community project

encouraging academics to share statistics support resources

All stcp resources are released under a Creative Commons licence

stcp-marshall-furtherRegressionS

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

Outliers, Durbin-Watson and interactions for regression in SPSS

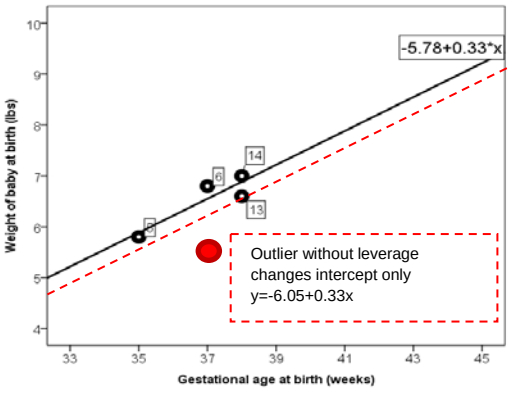
Dependent variable: Continuous (scale)
Independent variables: Continuous/ binary

Data: The data set '*Birthweight reduced.sav*' contains details of 42 babies and their parents at birth. The dependant variable is birthweight (pounds = lbs) and the two independent variables are the gestational age of the baby at birth (in weeks) and whether or not the mother smokes (0 = non-smoker, 1 = smoker).

Investigating outliers and influential observations

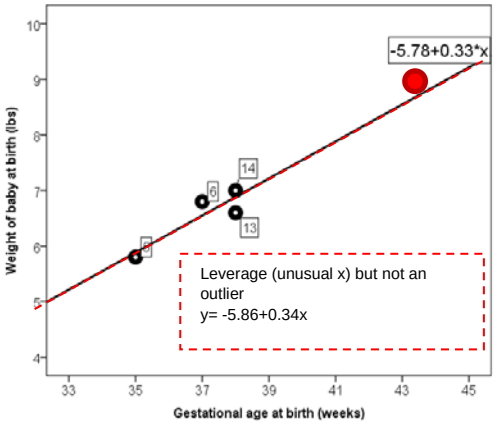
An assumption of regression is that there are no influential observations. These are extreme values which pull the regression line towards them therefore having a significant impact on the coefficients of the model.

Outliers: Outliers are observations where the observed dependent value does not follow the general trend given the independent value (unusual y given x). In this situation, the residual for that observation is likely to be large unless it is also influential and has pulled the line towards it. A residual is the difference between observed and predicted values and standardised residuals (with a mean of 0 and SD of 1) can be requested in SPSS. Approximately 5% of standardised residuals will be outside ± 1.96 and 0.3% of values are classified as extreme outliers which are outside ± 3. Large samples are more likely to contain extreme outliers just by chance.



Deleted residuals are the residuals obtained if the regression was repeated without the individual observation.

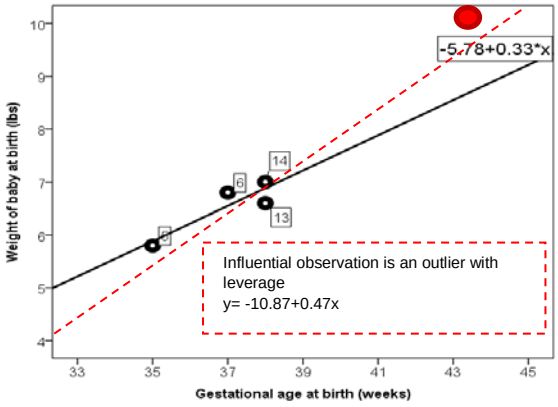
Leverage: Leverage relates to subjects with unusual values of the independent variable which have the potential to influence the slope greatly. An observation with high leverage will pull the regression line towards it. Calculations compare the independent values with their mean. The average leverage score is calculated as (k + 1)/ n where k is the number of independent variables in the model and n is the number of observations. Observations with high leverage will have leverage scores 2 or 3 times this value.



