statstutor

community project

encouraging academics to share statistics support resources

stcp-marshall-correlationX

The following resources are associated:
Excel dataset ‘Birthweight_reduced’, Simple regression in Excel

Scatterplots and correlation in Excel

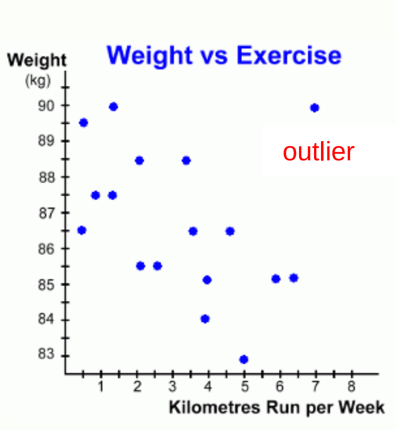
Dependent variable: Continuous (scale)

Independent variables: Continuous (scale)

Common Applications: Assessing the strength of a linear relationship between two continuous variables.

Scatterplots

When examining the relationship between two continuous variables always look at the scatterplot, to see visually the pattern of the relationship between them and look for outliers (observations lying away from the main body of points).

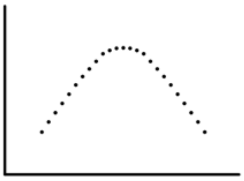


Look for these key things when interpreting a scatterplot:

- Is the relationship weak, moderate or strong
- Is the relationship linear?
- Is the relationship positive or negative?
- Are there any outliers?

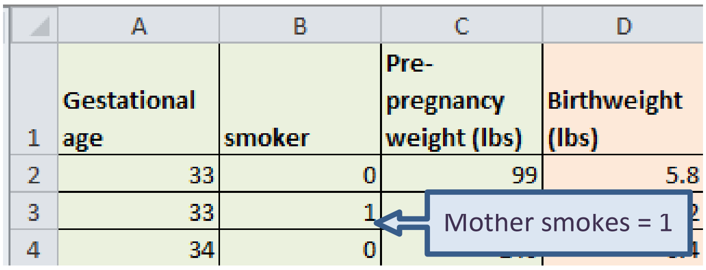
In this example, the relationship between kilometres run per week and weight in kilometres is investigated. Generally, there is a moderate negative relationship (as weight goes down as km per week goes up) which is approximately linear. There is one outlier but it is not extreme enough to be a data entry error.

Correlation measures the strength of a **linear** relationship which means the pattern looks roughly like a line. The graph to the right is an example of a non-linear relationship.



Data: The Excel data set ‘*Birthweight_reduced’* contains details of 42 babies and their parents at birth.

Research question: Which variables affect birth weight? The dependant variable is Birth weight (lbs) and the



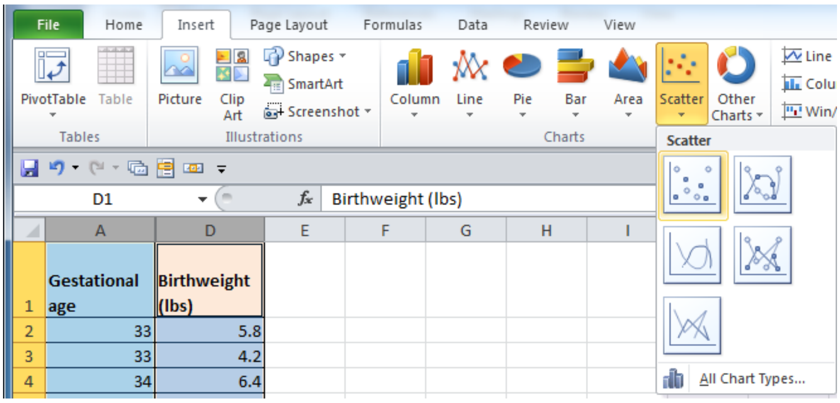
	A	B	C	D
1	Gestational age	smoker	Pre-pregnancy weight (lbs)	Birthweight (lbs)
2	33	0	99	5.8
3	33	1	[illegible]	[illegible]
4	34	0	[illegible]	[illegible]

independent variables include gestational age of the baby at birth (in weeks), pre-pregnancy weight of the mother and whether or not the mother smokes (Smoker = 1).

