*The following resources are associated:*

*Scatterplots and correlation, Checking normality in SPSS and the SPSS dataset Birthweight\_reduced.sav’*

Simple linear regression in SAS

Dependent (outcome) variable: Continuous (scale)

Independent (explanatory) variables: Continuous (scale)

Common Applications: Regression is used to (a) *look for significant relationships* between two variables or (b) *predict* a value of one variable for a given value of the other.



**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).

Download the data into SAS and save it in a specified folder e.g. home, under a given name e.g. birthweight\_data. You can give this an even shorter name in the temporary file e.g. birth and label the variables using the following.

**data** birth;

**set** home.birthweight\_data;

 label Birthweight='Birthweight (lbs)'

 gestation='Gestational age of baby at birth (weeks)'

Before carrying out any analysis, investigate the relationship between the independent and dependent variables by producing a scatterplot of the two variables (see ‘Scatterplots and correlation in SAS’ resource). The scatterplot suggests a strong positive linear relationship between gestational age and birthweight.

Residuals =

actual y – predicted y

Simple linear regression quantifies the relationship between two variables by producing an equation for a straight line of the form $y=a+βx$which uses the independent variable (x) to predict the dependent variable (y).

Regression involves estimating the values of the gradient and intercept  of the line that best fits the data. This is defined as the line which minimises the sum of the squared residuals.

A **residual** is the difference between an observed dependent value and one predicted from the regression equation.

**Assumptions for regression**

|  |  |  |
| --- | --- | --- |
| Assumptions | How to check | What to do if the assumption is not met |
| 1) The relationship between the independent and dependent variable is linear  | *Scatterplot*: scatter should form a line in the plot rather than a curve or other shape  | Transform either the independent or dependent variable |
| 2) Residuals should be approximately normally distributed  | Look at the *histogram of residuals* produced by SAS. Note: if there are no graphs check if you have written and run, ods graphics on;  | Transform the dependent variable |
| 3) Homoscedasticity: Scatterplot of standardised residuals and standardised predicted values shows no pattern (scatter is roughly the same width as y increases)  | This shape is bad since the variation in the residuals (up and down) is not constant (variance is increasing) | Transform the dependent variable |

**Other assumptions covered in more detail on the 'Further regression' resource:**

4) Independent observations (adjacent values are not related). This is only a possible problem if measurements are collected over time.

5) No observations have a large overall influence (leverage). Look at individual Cook’s and Leverage values. Interpretation of this is not included on this sheet.

**Steps in SAS**

In order to obtain the graphs to check the assumptions of regression, ensure that ods graphics is activated. 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.

ods graphics on;

title 'Regression output for gestation and birthweight';

**run**;

The following code should be used to run simple linear regression:

**proc** **reg** data = birth corr;

 model birthweight=gestation;

**run**;

 title;

**proc** **reg** is the procedure which produces simple or multiple linear regression output. The key prompt corr informs SAS to also calculate correlation coefficients.

The model statement tells SAS to create a model for a regression test with the variable on the left-hand side being the dependant (outcome) variable and all variables on the right-hand side being the independent variables.

**Output**

The correlation between gestational age and birthweight is 0.71 suggesting a strong positive relationship. The longer the pregnancy, the heavier babies tend to be.

The **Parameter Estimates** table below is the most important table in the regression output. It contains the coefficients for the regression equation and tests of significance.

The Parameter Estimate column, gives us the values of the gradient and intercept terms for the regression line.

The model is: **Birth weight (y) = -6.66 + 0.355 \*(Gestational age)**

The gradient is tested for significance. If there is no relationship, the gradient of the line  would be 0 and therefore every baby would be predicted to be the same weight. The main row of interest is the ‘Gestation’ row as this tests the relationship between gestation and the dependent variable birthweight. The test statistic is 6.31 and the p-value is calculated using a t distribution with 1 degree of freedom. The p-value for Gestational age (p < 0.0001) is contained in the column ‘$Pr>\left|t\right|$’. It is much less than the level of significance of 0.05 so there is significant evidence to suggest that the gradient is not 0 and therefore there is a relationship between gestational age and birthweight.

If the independent variable is significant, explain the relationship using the gradient in the parameter estimate column. This represents the increase in the dependent for a one unit increase in the independent variable. Here a baby increases in weight by 0.36 lbs for each extra week of gestation.

The key information from this table is the R2 value of 0.499. For simple linear regression this is just the correlation coefficient squared. This indicates that 50% of the variation in birth weight can be explained by the model containing only gestation. This is quite high so predictions from the regression equation are fairly reliable. It also means that 50% of the variation is still unexplained so adding other independent variables could improve the fit of the model.



**Checking the assumptions for this data**

The assumption of linearity is checked with the original scatterplot of the dependent and independent variables. Any non-linear patterns in the scatter mean that linear regression is not suitable for this data. The scatter is approximately linear here so linear regression is an appropriate technique.

If you have turned on the ods graphics before running regression, SAS will automatically generate a large number of plots for checking. For a basic linear regression only two are needed; the histogram of residuals (or QQ plot) and the scatterplot of predicted values and residuals so only include these in the Appendix.

The residuals (differences between the observed values and values predicted by the model for each individual) should be approximately normally distributed like these. They would need to be really skewed to be a problem (See ‘*Checking normality in SAS’* resource for more details on this).



The next plot assesses whether there is any pattern in the size of residuals as predicted values increase. We want to see random scatter to meet the assumption. There is no pattern in the scatter. The width of the scatter as predicted values increase is roughly the same so the assumption has been met.

If you wish to just produce these two rather than the block of 9 plots, this can be specified in the **proc** **reg** procedure using the following:

**proc** **reg** data=birth plots=residualhistogram plots=residualbypredicted

**Reporting regression**

Tips: Most of the tables produced by SAS have only one value of interest e.g. correlation coefficient or R2. Do not include these tables in a main report and just report the values to no more than two decimal places within a sentence with explanation. You can include the parameter estimates table particularly for multiple regression although here it is probably not necessary as key numbers can be included within the text. It is useful to include a labelled scatterplot of the dependent and independent variables but charts for checking assumptions should be included within the Appendix.

Suggested write up

*Simple linear regression was carried out to investigate the relationship between gestational age at birth (weeks) and birth weight (lbs). The scatterplot of gestation and birthweight and Pearson’s correlation coefficient of 0.71 suggest a strong positive linear relationship. Simple linear regression showed a significant relationship between gestation and birth weight (t(1)=6.31, p < 0.001) with the weight of a baby increases by 0.36 lbs for each extra week of gestation. The R2 value suggested that 50% of the variation in birth weight can be explained by the model containing only gestation.*

*The scatterplot of standardised predicted values verses standardised residuals showed that the data met the assumption of homogeneity of variance and the residuals were approximately normally distributed.*