statstutor

community project

encouraging academics to share statistics support resources

stcp-marshall-scatterS

The following resources are associated:
SPSS dataset 'Birthweight_reduced.sav', Simple and Multiple regression in SPSS

Scatterplots and correlation in SPSS

Dependent variable: Continuous (scale)

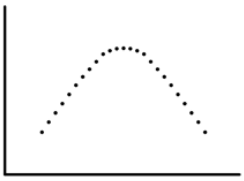
Independent variables: Continuous (scale)

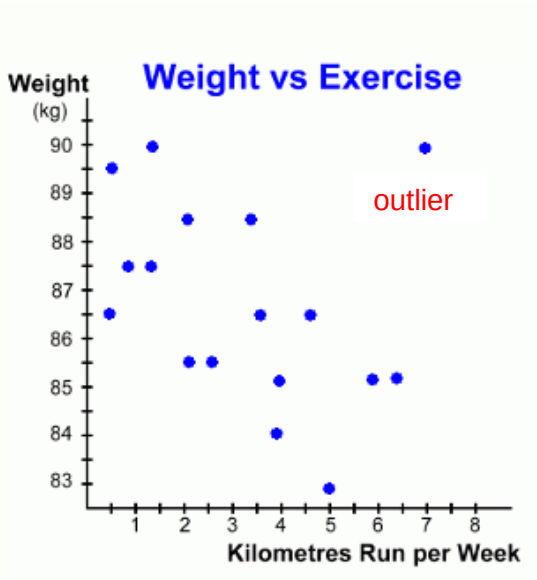
Common Applications: Assessing the strength of a linear relationship between two continuous variables.

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).

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.





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.

© Ellen Marshall
University of Sheffield

Reviewer: Jean Russell
University of Sheffield

Data: The data set '*Birthweight_reduced.sav*' contains details of 42 babies and their parents at birth. The research question is which factors affect birth weight. The dependant variable is Birth weight (lbs) and the independent variables for this sheet are gestational age of the baby at birth (in weeks) and whether or not the mother smokes.

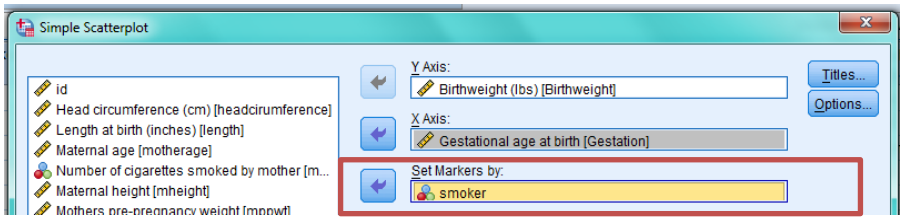
Birthweight	Gestation	smoker
5.8	33	
4.2	33	1
6.4	34	0

Mother smokes = 1

Steps in SPSS

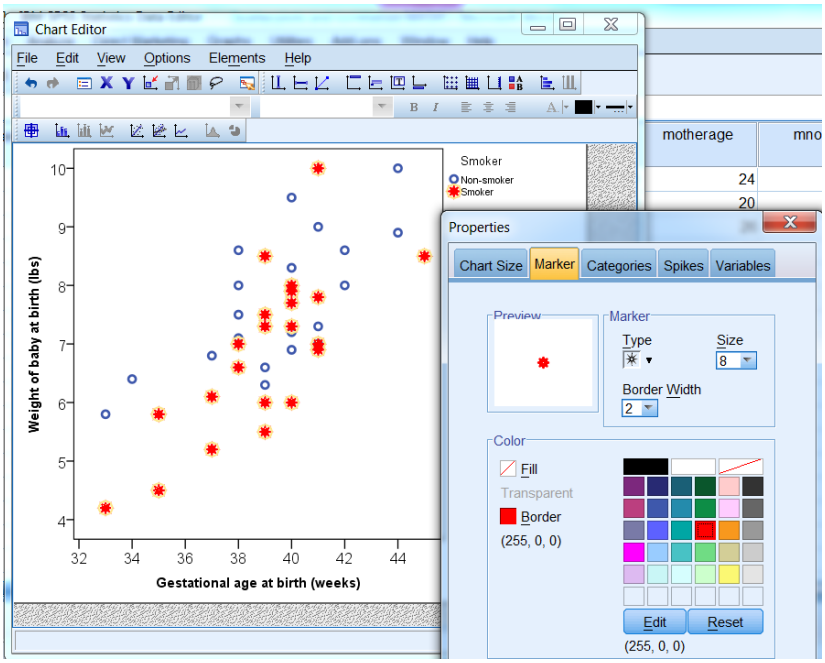
Scatterplots should be produced for each independent with the dependent so see if the relationship is linear (scatter forms a rough line). Binary variables can be distinguished by different markers on scatterplots which helps to investigate patterns within groups.