© Ellen Marshall and Sofia Maria Karadimitriou
University of Sheffield

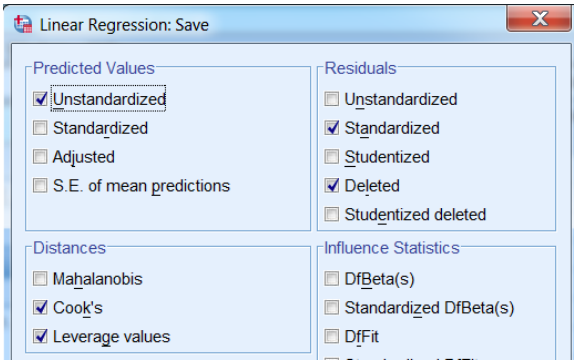
Reviewer: Jean Russell
University of Sheffield

Influence: An influential observation is one which is an outlier with leverage and affects the intercept and slope of a model significantly. Calculations are based on how the predictions would differ if the observation was not included.

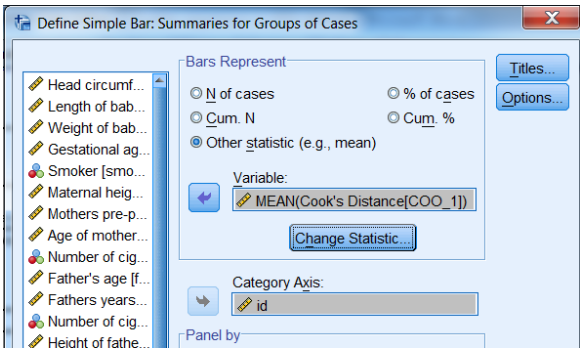


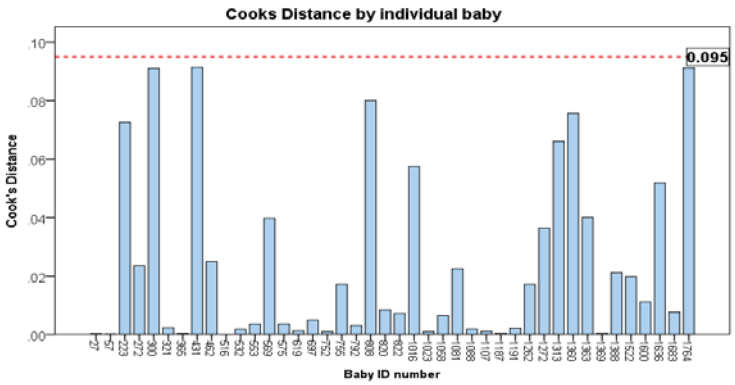
Cooks distance: This is calculated for each individual and is based on the squared differences between the predicted values from regression with and without an individual observation. A large Cook's Distance indicates an influential observation. Compare the Cooks value for each observation with 4/n where n is the number of observations. Values above this indicate observations which could be a problem.

Carry out simple linear regression through *Analyze → Regression → Linear* with Birthweight as the *Dependent* variable and Gestation as the *Independent*.
In the **Save** menu, select *Standardised residuals*, *Cook's* and *Leverage values*. The values for each individual will be added to the data set.

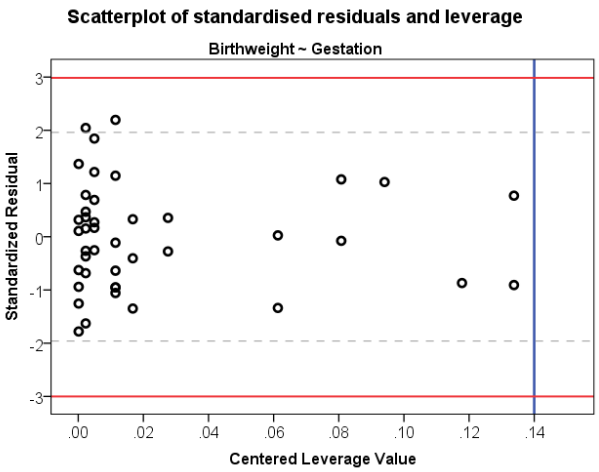


Then produce a bar chart of the cooks distances by ID. To produce a bar chart of Cook's distance for each observation, go to *Graphs → Legacy Dialogs → Bar,* choose *Other statistic (e.g. mean)* and move Cook's distance to the *Variable* box and id to the category axis. There's only one observation for each baby so the mean is the value.





The cut off for Cook's is 4/n so here it is 4/42 = 0.095 which can be added to the chart as a reference line to make it easier to see. All of the Cook's Distances are below this line.

