This handout explains the output generated by PROC REG

Here is the SAS code:

```
data fat;
  input triceps thigh midarm bodyfat;
  DATALINES;
  19.5  43.1  29.1  11.9
  24.7  49.8  28.2  22.8
  30.7  51.9  37.0  18.7
  29.8  54.3  31.1  20.1
  19.1  42.2  30.9  12.9
  25.6  53.9  23.7  21.7
  31.4  58.5  27.6  27.1
  27.9  52.1  30.6  25.4
  22.1  49.9  23.2  21.3
  25.5  53.5  24.8  19.3
  31.1  56.6  30.0  25.4
  30.4  56.7  28.3  27.2
  18.7  46.5  23.0  11.7
  19.7  44.2  28.6  17.8
  14.6  42.7  21.3  12.8
  29.5  54.4  30.1  23.9
  27.7  55.3  25.7  22.6
  30.2  58.6  24.6  25.4
  22.7  48.2  27.1  14.8
  25.2  51.0  27.5  21.1
;
run;

 proc reg data=fat plots=diagnostics;
 model bodyfat = triceps midarm / influence r;
 run;
```

The REG Procedure
Model: MODEL1
Dependent Variable: bodyfat

Number of Observations Read	20
Number of Observations Used	20

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	389.45533	194.72767	31.25	<.0001
Error	17	105.93417	6.23142		
Corrected Total	19	495.38950			

Root MSE	2.49628	**R-Square**	0.7862
Dependent Mean	20.19500	**Adj R-Sq**	0.7610
Coeff Var	12.36089		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	6.79163	4.48829	1.51	0.1486
triceps	1	1.00058	0.12823	7.80	<.0001
midarm	1	-0.43144	0.17662	-2.44	0.0258

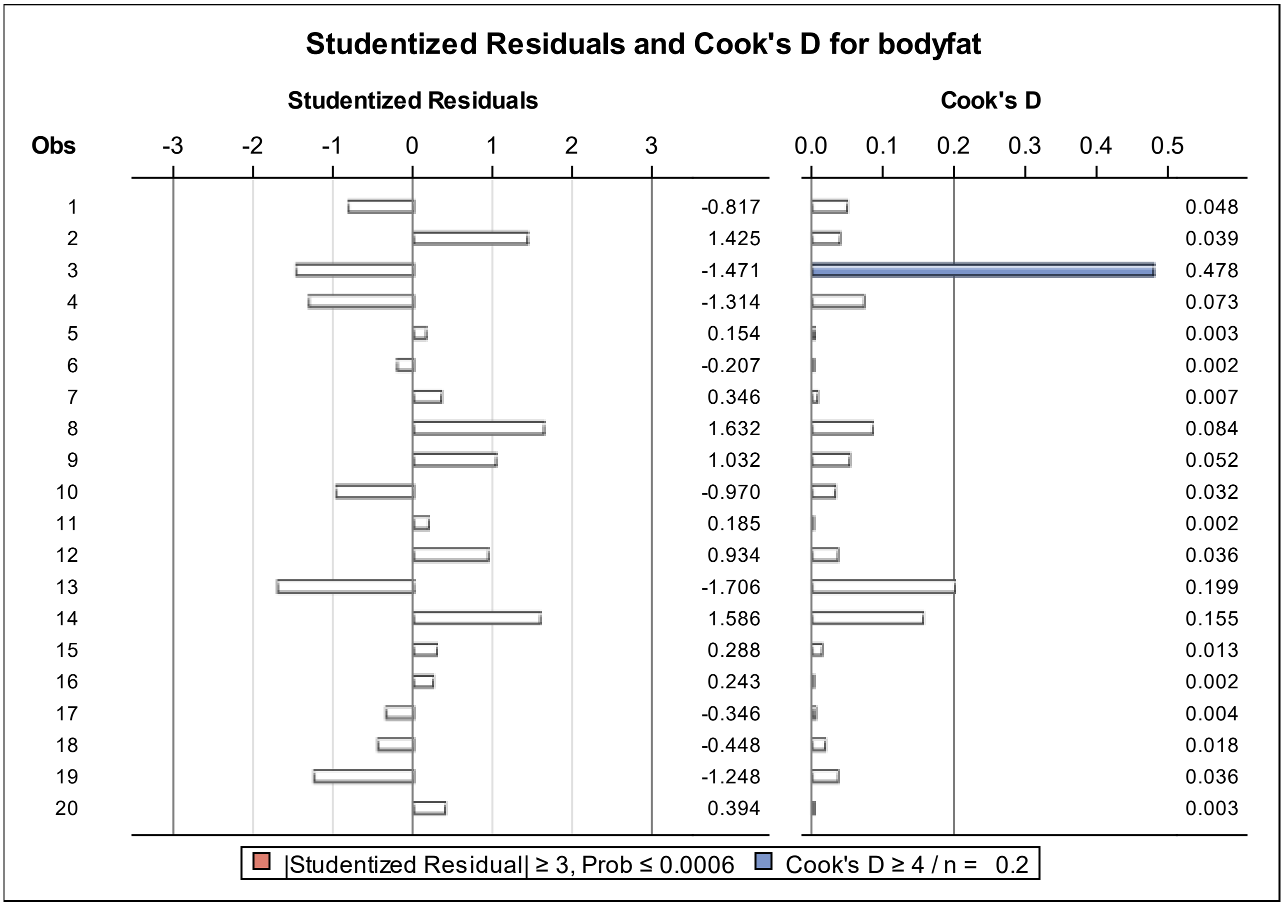
The REG Procedure
Model: MODEL1
Dependent Variable: bodyfat