Steps in Excel

Scatterplots should be produced for each continuous independent with the dependent to see if the relationship is linear (scatter forms a rough line). In Excel it is important to have the column with the independent variable before the column with the dependent.

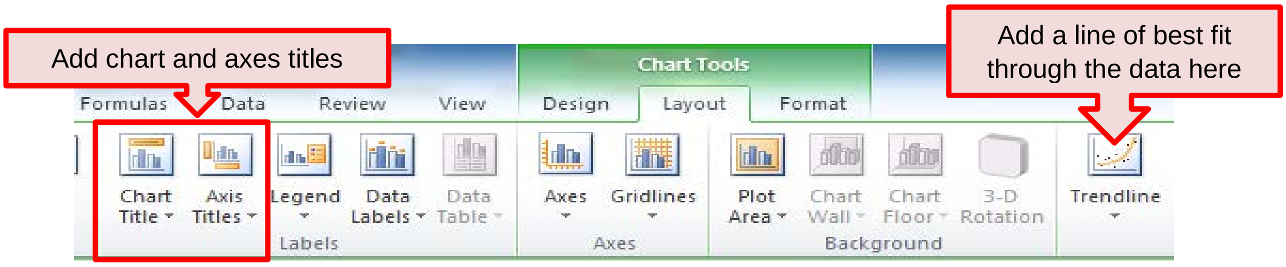
For a scatterplot, select the data for the independent (gestational age), press CTRL and then highlight the data for the dependent variable birthweight. Go to *Insert* → *Scatter* and choose the first option with just scatter.



D1 | Birthweight (lbs)

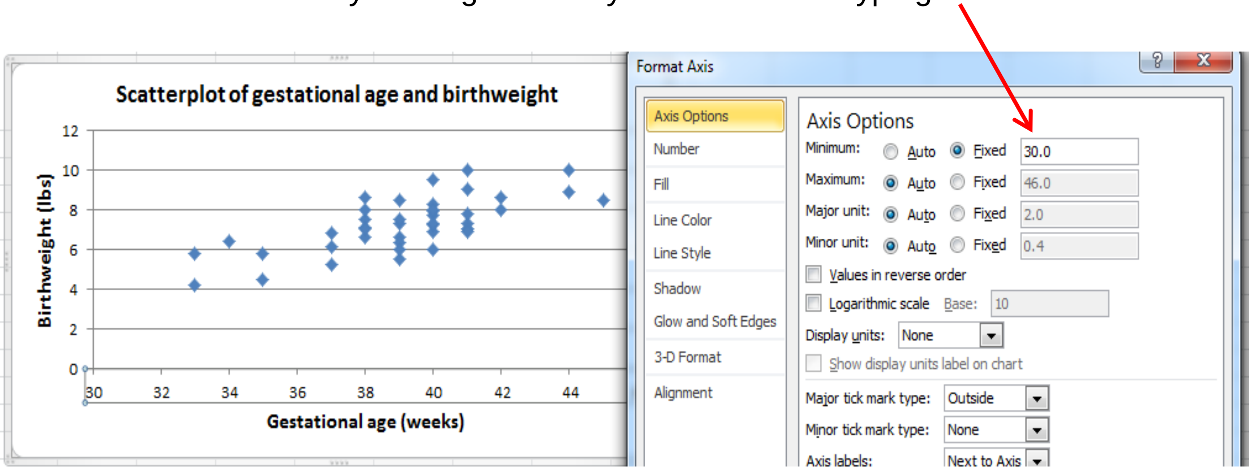
	A	D
1	Gestational age	Birthweight (lbs)
2	33	5.8
3	33	4.2
4	34	6.4

The chart produced needs editing to include titles and labels for the axes. Adjustments can be carried out using the '*Chart Tools'* tabs which are activated by clicking on the chart.


