The following resources are associated: ‘Simple linear regression in SAS’, ‘Scatterplots and correlation in SAS’, ‘Checking normality in SAS’ and the dataset ‘*Birthweight\_reduced.csv’*

Multiple linear regression in SAS

Dependent variable: Continuous (scale)

Independent variables: Continuous (scale) or binary (e.g. yes/no)

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



**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 are gestational age of the baby at birth (in weeks), mothers height (mheight) and pre-pregnancy weight (mppwt) and whether or not the mother smokes (smoker=1).

Download the data into SAS and label the variables and values of categorical 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**;

**Linear regression in SAS**

It is recommended that you read the ‘Simple linear regression in SAS’ resource in addition to this as it contains more details for certain aspects such as checking assumptions.

In order to obtain the graphs to check the assumptions of regression, ensure that ods graphics on is activated before running regression analysis. To run linear regression use the **proc** **reg** procedure:

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

model birthweight=gestation mppwt smoker/VIF;

**run**;

Specify the data, which is called birth here and request correlations using corr. SAS automatically produces a set of graphs to check assumptions but the only two specific plots required are the histogram of residuals and the scatterplot of residuals versus predicted values. The model statement tells SAS to create a regression model 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. The VIF, (*variance inflation factor*), is added after the / in the model statement to help check for similarity between independent variables.

**Correlation**

The first table just states how many individuals are in the data set. The second table shows the Pearson correlation coefficients (r) between each pair of variables. The interpretation of these are explained in more detail in ‘Scatterplots and correlation in SAS’ resource.

The output shows that gestational age has the strongest relationship with birthweight (r=0.71) but pre-pregnancy weight (r=0.39) also has a moderate positive relationship with birthweight. The relationship between the two continuous independents is weak (r=0.25).

Scatterplots should be produced for each independent with the dependent so see if the relationship is linear (scatter forms a rough line).

**Multiple regression**



The F-test in the ANOVA table tests the model as a whole, with all the independent variables. The p-value for the model (p < 0.0001) contained in the column ‘$Pr>F$’ is highly significant.

The **Parameter Estimates** table below is the most important table in the regression output and the only one which should appear in a report. It contains the coefficients for the regression equation and tests of significance for individual variables (parameters).

In multiple regression, significance is based on unique variation after removing the effect of the other variables, so any effect observed is down to that variable rather than other differences between groups. The test statistics are contained in the t value column and the p-values for the individual variables in the column ‘$Pr>\left|t\right|$’.

After controlling for the other variables, pre-pregnancy weight (p=0.03), whether the mother is a smoker (p=0.018) and gestation (p<0.0001) are significant predictors of birthweight. If the independent value is significant, explain the relationship between the independent and dependent variables using the *Parameter Estimates* B.

The ‘**B**’ column (parameter estimates) in the coefficients table, gives us the coefficients for each independent variable in the regression model. The model is:

**Birthweight (y) = -7.165 + 0.313 \*(Gestation) – 0.665\*(Smoker) + 0.02\*(mppwt)**

For gestation, there is a 0.313 lb increase in birthweight for each extra week of gestation. For each extra pound (lb) a mother weighs, the baby’s weight increases by 0.02 lbs. A binary variable such as Smoker coded as 0 and 1, the coefficient only applies for the group coded as 1 (smokers). Here smokers have babies who weigh 0.67 lbs less than non-smokers on average.

The only value needed from this table is the R2 value of 0.61 which indicates that 61% of the variation in birth weight can be explained by the model containing gestation, pre-pregnancy weight and whether or not the mother smokes. This is quite high so predictions from the regression equation are fairly reliable. It also means that 39% of the variation is still unexplained so adding other independent variables could improve the fit of the model. The model on the simple linear regression sheet which only has gestation explained 50% of the variation so adding pre-pregnancy weight and smoking status explains an extra 11% of the variation in birthweight for the individuals in the dataset.

Assumptions for multiple linear regression

All the assumptions for simple linear regression (with one independent variable) also apply for multiple regression with one addition. If two of the independent variables are highly related, this leads to a problem called multicollinearity. As the tests in the Parameter estimates table test unique variation explained by the variable, highly related variables will share a lot of variation meaning there are less likely to be significant.

To investigate possible multicollinearity, first look at the correlation coefficients for each pair of continuous (scale) **independent** variables **in the correlation table**. Correlations of 0.8 or above suggest a strong relationship and only one of the two variables are needed in the regression analysis.For this example, the correlation between gestational age at birth and the pre-pregnancy weight of the mother is 0.25 suggesting a weak positive correlation between them.

The correlations only test the relationships between pairs of variables but the **Variance Inflation** column in the main parameter estimates table gives a measure of how much each independent is related to all the other independents collectively. It uses the R2 of a model where the independent is the dependent.VIF= $\frac{1}{1-R^{2}}$

For example, the *variance inflation factor* (VIF) for gestational age of 1.08 has been calculated using the R2 of the model gestation = pre-pregnancy weight + smoker.

The VIF scores should be close to 1 but under 5 is fine and 10+ suggests high collinearity so the variable may not be needed. 2 indicates 50% of the variation of that variable is explained by the others, but 5 means it is not necessary. All the values in this analysis have scores close to 1 so no variables need to be removed.

**Checking the assumptions for this data**

As with simple linear regression, check the relationship between each continuous independent variable and the dependent variable is roughly linear using scatterplots.

If ods grahics has been turned on, numerous graphs are produced as part of the regression analysis however, you only need the histogram of the residuals (which shoud be approximately normal), and the scatterplot of predicted values with residuals (which should show random scatter) shown in this section. You can specify these charts using:

 plots=residualhistogram plots=residualbypredicted

|  |  |
| --- | --- |
| **Normality of residuals**The residuals are approximately normally distributed.  | **Homoscedasticity**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 the residuals are very skewed or a pattern appears in the second plot, you can try adding more variables to explain the variation or transform the dependent variable.  |

**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. The parameter estimates table is the only table which should appear in the main report with explanation of the key information underneath. 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.

*Multiple linear regression was carried out to investigate the relationship between gestational age at birth (weeks), mothers’ pre-pregnancy weight and whether she smokes and birth weight (lbs). There was a significant relationship between gestation and birth weight (t(1)=5.93, p < 0.001), smoking and birth weight (t(1)=-2.48, p = 0.017) and pre-pregnancy weight and birth weight (t(1)=2.26, p = 0.03). For gestation, there was a 0.31 lb increase in birthweight for each extra week of gestation. For each extra pound (lb) a mother weighs, the baby’s weight increases by 0.02 lbs and smokers have babies who weigh 0.67 lbs less than non-smokers.*

*The R2 value was 0.61 so 61% of the variation in birth weight can be explained by the model containing gestation, pre-pregnancy weight and whether the mother smokes or not.*

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