Output Statistics											
Obs	**Dependent Variable**	**Predicted Value**	**Std Error Mean Predict**	**Residual**	**Std Error Residual**	**Student Residual**	**Cook's D**	**RStudent**	**Hat Diag H**	**Cov Ratio**	**DFFITS**
1	11.9	13.7481	1.0546	-1.8481	2.263	-0.817	0.048	-0.8084	0.1785	1.2948	-0.3768
2	22.8	19.3394	0.5791	3.4606	2.428	1.425	0.039	1.4734	0.0538	0.8654	0.3514
3	18.7	21.5462	1.5765	-2.8462	1.935	-1.471	0.478	-1.5271	0.3988	1.3266	-1.2439
4	20.1	23.1912	0.8350	-3.0912	2.352	-1.314	0.073	-1.3449	0.1119	0.9794	-0.4774
5	12.9	12.5712	1.3047	0.3288	2.128	0.154	0.003	0.1500	0.2732	1.6433	0.0919
6	21.7	22.1814	0.9035	-0.4814	2.327	-0.207	0.002	-0.2010	0.1310	1.3699	-0.0780
7	27.1	26.3022	0.9618	0.7978	2.304	0.346	0.007	0.3372	0.1484	1.3789	0.1408
8	25.4	21.5058	0.7341	3.8942	2.386	1.632	0.084	1.7243	0.0865	0.7874	0.5306
9	21.3	18.8951	0.8923	2.4049	2.331	1.032	0.052	1.0336	0.1278	1.1328	0.3956
10	19.3	21.6068	0.7561	-2.3068	2.379	-0.970	0.032	-0.9678	0.0917	1.1134	-0.3076
11	25.4	24.9666	0.8686	0.4334	2.340	0.185	0.002	0.1799	0.1211	1.3565	0.0668
12	27.2	24.9996	0.8252	2.2004	2.356	0.934	0.036	0.9303	0.1093	1.1498	0.3259
13	11.7	15.5794	1.0305	-3.8794	2.274	-1.706	0.199	-1.8183	0.1704	0.8230	-0.8242
14	17.8	14.1639	0.9859	3.6361	2.293	1.586	0.155	1.6663	0.1560	0.8793	0.7163
15	12.8	12.2105	1.4279	0.5895	2.048	0.288	0.013	0.2800	0.3272	1.7569	0.1953
16	23.9	23.3225	0.7597	0.5775	2.378	0.243	0.002	0.2360	0.0926	1.3082	0.0754
17	22.6	23.4198	0.7850	-0.8198	2.370	-0.346	0.004	-0.3368	0.0989	1.3032	-0.1116
18	25.4	26.3958	1.1387	-0.9958	2.221	-0.448	0.018	-0.4375	0.2081	1.4615	-0.2242
19	14.8	17.8128	0.6352	-3.0128	2.414	-1.248	0.036	-1.2703	0.0648	0.9613	-0.3343
20	21.1	20.1417	0.5585	0.9583	2.433	0.394	0.003	0.3839	0.0501	1.2284	0.0881

