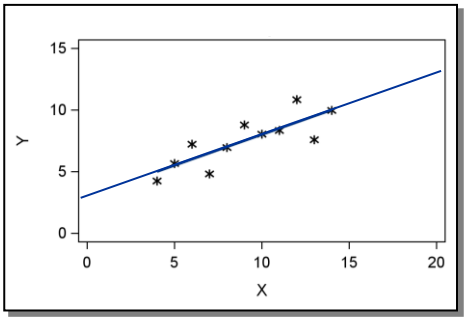
4. Regression Diagnostics: Influential Observations

As we have already explained, the diagnostics regarding the four assumptions of multiple linear regression models are essentially the same as the simple regression model. They start by inspecting data points through plots of response against predictors.

To illustrate the importance of plotting data, four examples were developed by Anscombe (1973). In each example, the scatter plot of the data values is different. However, the regression equation and the R^2 statistic are the same.

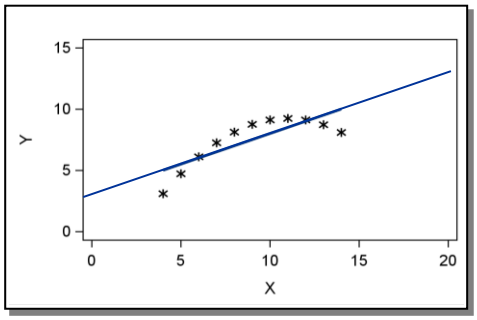
Scatter Plot of Correct Model



$\hat{y} = 3.0 + 0.5x$

$R^2 = 0.67$

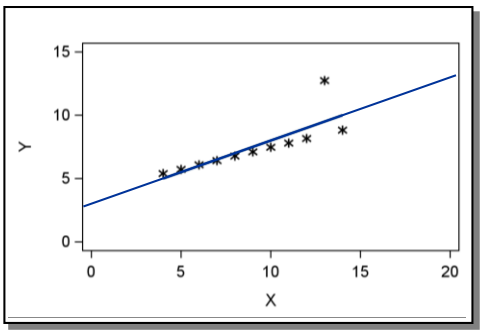
Scatter Plot of Curvilinear Model



$\hat{y} = 3.0 + 0.5x$

$R^2 = 0.67$

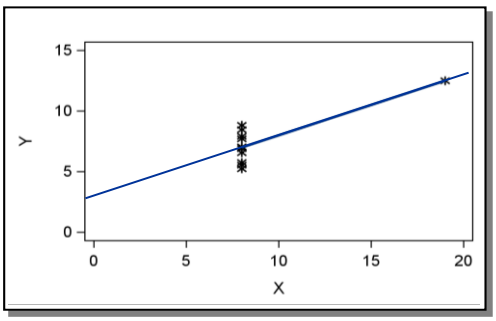
Scatter Plot of Outlier Model



$\hat{y} = 3.0 + 0.5x$

$R^2 = 0.67$

Scatter Plot of Influential Model



$\hat{y} = 3.0 + 0.5x$

$R^2 = 0.67$

The four plots illustrate that relying on the regression output to describe the relationship between your variables can be misleading. The regression equations and the R^2 statistics are the same even though the relationships between the two variables are different. Always produce a scatter plot before you conduct a regression analysis

Note: Outliers are data points that differ from the general trend of the data by more than is expected. An outlier might be an erroneous data point, or one that is atypical compared with the rest of the data. Influential observations are the ones that affect the model statistics (parameter estimates, standard errors of the parameter estimates, predicted values, studentised residuals, and so forth) when they are excluded from the analysis. Outliers might or might not be influential observations and vice versa.

Influential Observations

Identifying influential observations in multiple linear regression is more complex because you have more predictors to consider. The following measures and statistics are calculated to check such observations.

Standardised & Studentised Residuals

Recall that residuals (Unstandardized in SPSS) are the difference between observed values and predicted (fitted) values: $\boldsymbol{e = y - \hat{y}}$

Standardised residuals (ZRESID in SPSS) are the ordinary residuals divided by their standard errors: $\frac{e_i - \overline{e}}{\sqrt{MSE}}$.

