statstutor

community project

encouraging academics to share statistics support resources

stcp-karadimitriou-furtherRegressionR

The following resources are associated: the dataset 'Reduced birthweight.csv' and the Multiple linear regression in R script file. Simple and multiple linear regression in R resources.

Outliers, Durbin-Watson and interactions for regression in R

Dependent variable: Continuous (scale)

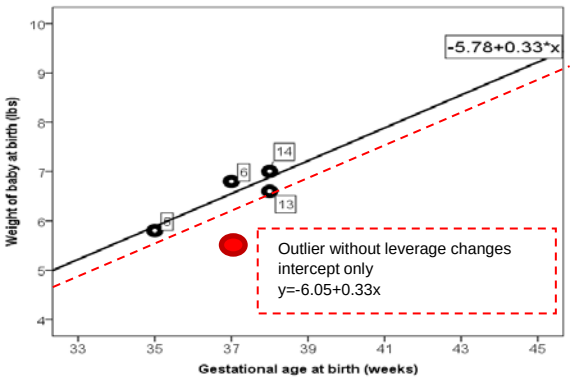
Independent variables: Continuous/ binary

Data: The data set '*Birthweight reduced.csv'* contains details of 42 babies and their parents at birth. The dependant variable is Birthweight (pounds = lbs) and the independent variables are the gestational age of the baby (weeks) and whether the mother smokes (0 = non-smoker, 1 = smoker).

Investigating outliers and influential observations

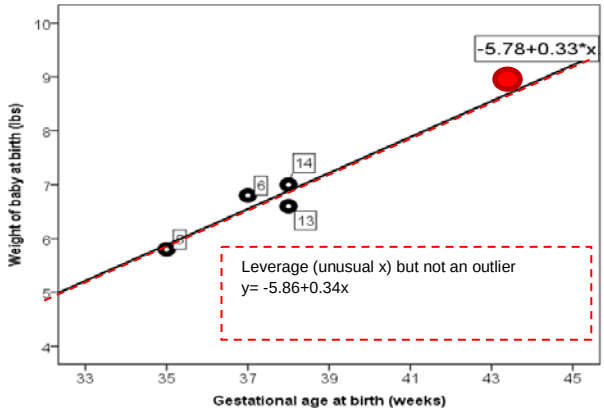
An assumption of regression is that there are no influential observations. These are extreme values which pull the regression line towards them therefore having a significant impact on the coefficients of the model.

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.



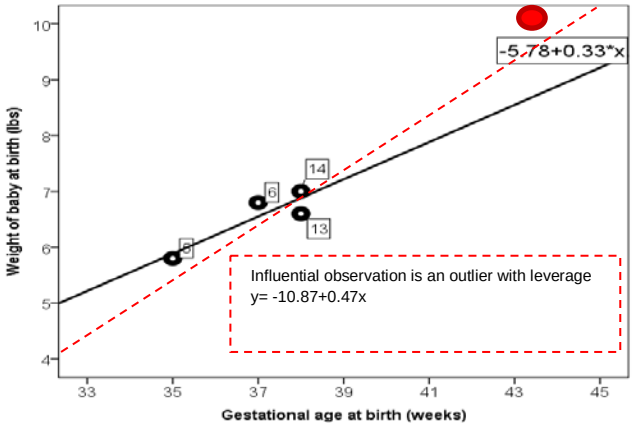
Deleted residuals are the residuals obtained if the regression was repeated without the individual observation.

Leverage: An observation with high leverage will pull the regression line towards it. 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.



© Sofia Maria Karadimitriou and Ellen Marshall
University of Sheffield

Reviewer: Jim Bull
University of Swansea



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.

Steps in R

Open the birthweight reduced dataset from a csv file, call it `birthweightR`, then attach the data.

```
birthweightR<-read.csv("D:\\Birthweight reduced.csv",header=T)
attach(birthweightR)
```

Tell R that 'smoker' is a factor and attach labels to the categories e.g. 1 is a smoker.