Output Statistics			
	DFBETAS		
Obs	**Intercept**	**triceps**	**midarm**
1	-0.0142	0.3087	-0.2152
2	0.0058	-0.0755	0.0837
3	1.0563	0.0525	-1.0572
4	0.3002	-0.1687	-0.2005
5	-0.0133	-0.0748	0.0665
6	-0.0505	-0.0306	0.0613
7	-0.0152	0.1146	-0.0529

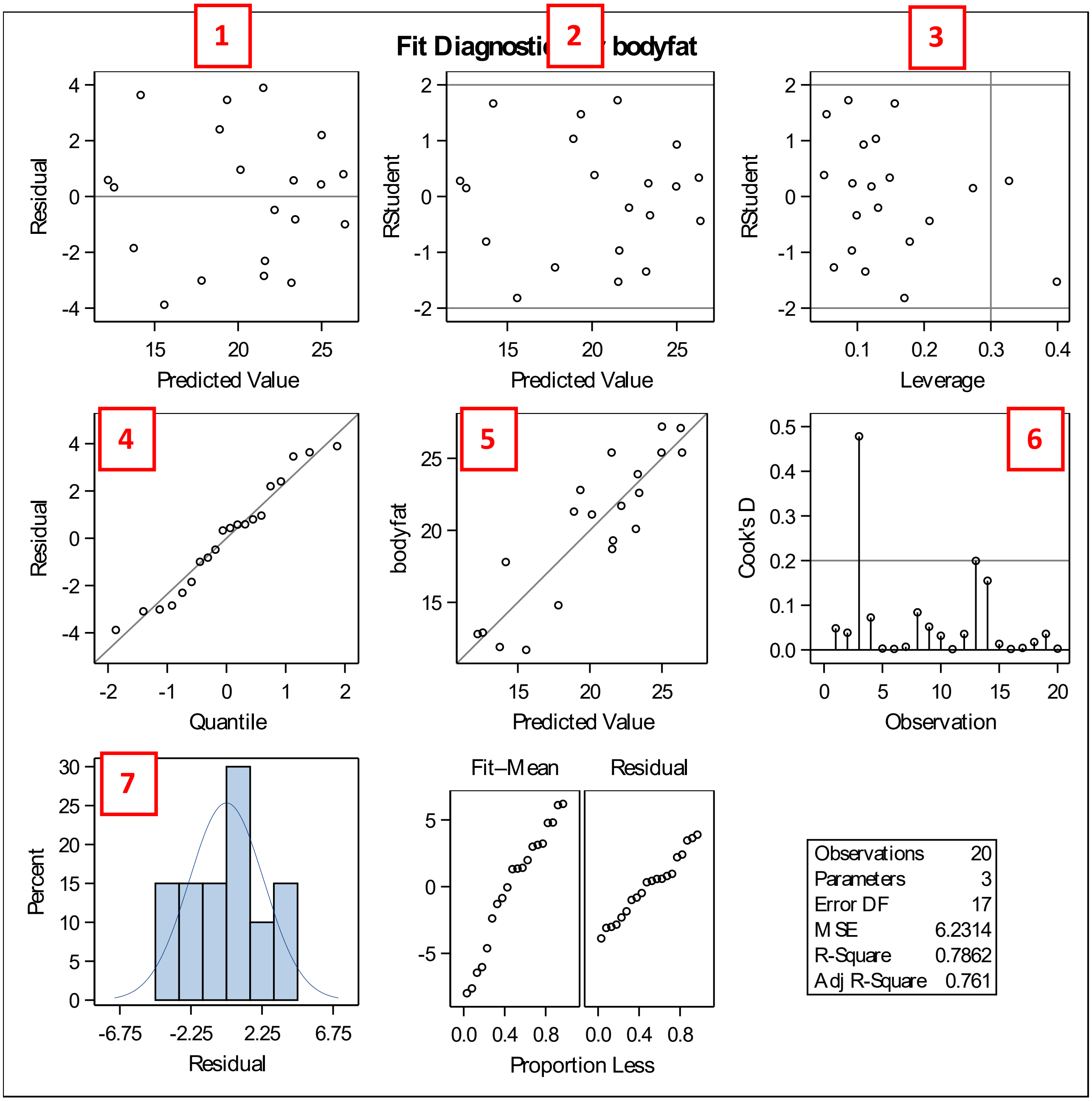
The REG Procedure
Model: MODEL1
Dependent Variable: bodyfat

