community project

encouraging academics to share statistics support resources



stcp-marshall-MultipleRegressionS

The following resources are associated: Simple linear regression in SPSS, Scatterplots and correlation in SPSS, Checking normality in SPSS and the SPSS dataset '*Birthweight_reduced.sav*'

## Multiple linear regression in SPSS

**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.sav' 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) and variables relating to the mother (mothers' height and weight as well as whether or not she smokes).

| Birthweight | Gestation | smoker | motherage | mnocig | mheight | mppwt |
|---|---|---|---|---|---|---|
| 5.8 | 33 | 0 | 24 | 0 | 58 | 99 |
| 4.2 | 33 | 1 | | 7 | 63 | 109 |

Mother smokes = 1

Weight of mother before pregnancy

The **Simple linear regression in SPSS** resource should be read before using this sheet.

## Assumptions for regression

All the assumptions for simple 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. This causes problems with the analysis and interpretation. To investigate possible multicollinearity, first look at the correlation coefficients for each pair of continuous (scale) variables. Correlations of 0.8 or above suggest a strong relationship and only one of the two variables is needed in the regression analysis. SPSS also provides *Collinearity diagnostics* within the **Statistics** menu of regression which assess the relationships between each independent variable and all the other variables.

© Ellen Marshall
Sheffield Hallam University

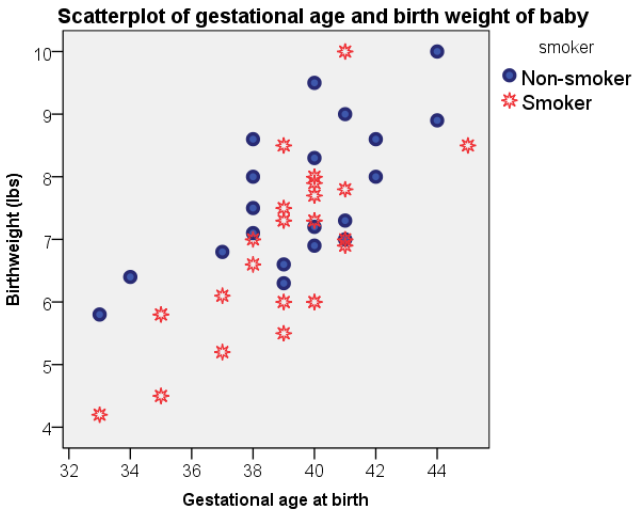
Reviewer: Jean Russell
University of Sheffield

To calculate Pearson's correlation co-efficients use
*Analyze → Correlate → Bivariate* and move Birthweight, Gestation, mheight and mppwt to the *variables* box.

The output shows that gestational age has a strong relationship with birthweight (r = 0.706), maternal height (r = 0.368) and pre-pregnancy weight (0.39) are moderately related with birthweight. The relationship between maternal height and weight is strong (r = 0.691) but not above 0.8.

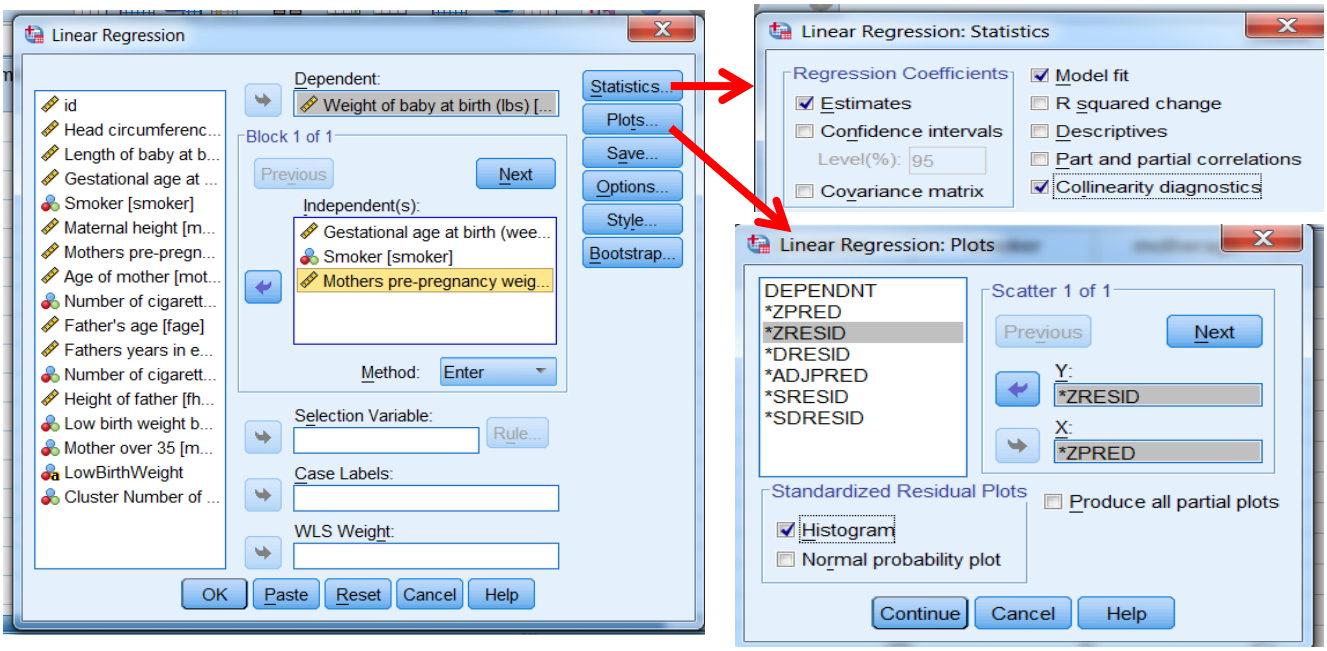
Scatterplots should be produced for each independent with the dependent so see if the relationship is linear (scatter forms a rough line). Binary variables can be distinguished by different markers on scatterplots which helps to investigate patterns within groups.