Add chart and axes titles

Add a line of best fit through the data here

You can also change the scale on the x and y axes by clicking on the axis, right clicking and selecting '*Format axis'* to open the following editing menu. Change the minimum x axis value to 30 weeks by clicking '*Fixed*' by 'Minimum' and typing 30 in the box.



Axis Options

Minimum: ○ Auto ● Fixed 30.0

Maximum: ● Auto ○ Fixed 46.0

Major unit: ● Auto ○ Fixed 2.0

Minor unit: ● Auto ○ Fixed 0.4

□ Values in reverse order

□ Logarithmic scale Base: 10

Display units: None

□ Show display units label on chart

Major tick mark type: Outside

Minor tick mark type: None

Axis labels: Next to Axis

The relationship between gestational age and birthweight is clearly linear and positive so an increase in gestation results in an increase in birthweight. Producing a correlation coefficient will give a better idea of the strength of the relationship.

Correlation

A correlation coefficient (r) measures the strength of a linear association between two variables and ranges between -1 (perfect negative correlation) to 1 (perfect positive correlation). There are several types of correlation but they are all interpreted in the same way. Cohen (1992) proposed these guidelines for the interpretation of a correlation coefficient:

Correlation coefficient value	Association
-0.3 to +0.3	Weak
-0.5 to -0.3 or 0.3 to 0.5	Moderate
-0.9 to -0.5 or 0.5 to 0.9	Strong
-1.0 to -0.9 or 0.9 to 1.0	Very strong

Cohen, L. (1992). Power Primer. Psychological Bulletin, 112(1) 155-159

Pearson's correlation coefficient

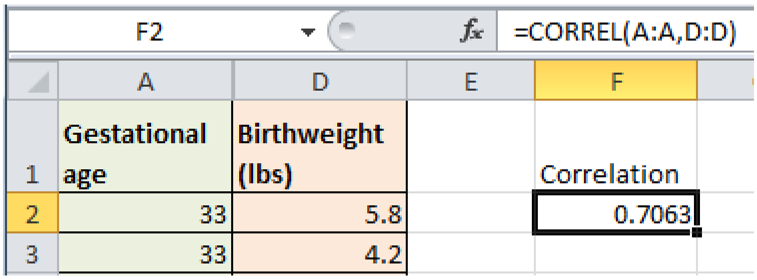
Pearson's correlation coefficient is the most common measure of correlation and is used when both variables are continuous (scale).

Assumptions	**How to check**	**What to do if assumption is not met**
Continuous data for each variable	Check data	If ordinal data use Spearman's or Kendall tau
Linearly related variables	Scatter plot	Transform data
Both variables are normally distributed	Histograms of variables/ Shapiro Wilk	Use rank correlation: Spearman's or Kendall tau

Spearman's rank correlation coefficient, $r_{s,}$ is a non-parametric statistical measure of the strength of a monotonic relationship between paired data. Kendall's τ ('tau') measures the degree to which a relationship is always positive or always negative and is useful for small data sets with a large number of tied ranks. Neither of these can be calculated using the standard Excel commands or toolpak.

Command in Excel

The Excel command for calculating the correlation between two variables is `=CORREL(variable1, variable2).` The example below shows the command for calculating the correlation coefficient for gestational age and birthweight which is 0.71. (Note: Excel always reports to too many decimal places. Two decimal places are enough for a correlation coefficient).