Output Statistics			
	DFBETAS		
Obs	Intercept	triceps	midarm
8	-0.2916	0.0664	0.2703
9	0.3335	-0.0238	-0.2627
10	-0.1791	-0.1029	0.2073
11	-0.0320	0.0423	0.0062
12	-0.0666	0.2363	-0.0707
13	-0.7015	0.3786	0.3425
14	0.0731	-0.5798	0.3648
15	0.1672	-0.1179	-0.0667
16	-0.0380	0.0335	0.0190
17	-0.0365	-0.0657	0.0682
18	-0.0652	-0.1717	0.1616
19	-0.1127	0.1537	-0.0321
20	0.0140	-0.0006	-0.0024

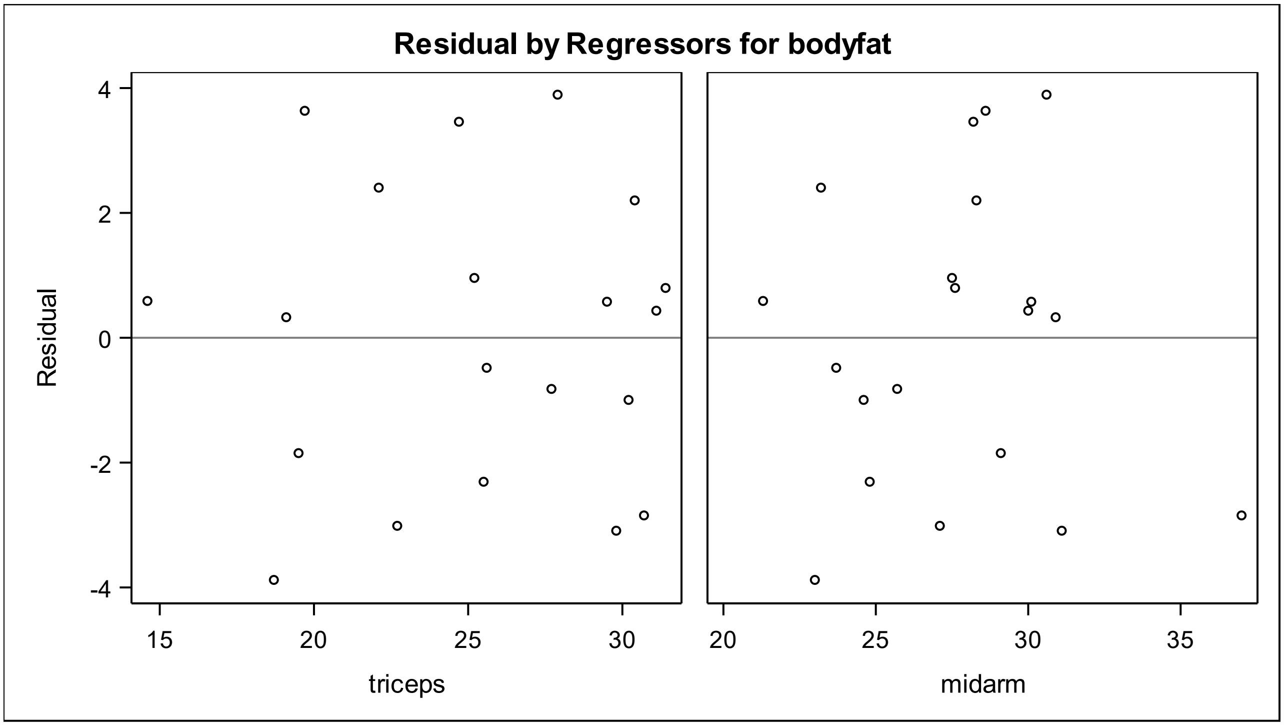


The REG Procedure
Model: MODEL1
Dependent Variable: bodyfat

Sum of Residuals	0
Sum of Squared Residuals	105.93417
Predicted Residual SS (PRESS)	148.33348



The REG Procedure
Model: MODEL1



Explanation of the Table on Page 2

This table is generated by the INFLUENCE and R options on the model statement. It looks like there are two tables (that go into page 3), but actually this is all one table. There are too many columns to fit on one page, so SAS prints it as two separate tables.

As usual, SAS gives us more information that we will use. For each observation (row), the columns we will examine are

Obs :	Observation (row) number in the data set
Dependent Variable :	Observed value for Y (i.e., the Y value from the original data)
Predicted Value :	Predicted value for Y (i.e., the fitted value)
Residual :	Observed – Predicted
Cook's D :	Cook’s D
RStudent :	Studentized residual
Hat Diag H :	Leverage
DFFITS :	Influence this observation has on the fitted (predicted) values
DFBETAS intercept :	Influence this observation has on the estimate for the intercept
DFBETAS triceps :	Influence this observation has on the estimate for the slope on triceps
DFBETAS midarm :	Influence this observation has on the estimate for the slope on midarm

Explanation of the Graphs on Page 3

These two graphs give a visual representation of two columns in the table – studentized residuals and Cook’s D. These are given side-by-side so we can see the combined effect of outliers (on the left) and influence (on the right) for each observation in the data set. These graphs provide exactly the same information that is given in the table, but it is easier to comprehend the information in graphical form.

Explanation of the Plots on Page 4

Consider the 8 plots on the fourth page of the SAS output.

- Plot 1 is the residual plot. The x-axis shows the fitted (i.e., predicted) values from the model and the y-axis shows the residuals (i.e., observed – predicted values). We have already used this plot to assess model assumptions. We want the points on this plot to have no obvious pattern. For example, a wedge-shaped pattern would indicate that the assumption of equal variances has been violated.