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




## Steps in SPSS

To run a regression, go to *Analyze → Regression → Linear*
Move 'Birth weight' to the *Dependent* box and 'Gestational age at birth', 'Smoker' and 'mppwt' (mothers' pre-pregnancy weight) to the *Independent(s)* box. Multicollinearity can be checked using the *Collinearity diagnostics* in the **Statistics** menu. In the **Plots** menu, move ZRESID to the *Y* box and ZPRED to the *X* box to check the assumption of homoscedasticity. Request the *Histogram* to check the normality of residuals.



## Output

The Coefficients table contains the coefficients for the regression equation (model), tests of significance for each variable and collinearity statistics.

**Coefficients[a]**

| Model | | Unstandardized Coefficients | | Standardized Coefficients | t | Sig. | Collinearity Statistics | |
|---|---|---|---|---|---|---|---|---|
| | | B | Std. Error | Beta | | | Tolerance | VIF |
| 1 | (Constant) | **-7.165** | 2.107 | | -3.400 | **.002** | | |
| | Gestational age at birth (weeks) | **.313** | .053 | .623 | 5.926 | **.000** | .928 | 1.078 |
| | Smoker | **-.665** | .268 | -.253 | -2.485 | **.017** | .990 | 1.010 |
| | Mothers pre-pregnancy weight (lbs) | **.020** | .009 | .237 | 2.261 | **.030** | .936 | 1.068 |

a. Dependent Variable: Weight of baby at birth (lbs)

The **Sig** column contains the p-values for each of the independent variables. The hypothesis being tested for each is that the coefficient (B) is 0 after controlling for the other variables. For example, the effects of gestational age and smoking are removed before assessing the relationship between the weight of the mother and the weight of the baby. A p-value < 0.05, provides evidence that the coefficient is different to 0. Gestational age ($p < 0.001$), smoker ($p = 0.017$) and mothers' pre-pregnancy weight ($p = 0.03$) are all significant predictors of birthweight. If the independent value is significant, explain the relationship between the independent and dependent variables using the *Unstandardized Coefficient* B.

The '**B**' column 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. Here smokers have babies who weigh 0.665 lbs less than non-smokers.

Collinearity statistics measure the relationship between multiple independent variables by giving a score for each independent. The "tolerance" is an indication of the percent of variance in an independent that cannot be accounted for by the other independent variables, hence very small values indicate that an independent variable is redundant. The VIF, which stands for *variance inflation factor*, is (1 / tolerance). 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. All the values in this analysis have scores close to 1.

The $R^2$ value of 0.61 indicates that 61% of the variation in birth weight can be explained by the model containing gestation, smoker and pre-pregnancy weight. 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.

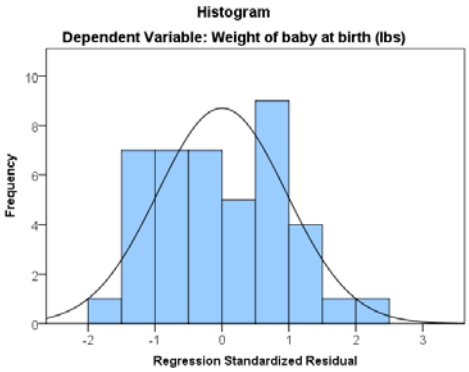
**Model Summary[b]**

| Model | R | R Square | Adjusted R Square | Std. Error of the Estimate |
|---|---|---|---|---|
| 1 | .781[a] | **.610** | .580 | .8622 |

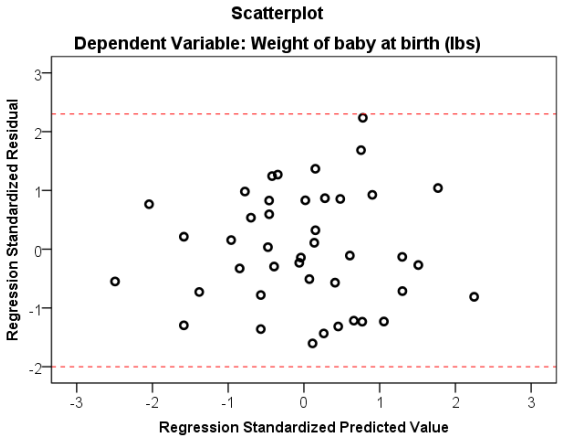
## Checking the assumptions for this data

### Normality of residuals

The residuals are approximately normally distributed. Tip: If you are not sure if the histogram is normally distributed, change the number of intervals in the *Binning* section of the Bar **Properties**. Try 7 here.




### 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.

## Reporting regression

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 ($p < 0.001$), smoking and birth weight ($p = 0.017$) and pre-pregnacy weight and birth weight ($p = 0.03$). For gestation, there was 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 and smokers have babies who weigh 0.665 lbs less than non-smokers.

The $R^2$ 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.