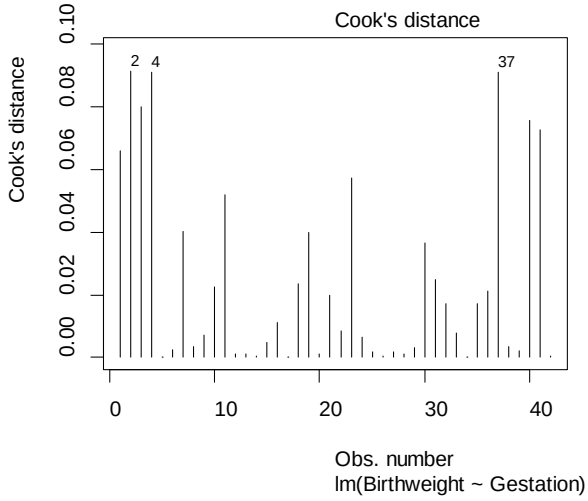
```
smoker<-factor(smoker,c(0,1),labels=c('Non-smoker','Smoker'))
```

Fit the regression model from the Simple linear regression resource using the `lm(dependent~Independent)` command and give it a name (reg1).

```
reg1<-lm(Birthweight~Gestation)
```

To produce a bar chart of Cook's distance for each observation:

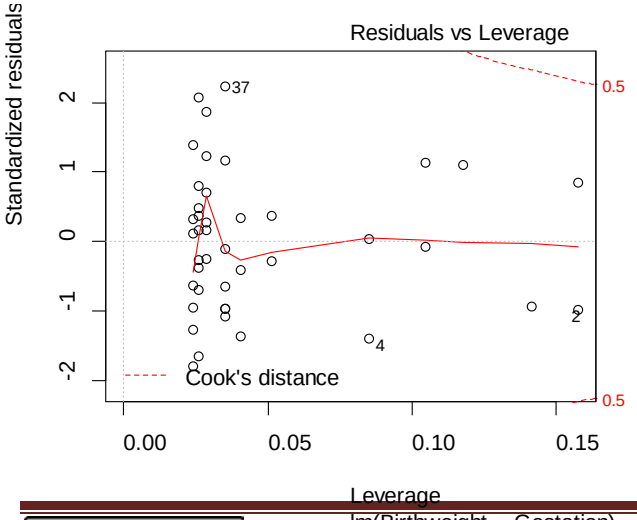
```
plot(reg1, which = 4)
```



R identifies observation with Cooks > 4/n where n = number of observations.
The bottom left graph shows the values of the Cook's Distances and we can see that three observations could be problematic for our model.

To produce a scatterplot of leverage values against standardised residuals:

```
plot(reg1, which = 5)
```



R identifies outliers outside ± 1.96 but extreme outliers will have standardised residuals outside ± 3. There are none here.
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. R identifies observation 2 as an observation with high leverage. 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. The Durbin Watson statistic tests the hypothesis that there is no autocorrelation. If there is no autocorrelation (where subsequent observations are related), the Durbin-Watson statistic should be between 1.5 and 2.5 and the p-value will be above 0.05.

Fit a regression model using the lm(dependent~Independent) command and give it a name (reg1).

```
reg1<-lm(Birthweight~Gestation)
```

To carry out the Durbin Watson Statistic for autocorrelation the library car must be loaded.

```
library(car)
```

If this command does not work, go to the Packages --> Install package(s) and select the UK (London)CRAN mirror. Then look for the package 'car' and click. For Rstudio, use Tools → Install packages. You may find that some earlier versions of Rstudio may not run the following command.

