Open topic with navigation

Simple Linear Regression and Correlation

Menu location: **Analysis_Regression and Correlation_Simple Linear and Correlation**.

This function provides simple linear regression and Pearson's correlation.

Regression parameters for a straight line model (Y = a + bx) are calculated by the least squares method (minimisation of the sum of squares of deviations from a straight line). This differentiates to the following formulae for the slope (b) and the Y intercept (a) of the line:

$$b = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i-1}^{n}(x_i - \bar{x})^2}$$

$$a = \bar{Y} - b\bar{x}$$

Regression assumptions:

- Y is linearly related to x or a transformation of x
- deviations from the regression line (residuals) follow a normal distribution
- deviations from the regression line (residuals) have uniform variance

A residual for a Y point is the difference between the observed and fitted value for that point, i.e. it is the distance of the point from the fitted regression line. If the pattern of residuals changes along the regression line then consider using rank methods or linear regression after an appropriate transformation of your data.

Pearson's product moment correlation coefficient (r) is given as a measure of linear association between the two variables:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(Y_i - \bar{Y})^2}}$$

r^2 is the proportion of the total variance (s^2) of Y that can be explained by the linear regression of Y on x. $1-r^2$ is the proportion that is not explained by the regression. Thus $1-r^2 = s^2xY / s^2Y$.

Confidence limits are constructed for r using Fisher's z transformation. The null hypothesis that r = 0 (i.e. no association) is evaluated using a modified t test (Armitage and Berry, 1994; Altman, 1991).

Pearson's correlation assumption:

· at least one variable must follow a normal distribution

The estimated regression line may be plotted and belts representing the standard error and confidence interval for the population value of the slope can be displayed. These belts represent the reliability of the regression estimate, the tighter the belt the more reliable the estimate (Gardner and Altman, 1989).

N.B. If you require a weighted linear regression then please use the multiple linear regression function in StatsDirect; it will allow you to use just one predictor variable i.e. the simple linear regression situation. Note also that the multiple regression option will also enable you to estimate a regression without an intercept i.e. forced through the origin.

Example

From Armitage and Berry (1994, p. 161).
Test workbook (Regression worksheet: Birth Weight, % Increase).

The following data represent birth weights (oz) of babies and their percentage increase between 70 and 100 days after birth.

Birth Weight	% Increase
72	68
112	63
111	66
107	72
119	52
92	75
126	76
80	118
81	120
84	114
115	29
118	42
128	48
128	50
123	69
116	59
125	27
126	60
122	71
126	88
127	63
86	88
142	53
132	50
87	111
123	59
133	76
106	72
103	90
118	68
114	93
94	91

To analyse these data in StatsDirect you must first enter them into two columns in the workbook appropriately labelled. Alternatively, open the test workbook using the file open function of the file menu. Then select Simple Linear and Correlation from the Regression and Correlation section of the analysis menu. Select the column marked "% Increase" when prompted for the response (Y) variable and then select "Birth weight" when prompted for the predictor (x) variable.

For this example:

Simple linear regression

Equation: % Increase = -0.86433 Birth Weight +167.870079

Standard Error of slope = 0.175684
95% CI for population value of slope = -1.223125 to -0.505535

Correlation coefficient (r) = -0.668236 (r^2= 0.446539)

95% CI for r (Fisher's z transformed) = -0.824754 to -0.416618

t with 30 DF = -4.919791
Two sided P < .0001
Power (for 5% significance) = 99.01%

Correlation coefficient is significantly different from zero

From this analysis we have gained the equation for a straight line forced through our data i.e. % increase in weight = 167.87 - 0.864 * birth weight. The r square value tells us that about 42% of the total variation about the Y mean is explained by the regression line. The analysis of variance test for the regression, summarised by the ratio F, shows that the regression itself was statistically highly significant. This is equivalent to a t test with the null hypothesis that the slope is equal to zero. The confidence interval for the slope shows that with 95% confidence the population value for the slope lies somewhere between -0.5 and -1.2. The correlation coefficient r was statistically highly significantly different from zero. Its negative value indicates that there is an inverse relationship between X and Y i.e. lower birth weight babies show greater % increases in weight at 70 to 100 days after birth. With 95% confidence the population value for r lies somewhere between -0.4 and -0.8.

regression and correlation
P values
confidence intervals