For a scatterplot: *Graphs → Legacy Dialogs → Scatter/Dot*, then choose *Simple Scatter*. Move the dependent 'Birthweight' to the *Y Axis* box, the independent Gestation to the *X Axis* box and the binary variable Smoker to the *Set Markers by*.



It is a good idea to change the shape of the scatter for one group to make group comparison clearer and increase the size of the scatter so that it can be seen more clearly in a report.

This can be done in the chart editor window which opens if you double-click on the part of the chart you wish to edit. To edit just one group, leave a gap between the two clicks so only one group is highlighted. *Type* changes the symbol being use, *Size* adjusts the overall size of the object (changed to 8) and *Border width* changes the width of the line (changed to 2 here).



The relationship between gestational age and birthweight is clearly linear. The babies of smokers tend to be lighter at each gestational age.

Change the font for all axes to 12 point and add a title before putting the chart in a report.

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

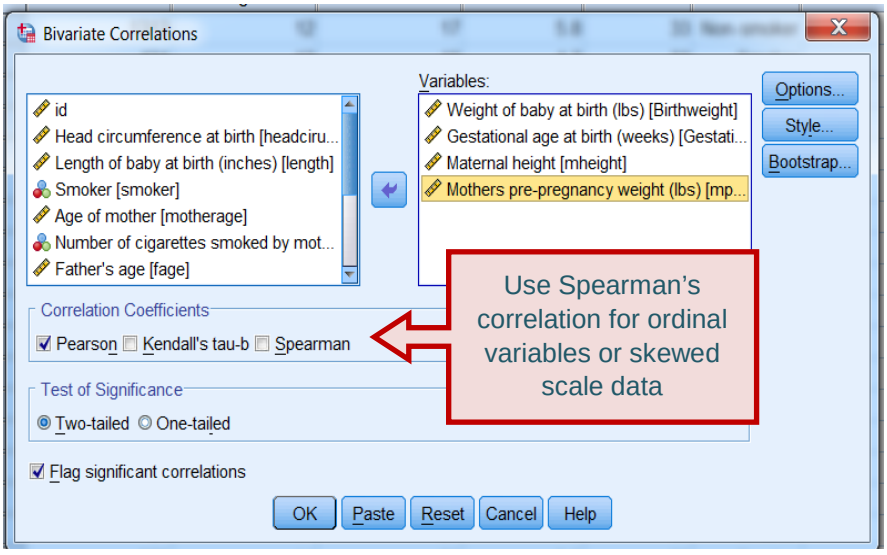
Pearson's correlation coefficient

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

Steps in SPSS

SPSS: Analyse → Correlate → Bivariate Correlation



SPSS can produce multiple correlations at the same time. Using the birth weight dataset, move the variables birthweight, Gestation, mheight and mppwt to the box on the right.

As they are all scale variables, choose the default test *Pearson's* from the **Correlation Coefficients** options.

The output

Correlations

		Weight of baby at birth (lbs)	Gestational age at birth (weeks)	Maternal height	Mothers pre-pregnancy weight (lbs)
Weight of baby at birth (lbs)	Pearson Correlation	1	**.706****	.368*	.390*
	Sig. (2-tailed)		.000	.017	.011
	N	42	42	42	42
Gestational age at birth (weeks)	Pearson Correlation	.706**	1	.231	.251
	Sig. (2-tailed)	.000		.141	.110
	N	42	42	42	42
Maternal height	Pearson Correlation	.368*	.231	1	**.671****
	Sig. (2-tailed)	.017	.141		.000
	N	42	42	42	42
Mothers pre-pregnancy weight (lbs)	Pearson Correlation	.390*	.251	.671**	1
	Sig. (2-tailed)	.011	.110	.000	
	N	42	42	42	42

**. Correlation is significant at the 0.01 level (2-tailed).

*. Correlation is significant at the 0.05 level (2-tailed).

P-value = 0.011
If p < 0.05, there's evidence that r is NOT 0

There's a strong relationship between height and weight of the mother (r=0.671)

Reporting Correlation

The test for correlation tests the null hypothesis that r = 0 not whether or not there is a strong relationship and is highly influenced by sample size. This means that for large samples, a weak correlation can be classified as significant. When writing up use the p-value to identify the existence of a relationship and the correlation coefficient to measure the strength of the relationship.

Pearson's correlation was carried out to look for relationships between the variables birthweight, gestational age, height and weight of mother. There was significant evidence of a relationship between birthweight and gestational age (r = 0.709, p< 0.001), height of mother (r = 0.368, p = 0.017) and weight of mother (r = 0.39, p = 0.011). Gestational age is strongly related to birthweight and is moderately related to the others. There was also evidence of a relationship between the weight and height of the mothers (r = 0.671, p < 0.001) which is moderate.

If the data is not normally distributed or ordinal there are alternative methods which can be used. Spearman's rank correlation coefficient is a non-parametric statistical measure of the strength of a monotonic relationship between paired data. The notation used for the sample correlation is r_s. 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.