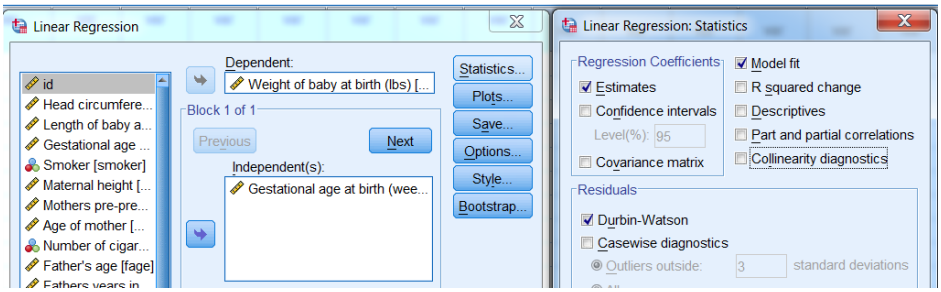


To check for outliers and leverage, produce a scatterplot of the Centred Leverage Values and the standardised residuals. There are two observations with standardised residuals outside ± 1.96 but there are no extreme outliers with standardised residuals outside ± 3. Leverage values 3 times (k + 1)/ n are large where k = number of independent variables. The cut off here is 3*(1+1)/42 = 0.14. No observations have leverage values above 0.14
If an observation has a very large leverage score, try running the model with and without the value to

see how much the coefficients in the model change.

The Durbin Watson test

One of the assumptions of regression is that the observations are independent. If observations are made over time, it is likely that successive observations are related. If there is no autocorrelation (where subsequent observations are related), the Durbin-Watson statistic should be between 1.5 and 2.5. Carry out simple linear regression through *Analyze → Regression → Linear* with Birthweight as the *Dependent* variable and Gestation, the *Independent*. The *Durbin-Watson* Statistic is found in the **Statistics** menu.



Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	**Durbin-Watson**
1	.706[a]	.499	.486	.9530	**2.390**

The Durbin-Watson statistic is 2.39 which is between 1.5 and 2.5 and therefore the data is not autocorrelated.

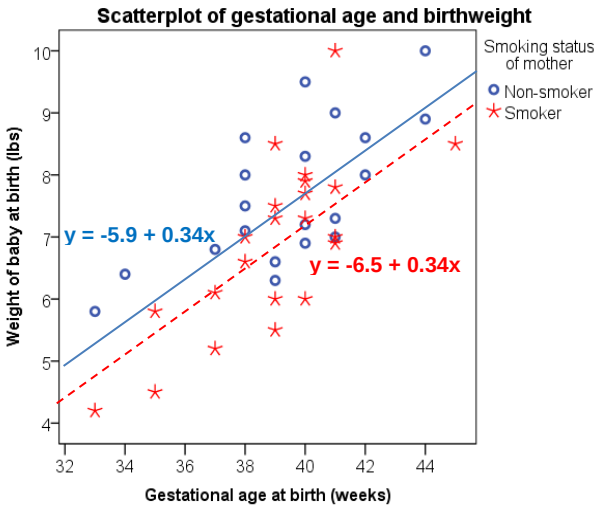
Interactions in regression

An interaction is the combined effect of two independent variables on one dependent variable. Interactions in SPSS must be calculated before including in a model. The following example uses the birthweight data with birthweight as the dependent variable and gestation and whether or not the mother smokes (0 = no, 1 = yes) as the independent variables.

The scatterplot to the right shows the regression lines for birthweight (y) **without an interaction** between the two independents in the model.

The continuous x variable 'Gestational age' contributes to the slope of the line. For both lines, the slope is 0.34 so a baby increases in weight by 0.34 lbs for each extra week of gestation. The binary variable 'Smoking status of mother' changes the intercept so smokers/ non-smokers have a different intercept.

