotus 123 are both capable of performing multiple linear regression. Excel includes an add-in and 123 has functions to perform regression. In Excel, select *Data Analysis* from the *Tools* menu and select the regression option. If *Data Analysis* is not on the menu the *Analysis Tool Pack* has not been activated. To activate the *Analysis Tool Pack* click the *Tools* menu and select *Add-Ins*, then check the *Analysis Tool Pack* box and click *OK*.

In 123, the function "@Regression(X-range, Y-range, Attribute). The X-range specifies the independent variable(s) and includes more than 1 column for multiple regression. The X-range specifies the dependent variable. Setting the attribute to 1 returns, b0, setting the attribute to 3 returns the coefficient of determination, setting the attribute to 101 returns the coefficient of the first independent variable, setting the attribute to 102 returns the coefficient of the second independent variable, setting the attribute to 103 returns the coefficient of the third independent variable, etc.

Detailed calculations are available in the Microsoft Excel file “**Chapt-8 Regression.xlsx**”.

Co-Linearity

When data mining existing data sets that were not carefully produced from experimentation, variables are often co-linear. Consider the production data for a sheet metal press shown below.

Press Force	Thickness	Flatness
487	3.8	0.2
507	4.1	0.4
595	4.4	0.7
622	5.1	0.9

A casual observation of the data shows a relationship between thickness and flatness and a relationship between press force and flatness. An inexperience data analyst may use both press force and thickness as independent variable to predict the dependant variable flatness. Before building multiple regression models, the independent variables should be analysed to ensure their independence. In the example above the independent variables press force and thickness are not independent, thus they are co-linear. Co-linearity means an independent variable is a linear combination of one or more other independent variables. The existence of co-linearity can be detected with a covariance matrix.

Consider the following matrix.

X = 4.0 2.0 0.6

4.2 2.1 0.59

3.9 2.0 0.58

4.3 2.1 0.62

4.1 2.2 0.63

The set of 5 observations, measuring 3 variables, can be described by its mean vector and variance-covariance matrix. The three variables, from left to right are length, width, and height of a certain object, for example. Each row vector x_i is another observation of the three variables (or components).

The mean vector consists of the means of each variable and the variance-covariance matrix consists of the variances of the variables along the main diagonal and the covariances between each pair of variables in the other matrix positions.

The results are:

x-bar = [4.10 2.08 0.604]

0.025   0.0075   0.00175

S = 0.0075   0.0070   0.00135

0.00175 0.00135 0.00043

where the mean vector contains the arithmetic averages of the three variables and the (unbiased) variance-covariance matrix S is calculated by

$$S = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X_i}\right)'\left(X_i - \overline{X_i}\right)$$

where $n = 5$ for this example. Thus, 0.025 is the variance of the length variable, 0.0075 is the covariance between the length and the width variables, 0.00175 is the covariance between the length and the weight variables, 0.007 is the variance of the width variable, 0.00135 is the covariance between the width and weight variables and .00043 is the variance of the weight variable.

The mean vector is often referred to as the centroid and the variance-covariance matrix as the dispersion or dispersion matrix. Also, the terms variance-covariance matrix and covariance matrix are used interchangeably.

List of excel functions used in Regression

Excel, a widely used spreadsheet software, provides several functions that are commonly used in regression analysis. Here is a list of some of the key Excel functions used in regression:

1. `LINEST`: This function returns an array of regression statistics including the slope, intercept, and R-squared value for a linear regression model. It can also provide additional statistics such as standard errors and t-values.
2. `FORECAST`: This function predicts a future value for a given x using a linear regression model. It takes the known x and y values and uses the regression coefficients obtained from `LINEST` to make the prediction.
3. `RSQ`: This function returns the R-squared value, which is a measure of the proportion of the total variation in the dependent variable (y) that can be explained by the regression model. It can be used to assess the goodness of fit of the regression model.
4. `SLOPE`: This function returns the slope of the regression line, which represents the change in the dependent variable (y) for a unit change in the independent variable (x). It can be used to calculate the slope coefficient in a linear regression model.
5. `INTERCEPT`: This function returns the y-intercept of the regression line, which represents the value of the dependent variable (y) when the independent variable (x) is zero. It can be used to calculate the intercept coefficient in a linear regression model.
6. `COVAR`: This function calculates the covariance between two sets of values, such as x and y, which is a measure of how much the two variables change together. It can be used to calculate the covariance between x and y for regression analysis.
7. `STEYX`: This function returns the standard error of the predicted y-values for each x in a regression analysis. It can be used to assess the accuracy of the predicted y-values based on the regression model.
8. `TREND`: This function calculates new y-values based on a linear trendline fitted to a set of x and y values. It can be used to generate predicted y-values using the regression model.

These are just a few examples of Excel functions commonly used in regression analysis. Excel provides many other statistical functions that can be used for regression analysis, depending on the specific requirements of the analysis.