- Plot 2 is another residual plot. The x-axis is the same as in Plot 1, but the y-axis shows the studentized residuals (instead of the regular residuals). The advantage of using Plot 2 instead of Plot 1 is that we can use Plot 2 to easily identify potential outliers. Points that have studentized residuals above +2 or below –2 are considered potential outliers. The values 2 and –2 are marked with horizontal lines on Plot 2, so it is easy to see if any points fall above the top line or below the bottom line.
- Plot 3 shows leverage on the x-axis and studentized residuals on the y-axis. Since the y-axis is the same as in Plot 2, we can use the two horizontal lines (at 2 and –2) to identify potential outliers. The horizontal axis contains the leverage. Recall that leverage is calculated using only the X values for each observation – the Y values are not involved in leverage. The vertical line at 0.3 marks the boundary for leverage. Points to the right of this line have high leverage. Points with high leverage have an unusual combination of X values, but this may or may not be a mistake in the data.
- Plot 4 is the QQ plot, which we have been calling the normal probability plot. We have used this plot to assess the assumption of normal errors. We want the points to follow the line. Extreme deviations from the line indicate that the errors are not normal.
- Plot 5 shows how well the model fits the data. The x-axis is the value for Y that is predicted by the model and the y-axis is the observed value for Y. If the model was “perfect”, then every point would fall directly on the line. Since no model is perfect, we expect the points to vary around the line. The amount of variation is related to the values for R^2 and RMSE. If the points closely follow the diagonal line, then the value for R^2 should be high and RMSE should be low, indicating the model fits the data very well.
- Plot 6 shows the values for Cook's D, which measures the overall influence each observation has on the final fitted model. Observations with large influence will be shown as a point above the horizontal line. These points should be investigated to make sure they are not mistakes in the data. The observation number (row in the data set) is on the x-axis.
- Plot 7 is a histogram of the residuals. This is related to Plot 4 in that they both represent how well the residuals follow a normal distribution. In the QQ plot (Plot 4), we want the points to follow the line. In the histogram (Plot 7), we want the histogram of the calculated residuals (the bars) to follow the theoretical normal distribution (the curve).
- We will not discuss the last, un-numbered, graph in this section.