The Studentised residuals (SRESID in SPSS)

$\boldsymbol{e = y - \hat{y} = y - Xb = y - X(X'X)^{-1}X'y = My}$, where $\boldsymbol{M = I - X(X'X)^{-1}X'}$ or

$\boldsymbol{e = y - X(X'X)^{-1}X'y = y - Hy}$, where $\boldsymbol{H = X(X'X)^{-1}X'}$ is the $n \times n$ hat (projection) matrix.

$\boldsymbol{H}$ is important for several reasons as it appears often in regression formulas. One important implication of $\boldsymbol{H}$ is that it is a projection matrix, meaning that it projects the response vector, $\boldsymbol{y}$, as a linear combination of the columns of the $\boldsymbol{X}$ matrix in order to obtain the vector of fitted values, $\boldsymbol{\hat{y} = Hy}$. Also, the diagonal of this matrix contains h_{ii} values, which is used to define studentised residuals: $\frac{e_i}{\sqrt{MSE(1-h_{ii})}}$.

Studentised deleted residuals. Another refinement that may be made in the study of the fitted model residuals is to measure the i^{th} residual when the fitted regression is based on all observations except the i^{th} observation, ie deleting the i^{th} observation: $\frac{e_i}{\sqrt{MSE_{(i)}(1-h_{ii})}}$.

In practice, if the absolute value of studentised deleted residuals is greater than 2, the observation might be influential.

Influence Statistics

These are based on the changes in regression values resulted from exclusion of a particular case. Standardised values are also calculated and used.

DfBeta (standardised): It measures the change in the parameter estimates when an observation is deleted from the analysis. $DfBeta_{j(i)} = \frac{b_j - b_{j(i)}}{SE(b_j)}$. The suggested cut-off is when the absolute value of the DfBeta is greater than $\frac{2}{\sqrt{n}}$.

DfFit (standardised): It measures the change in the predicted value when an observation is deleted. $DfFit_i = \frac{\hat{y}_i - \hat{y}_{(i)}}{SE(\hat{y}_i)}$. If the absolute value of the DfFit statistic is greater than $2\sqrt{\frac{k}{n}}$, the observation is considered influential.

Covariance Ratio: The ratio of the determinant of the covariance matrix with a particular case excluded from the calculation of the regression coefficients to the determinant of the covariance matrix with all cases included. $CovRatio_i = \frac{\left| s_i^2 (\boldsymbol{X}_i' \boldsymbol{X}_i)^{-1} \right|}{\left| s^2 (\boldsymbol{X}' \boldsymbol{X})^{-1} \right|}$. A ratio near one indicate that the i^{th} observation has little effect on the precision of the estimates. Suggested cut-off values for the covariance ratio are given by $CovRatio_i < 1 - 3\frac{k}{n}$ and $CovRatio_i > 1 + 3\frac{k}{n}$.

Distances

Mahalanobis: It measures how much a case's value of the independent variable differ from the mean of all cases: $D_i^2 = (\boldsymbol{X}_i - \bar{\boldsymbol{X}})' \boldsymbol{S}^{-1} (\boldsymbol{X}_i - \bar{\boldsymbol{X}})$, where $\boldsymbol{X}_i$, is the vector of the data for case *i* and $\bar{\boldsymbol{X}}$ is the vector of means (centroid) for the predictors. There is not a simple cut-off point for this measure. However, a table of critical values was provided by Barnett & Lewis (1978). As examples, for (n = 500, k = 6) values greater than 25; for (n = 100, k = 4) values above 15; and for (n = 30, k = 3) values greater than 11 can be problematic.

Cook's D: is a measure of the simultaneous change in the parameter estimates when an observation is deleted from the analysis. $D_i = \frac{\sum_{j=1}^{n} (\hat{y}_j - \hat{y}_{j(i)})^2}{k.MSE}$. An observation might have an adverse effect on the analysis if the Cook's D statistic is greater than $\frac{4}{n}$.

