The following resources are associated:

Dataset ‘*Birthweight\_reduced.csv’*, *Simple and Multiple linear regression in SPSS* resources

Scatterplots and correlation in SAS

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

outlier

* 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.

**Data**: The data set ‘*Birthweight\_reduced.csv’* contains details of 42 babies and their parents at birth. The dependant variable is Birth weight (lbs) and the independent variables on this sheet are gestational age of the baby at birth (in weeks), mothers height (mheight) and weight (mppwt) and whether or not the mother smokes (smoker=1).

Download the data into SAS and give it a name e.g. birthweight\_data. Use **proc** **format** to set up the value labels for whether the mother smokes or not: **proc** **format**; value smokeS **1**='Smoker' **0**='Non-smoker'; run;

**Producing scatterplots in SAS**

To produce a basic scatterplot in SAS with a regression line use:

title 'Scatterplot of gestational age and birthweight';

**proc** **sgplot** data = birth;

 scatter x=Gestation y=Birthweight;

 reg x=Gestation y=Birthweight;

run;

title;

proc sgplot is the procedure that uses SASs' more advanced graphing software and data = tells SAS which data is being used (in this case the pre-defined dataset birth).

scatter requests a scatterplot with the specified x and y variables

reg adds a regression line (line of best fit) which helps visualise the relationship for the (x,y) scatter graph. There is a strong positive relationship between gestational age and birthweight.

The title command simply adds the specified title to all results produced by SAS so long as it hasn't been told to stop. Therefore, remember to end the command using title.

Binary variables can be distinguished by different markers on scatterplots which helps to investigate patterns within groups. To differentiate between groups within the scatter plot data use the group= command and the categorical variable of interest on the scatter line. Here it is whether or not the mother is a 'smoker'.

**proc** **sgplot** data = birth;

 scatter x = gestation y =birthweight/group=smoker

 markerattrs=(symbol=CircleFilled size=**14**) filledoutlinedmarkers;

 reg x = gestation y = birthweight/group=smoker;

 format smoker smokeS.;

**run**;

Separate regression lines by group can be requested by adding the group= statement to the reg line. Here different regression lines are shown; one for Smokers and one for Non-smokers.

Using format smoker smokeS.; allows the key to produce ‘Smoker’ and ‘Non-smoker’ in the legend instead of 1 and 0.

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.

markerattrs allows you to modify the points shape and filledoutlinedmarkers adds a border around markers with and only with a filled-out condition applied i.e. symbol=CircleFilled

A list of symbol properties can be found here <http://support.sas.com/documentation/cdl/en/grstatgraph/63878/HTML/default/viewer.htm#p0icsws0thpaghn1t12m27fyqnjc.htm>

**Interpretation**: The relationship between gestational age and birthweight is clearly linear but the impact of gestational age is stronger for smokers. The babies of smokers tend to be lighter at each gestational age but this difference is larger for lower gestational ages.

Note: When carrying out regression, scatterplots should be produced for each independent with the dependent so see if the relationship is linear (scatter forms a rough line).

**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) and approximately normally distributed.

|  |  |  |
| --- | --- | --- |
| **Assumptions** | **How to check** | **What to do if assumption is not met** |
| Continuous data for each variable | Check data types | If ordinal data use Spearman’s correlation |
| Linearly related variables | Scatter plot | Transform data |
| Both variables are normally distributed | Histograms or QQ plots of variables (see ‘*Checking normality in SAS resource’*) | Use Spearman’s correlation |

**Steps in SAS**

Here relationships between mothers weight, height and smoking status and birthweight of babies are investigated. As they are all scale variables, use the *Pearson’s* correlation coefficient which is the default. The **proc** **corr** procedure allows for multiple correlation coefficients to be calculated simultaneously and if ods graphics is on, scatterplots for each combination are produced.

ods graphics on; **run**;

**proc** **corr** data= data = birth nosimple plots=matrix;

var Birthweight Gestation mheight mppwt; **run**;

The data being used is the birthweight\_data set and nosimple tells SAS not to present the simple statistics. The plots=matrix produces scatterplots for every combination to check linearity.

The var statement allows the calculation of multiple correlations that encompasses all combinations of all the listed variables.

Note: If you wish to calculate correlations with one variable (in the var line) and all the others, e.g. birthweight with all the other variables, use the with statement:

**proc** **corr** data= birth nosimple;

var Birthweight;

with Gestation mheight mppwt; **run**;

**The output**

Although not shown here, check that the relationships between variables of interest are linear using the scatterplots produced. These scatterplots are for checking an assumption rather than display. Refer to the scatterplots section of the handout if you wish to use them in reports.



There’s a strong relationship between height and weight of the mother (r=0.67)

P-value < 0.001

If p < 0.05, there’s evidence that r is NOT 0

**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. When reporting correlation coefficients, two decimal places should be used.

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 = 0.71, p< 0.001), height of mother (r = 0.37, 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 = 0.67, p < 0.001) which is moderate.

Skewed or ordinal data

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 rs and it is interpreted in the same way as Pearson’s correlation.

Extra information Spearmen’s test is conducted from,

**proc** **corr** data= birth spearman;

Use Spearman’s correlation for ordinal variables or skewed scale data

var Birthweight Gestation mheight mppwt ;

**run**;