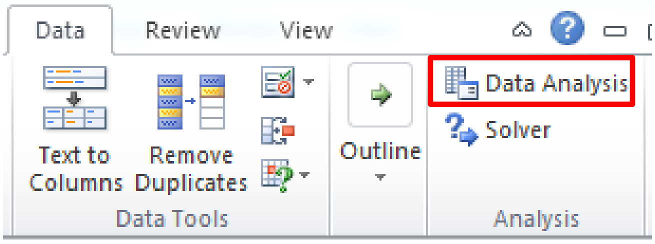


F2 =CORREL(A:A,D:D)

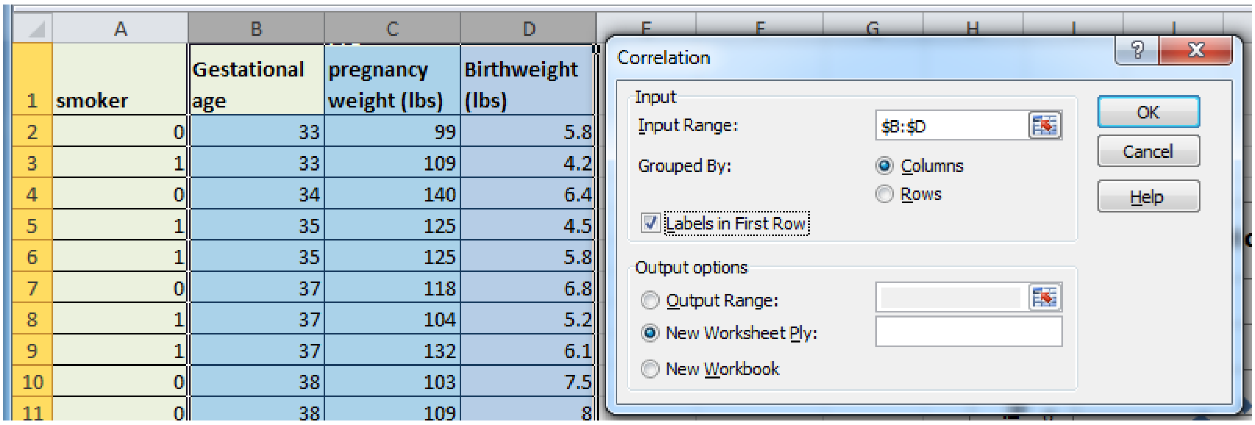
	A	D	E	F
1	Gestational age	Birthweight (lbs)		Correlation
2	33	5.8		0.7063
3	33	4.2		

Using the data analysis toolpak

Excel has an add-in package which needs to be activated through the Excel options (see the '*Additional toolpaks in Excel*' sheet for more details). Once the data analysis toolpak is loaded, it will appear in the **Data** tab.



Click on the **Data Analysis** button and choose **Correlation** from the options. In the Input Range box, select all the columns you wish to calculate correlations for (here it's columns B – D). Select the '*Labels in First Row*' option and then click **OK**.



	A	B	C	D
1	smoker	Gestational age	pregnancy weight (lbs)	Birthweight (lbs)
2	0	33	99	5.8
3	1	33	109	4.2
4	0	34	140	6.4
5	1	35	125	4.5
6	1	35	125	5.8
7	0	37	118	6.8
8	1	37	104	5.2
9	1	37	132	6.1
10	0	38	103	7.5
11	0	38	109	8

Correlation

Input

Input Range: $B:$D

Grouped By: ● Columns ○ Rows

☑ Labels in First Row

Output options

○ Output Range:

● New Worksheet Ply:

○ New Workbook

The output

Excel will produce correlations for each pair of variables selected e.g. the correlation coefficient for pre-pregnancy weight and birthweight is 0.39 which is a moderate positive relationship.

	Gestational age	*Pre-pregnancy weight (lbs)*	*Birthweight (lbs)*
Gestational age	1		
Pre-pregnancy weight (lbs)	0.251	1	
Birthweight (lbs)	0.706	0.390	1

Reporting Correlation

Pearson's correlation was carried out to look for relationships between the variables birthweight, gestational age and weight of mother. There was strong positive relationship between birthweight and gestational age (r = 0.709) and a moderate positive relationship between birthweight and the pre-pregnancy weight of mother (r = 0.39). Pre-pregnancy weight and gestational age at birth were only weakly positively related (r = 0.251).