Leverage: The leverage value for an observation is the diagonal element of matrix $\boldsymbol{H}$ and measures how far the observation is from the centre of the $\boldsymbol{X}$-space. Values greater than $\frac{2k}{n}$ indicate observations that may be influential.

Example 4.1

In exercise physiology, an objective measure of aerobic fitness is how fast the body can absorb and use oxygen (oxygen consumption). Subjects participated in a predetermined exercise run of 1.5 miles. Measurements of oxygen consumption as well as several other continuous measurements such as age, pulse, and weight were recorded. The researchers are interested in determining whether any of these other variables can help predict oxygen consumption. Rawlings (1998). The predictors are:

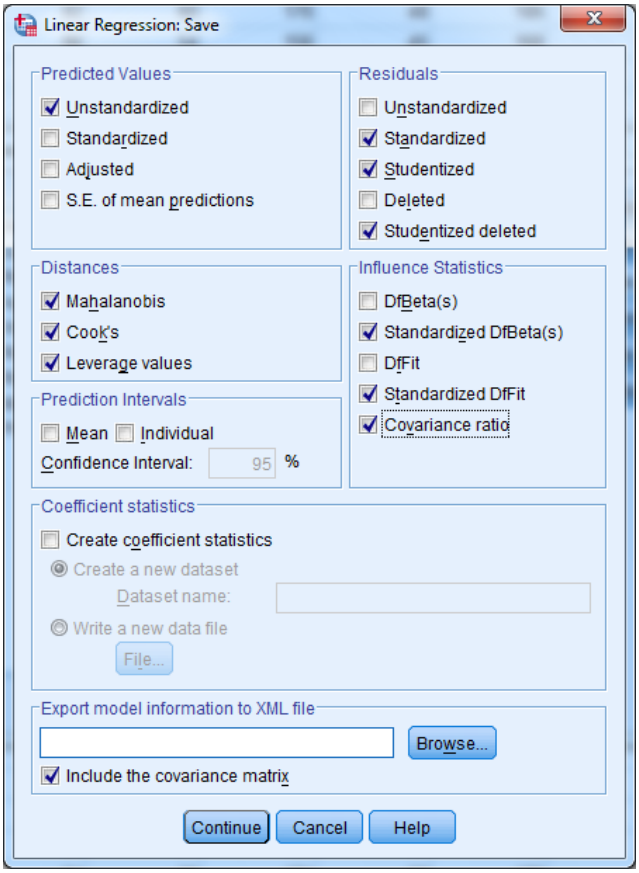
Runtime time to run 1.5 miles (in minutes)

Age age of the member (in years)

Run_Pulse pulse rate at the end of the run

Maximum_Pulse maximum pulse rate during the run

We can save the relevant measures by checking the values in SAVE dialogue box of SPSS:



Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.914[a]	.835	.810	2.321

a. Predictors: (Constant), Maximum_Pulse, RunTime, Age, Run_Pulse

b. Dependent Variable: Oxygen_Consumption

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	711.451	4	177.863	33.007	.000[b]
	Residual	140.104	26	5.389		
	Total	851.555	30			

a. Dependent Variable: Oxygen_Consumption

b. Predictors: (Constant), Maximum_Pulse, RunTime, Age, Run_Pulse

Coefficients[a]

	Unstandardized Coefficients		Standardized Coefficients		
Model	B	Std. Error	Beta	t	Sig.
1 (Constant)	97.170	11.657		8.336	.000
RunTime	-2.776	.342	-.723	-8.126	.000
Age	-.189	.094	-.187	-2.003	.056
Run_Pulse	-.346	.118	-.665	-2.924	.007
Maximum_Pulse	.272	.134	.468	2.023	.053

a. Dependent Variable: Oxygen_Consumption

Residuals Statistics[a]