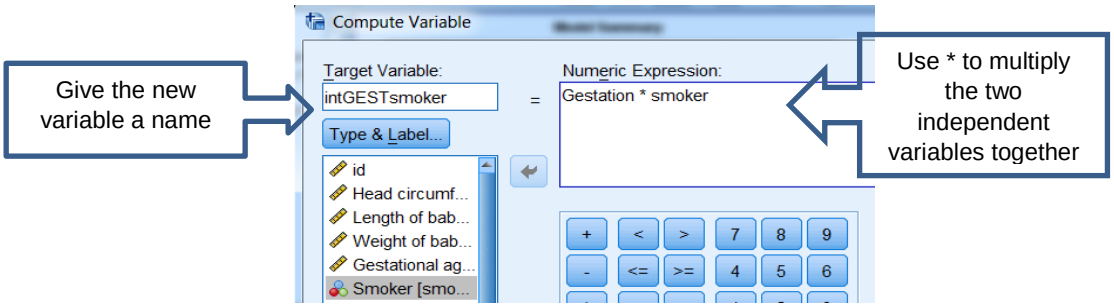
The lines are parallel but smokers tend to have lighter babies at each gestational age (intercept is 0.6 lbs lower).



If there is an interaction between gestational age and smoking status, the slopes of the two lines would be different. This means that the effect of gestational age (x) on birthweight (y) is different depending on whether or not the mother smokes.

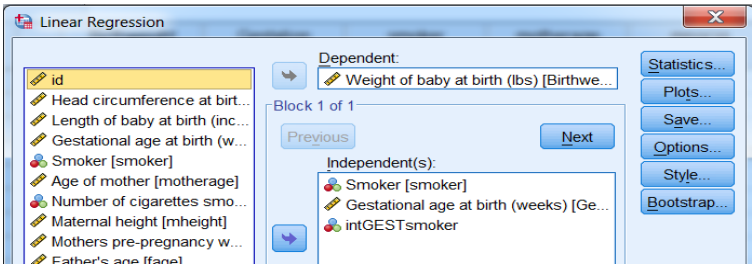
Including interaction terms in regression

For standard multiple regression, an interaction variable has to be added to the dataset by multiplying the two independents using *Transform → Compute variable*



To run a regression model: *Analyze → Regression → Linear*

Run the regression model with 'Birth weight' as the *Dependent* and gestational age, smoker and the new interaction variable intGESTsmoker as *Independent(s)*.



The **Coefficients** table contains the coefficients for the model (regression equation) and p-values for each independent variable. The output shows that the interaction is not significant so the main effects can be interpreted. Only gestation is significant ($p < 0.001$) whilst the interaction term is in the model. The regression analysis can be repeated without the interaction term if it is not significant.

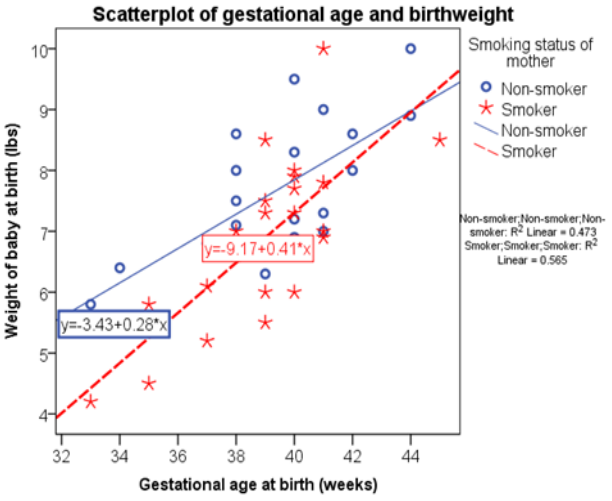
Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	-3.431	2.920		-1.175	.247
	Smoker	-5.734	4.206	-2.180	-1.363	.181
	Gestational age at birth (weeks)	.282	.074	.560	3.818	.000
	intGESTsmoker	.130	.107	1.932	1.213	.233

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

Calculations for the equations of the lines with an interaction term

The regression model uses the *Unstandardized Coefficients*



Birth weight y = -3.431 –5.734*(smoker) + 0.282*(Gest) + 0.13*(Smoker*Gest)

For non-smokers, smoker = 0 so the model becomes y = -3.431 + 0.282(Gest)

For smokers, smoker = 1: y = -3.431 –5.734*(1) + 0.282*(Gest) + 0.13*(1*Gest)

= -9.165 + 0.412*(Gest)

Note: Where there are interactions between two scale variables, the coefficient of the interaction can be quite small and more difficult to interpret.