It is worth noting that NONE of the graphs on page 4 are scatter plots. In a scatter plot, the x-axis has the values of the predictor variable and the y-axis contains the values of the response variable. A multiple linear regression model has more than one predictor variable, so it is impossible to visualize the data in a single scatter plot. To accomplish this, we would need to generate one scatter plot for each predictor variable. In SAS, this can be accomplished by using PROC SGPLOT or by using the "plots=matrix" option on PROC CORR.

Explanation of the Plots on Page 5

The two graphs on page 5 of the SAS output contain the partial residual plots. There will be one partial residual plot for each predictor variable. For the current body fat example, the predictors are triceps and midarm, so there are two partial residual plots. These plots are interpreted the same way as the ordinary residual plot (Plot 1 on page 3). We want the points to appear scattered, with no obvious pattern. The usefulness of the partial residual plots occurs when the points DO follow some pattern, because this could indicate what changes need to be made to the model to improve the goodness of fit. In the current example, both of these plots seem to have a random scatter of points, so no adjustment to the model is warranted. To illustrate how these graphs can be used, suppose that the partial residual plot for triceps (the first plot on page 4) showed a distinct quadratic shape. Then this would indicate that we need to include a "triceps squared" term in the model.

How to Use This Information

The graphs can be used to gain insight into measures of influence, outliers and potential violations of model assumptions. The information we obtain from any graph is merely for insight -- we generally do not make formal statistical decisions based solely on a graph. Formal decisions should be made based on numeric measurements (like Cook's D) or an official hypothesis test. Sometimes, these formal methods do not exist so we rely on the graphs to identify potential difficulties (or the lack thereof). Even when the formal methods do exist, the graphs provide quick method to identify potential problems with the data and/or the model.

⇒ Outliers

From Plot 2, there are no points above 2 or below –2, so there are no outliers in this particular dataset. If there had been any potential outliers, we would go to the big table on page 2 and find the specific observations. These will have RStudent values above 2 or below –2.

⇒ Leverage

From Plot 3, there are two observations with high leverage. These correspond to the two points to the right of the vertical line at 0.3. Use the big table on page 2 to identify the specific observations. They will have "Hat Diag H" values larger than 0.3. For this dataset, the observations with high leverage are observations 3 and 15. These two observations should be checked to make sure they do not contain errors. Any mistakes should be corrected, but otherwise these observations should not be changed or removed from the dataset.

⇒ **Influence**

Each observation (row) in the dataset can influence the estimation of several quantities associated with the model. It can influence the fitted values (as measured by DFFFITS), or it can influence the estimates for intercept or the slopes (DFBETAS). We will rely on Cook's D, which measures the overall influence of each observation. Plot 6, which shows the plotted values for Cook's D, has one point above the horizontal line. This indicates one influential observation. The row number for this observation is found on the x-axis of Plot 6 – it is observation 3. In the SAS code, the third line of data shows relatively high values for triceps and midarm (30.7 and 37.0, respectively), but a relatively low value for body fat (18.7). It is this combination of values that makes observation #3 influential. If we had collected this data, we want to investigate this observation to determine if any of these values are incorrect. For the analyses we encounter in this class, the data will be given to us and we will assume the data is correct (so we cannot change it).

Conclusion

There is one observation (#15) that has high leverage and one observation (#3) that has both high influence and high leverage, but there are no outliers in the data. The diagnostic plots indicate that no assumptions have been violated, so this appears to be a reasonable model.