Request the Durbin Watson test

```
dwt(reg1)
```

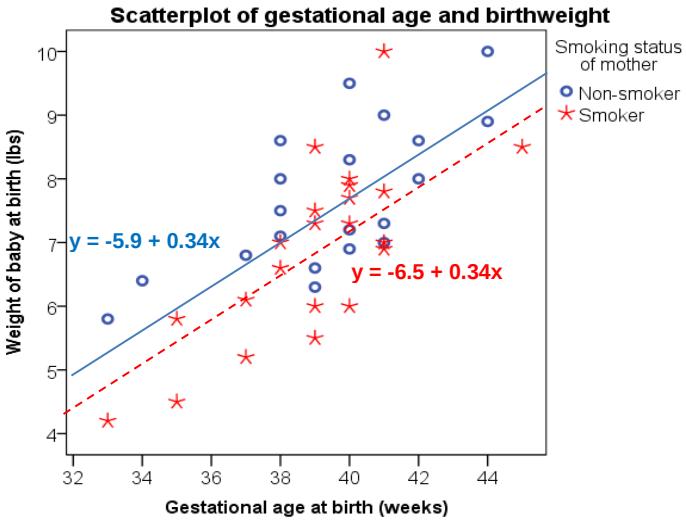
```
 lag Autocorrelation D-W Statistic p-value
   1       -0.202439      2.389853   0.262
 Alternative hypothesis: rho != 0
```

The Durbin Watson test statistic is 2.38 and the p-value is 0.262 so the hypothesis of no autocorrelation is not rejected and the observations can be classed as independent.

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 (smoker) 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

To run a regression model with only the main effects of gestation and smoker use the command `lm(Birthweight~Gestation+smoker)`.
Placing a '**:**' symbol between the two independents `lm(Birthweight~Gestation:smoker)` means that only the interaction is contained in the model.
Finally, using an '*****' between the two variables means that both the interaction and the main effects are included in the model.

To run the regression with both the interaction and the main effects of gestation and smoker
`Reg4<-lm(Birthweight~Gestation*smoker)`

The table contains the coefficients (Estimate) for the model (regression equation), their standard errors, the t-test values and p-values for each independent variable. Furthermore, the R-squared and the F test for the model is given on the second part. The output shows that only gestation is significant ($p < 0.001$) once the interaction term is added.

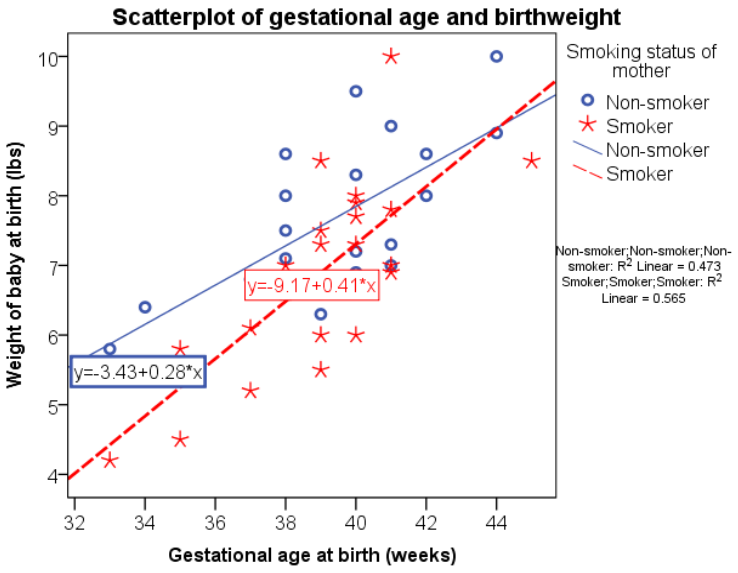
```
Call:
lm(formula = Birthweight ~ Gestation * smoker)

Residuals:
     Min       1Q   Median       3Q      Max
-1.39599 -0.80190  0.00992  0.54243  2.28036

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)            -3.43122    2.91950  -1.175 0.247197
Gestation               0.28191    0.07383   3.818 0.000482 ***
smokerSmoker           -5.73385    4.20587  -1.363 0.180812
Gestation:smokerSmoker  0.12992    0.10708   1.213 0.232530
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.901 on 38 degrees of freedom
Multiple R-squared:  0.5744,    Adjusted R-squared:  0.5408
F-statistic:  17.1 on 3 and 38 DF,  p-value: 3.451e-07
```

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.