	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	37.62	57.84	47.38	4.870	31
Std. Predicted Value	-2.002	2.148	.000	1.000	31
Standard Error of Predicted Value	.501	1.612	.898	.254	31
Adjusted Predicted Value	37.76	57.46	47.34	4.778	31
Residual	-4.984	4.724	.000	2.161	31
Std. Residual	-2.147	2.035	.000	.931	31
Stud. Residual	-2.199	2.132	.006	1.001	31
Deleted Residual	-5.228	5.185	.037	2.522	31
Stud. Deleted Residual	-2.390	2.302	.006	1.037	31
Mahal. Distance	.431	13.505	3.871	2.829	31
Cook's Distance	.000	.331	.035	.063	31
Centered Leverage Value	.014	.450	.129	.094	31

a. Dependent Variable: Oxygen_Consumption

The cut-off points for measures of influence are:

DfFit: $2\sqrt{\frac{k}{n}} = .8$

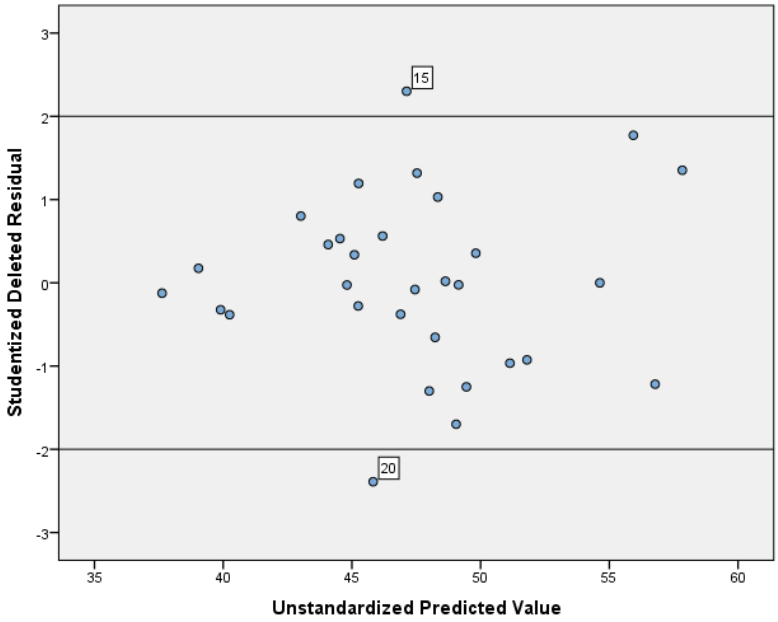
DfBetas: $\frac{2}{\sqrt{n}} = .36$

Mahalanobis: 13

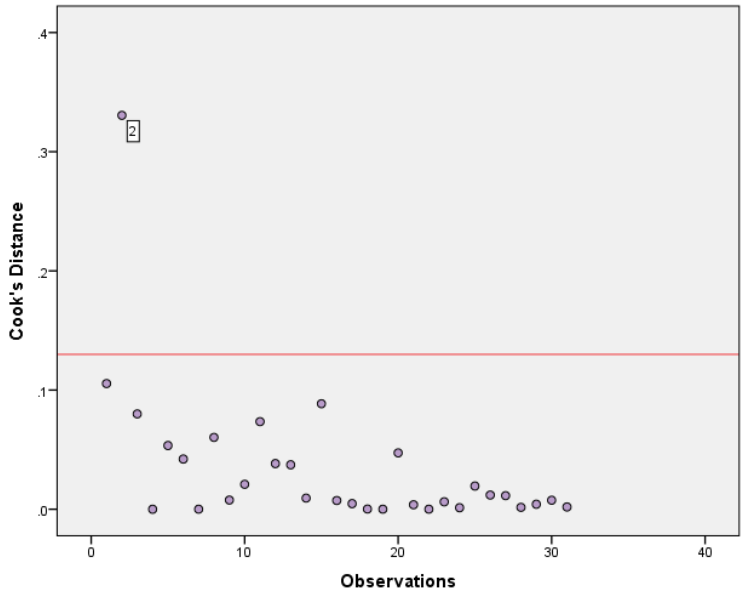
Cook's D: $\frac{4}{n} = .13$

Leverage: $\frac{2k}{n} = .32$

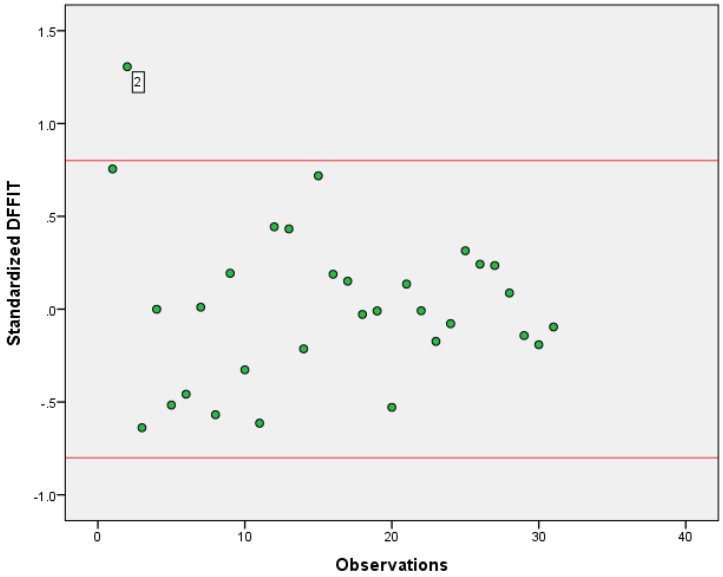
CovRatio: $1-3\frac{k}{n} = .52$, $1+3\frac{k}{n} = 1.48$



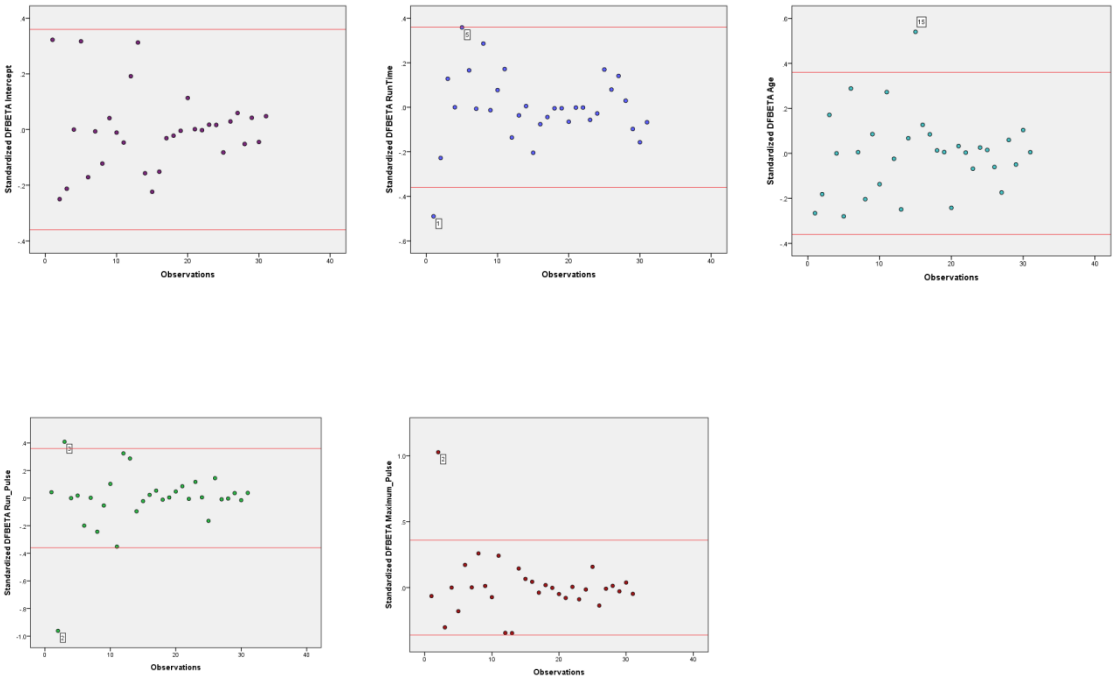
There are two observations beyond 2 standard errors from the mean of 0. Because we expect 5% of values to be beyond 2 standard errors from the mean (remember that these residuals are assumed to be normally distributed), the fact that we have 2 that far out gives no cause for concern (5% of 31 is 1.55 expected observations).



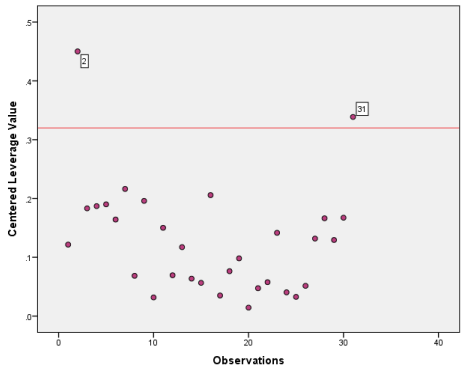
The Cook's D plot shows case number **2** to be an influential point.

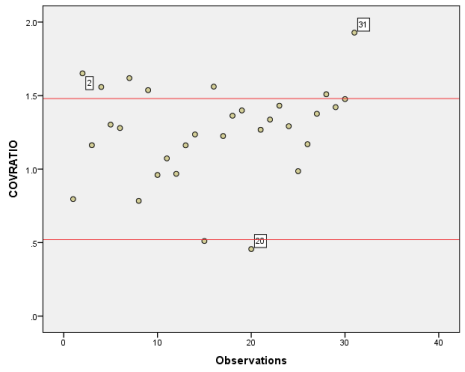


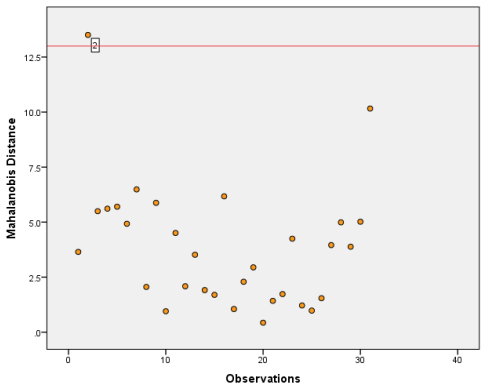
Case number **2** appears once again as an influential point. At this point it might be helpful to see which parameters this individual might be most influencing. DfBetas will provide that information.



Apparently, Individual **2** is influential because of the effects on the estimates of both **Run_Pulse** and **Maximum_Pulse**. Interestingly, this effect is opposite for **Maximum_Pulse** (where it is positive) compared with **Run_Pulse** (where it is negative).







The influential impact of individual **2** is further confirmed by the plots of distances above.

How to Treat Influential Observations

- Re-check the data to see if data entry errors occurred.
- One possible explanation is that the model is not adequate.
- In general, do not exclude data. In many circumstances, some of the unusual observations contain important information.
- If you do choose to exclude some observations, include a description of the types of observations you exclude and provide an explanation.
- Sensitivity analysis. This refers to performing the analysis for different scenarios and comparing the results. For example, you may perform the analysis for data with and without the influential observations and evaluate how the results are affected by inclusion and exclusion of the influential observations.

Practical Week 4

This week practical will be part of your assessment and carries 20% of the total mark.

Retrieve the data set **BodyFat.sav,** also used in week 3. Run a regression model, using **PctBodyFat** on **Abdomen**, **Weight**, **Wrist** and **Forearm**.

1. Comment on the regression output.
2. Check and comment on all relevant diagnostics, discussed in week 3.
3. Calculate all relevant measures, described in week 4 and use related plots to identify potential influential observations based on the suggested cutoff values.

Please return your work by Friday week 5, 27 October 2017.