

## **Assignment 6**

**1. (6 points) Potential problems.** On pages 92 – 103, the authors discuss several problems we may encounter. In your own words, give a brief description of the problem and, as best you can, explain what we can do to identify/correct the problem.

**a. Non-linearity of the data.**

The non-linearity problem occurs when the connection between variables in our data doesn't follow a straight line, but our model assumes a linear relationship. To identify this issue, we can check a residual plot, which reveals patterns in the errors. To fix it, we might use interaction terms or power functions for the predictors. This means tweaking our model to better capture the non-linear patterns in the data, ensuring our predictions align more accurately with the real-world relationships among the variables.

**b. Correlation of error terms.**

One critical assumption in our model is that errors are not correlated; they should be independent. However, if errors show correlation, our model might wrongly appear more precise than it actually is. To detect this, we can create a simple scatter plot of residuals against predicted values. If this plot displays a noticeable pattern or trend, it indicates a correlation problem among errors. Why does it matter? If errors are correlated, our estimated standard errors might underestimate the true standard errors. This results in confidence intervals that are narrower than they should be, and p-values that are lower than they should be. Essentially, we may have an unjustified sense of confidence in our model.

**c. Non-constant variation of error terms.**

Non-constant errors, or heteroscedasticity, manifest when the variances of the error terms in a model are not consistent and possibly increase over time, often depicted by a funnel shape in the residual plot. When this pattern is observed, it suggests that the variability of errors is changing across different levels of the independent variable(s). To rectify this issue, a common approach is to employ a transformation on the response variable, Y. One effective strategy is to use a concave function, such as taking the logarithm ( $\log Y$ ) or square root ( $\sqrt{Y}$ ) of the response variable. These transformations help stabilize the variance across the range of predictors, mitigating the funnel-shaped residuals.

**d. Outliers.**

An outlier is a data point that lies considerably far from the least squares regression line on a graph. While outliers can impact the model, in some cases, removing them may not significantly alter the slope of the line. Residual plots are often effective in identifying these outliers. However, caution is advised when considering their removal, as outliers might carry valuable insights or indicate unique patterns in the data.

### e. High leverage points.

High leverage points are observations that have a high influence on the fit of a regression model. These points can strongly influence the slope (regression coefficient) of the model, especially if they have extreme values in the independent variable. (How to identify the problem)

In order to quantify an observation predictor to understand if it is a high leverage point we can compute the leverage statistic.

### f. Collinearity

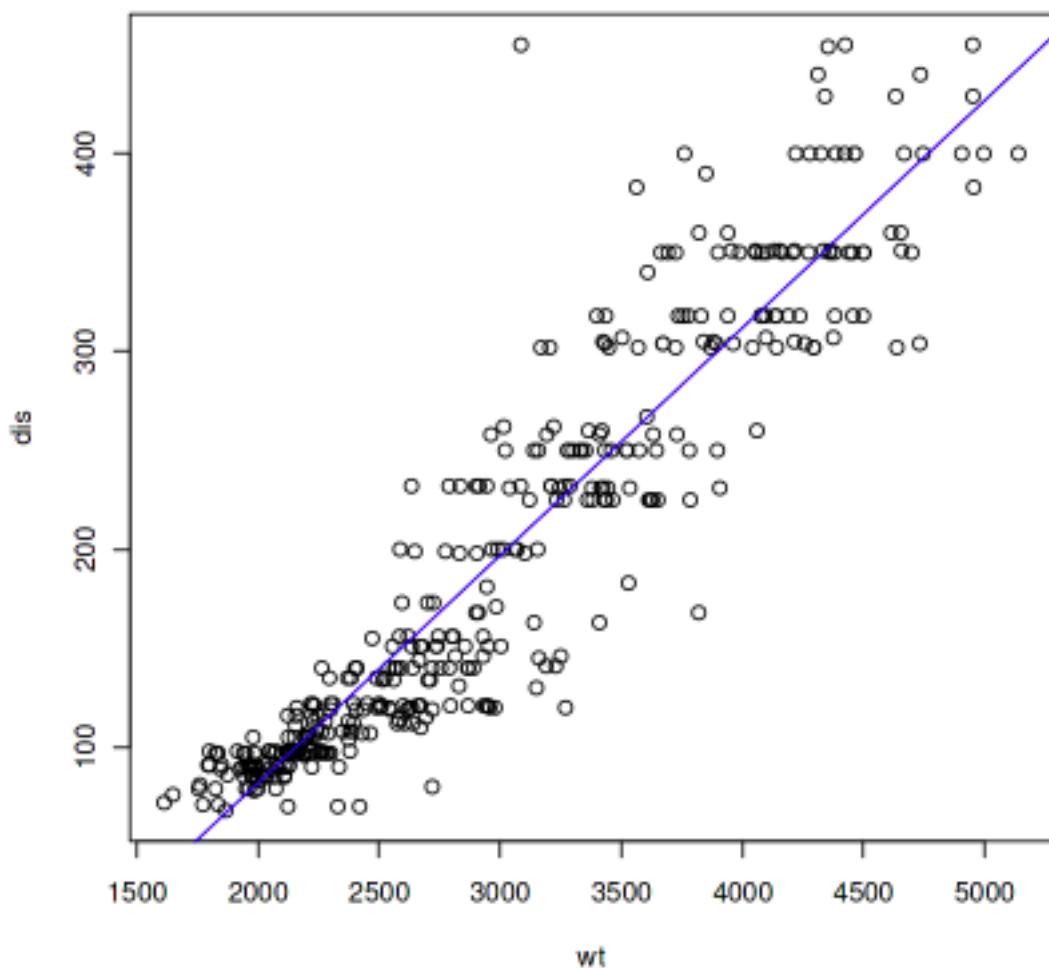
Whenever two predictor variables are strongly related to each other, they are highly correlated. This could cause problems in understanding the trends relation between each variable and the response. Collinearity most of the time will cause the p-value for each predictor to increase causing a reduction in the power of the hypothesis of the test. In order to detect collinearity we need to look at the correlation matrix of the predictors. If an element of the matrix is large in value, it will indicate a pair of highly correlated variables, showing high collinearity. If collinearity exists between 3 or more variables the VIF is used to detect it.

**2. (6 points) Examining residual plots. In Assignments 2 and 3, we studied the Auto data set. We found a linear model for displacement as a function of weight.**

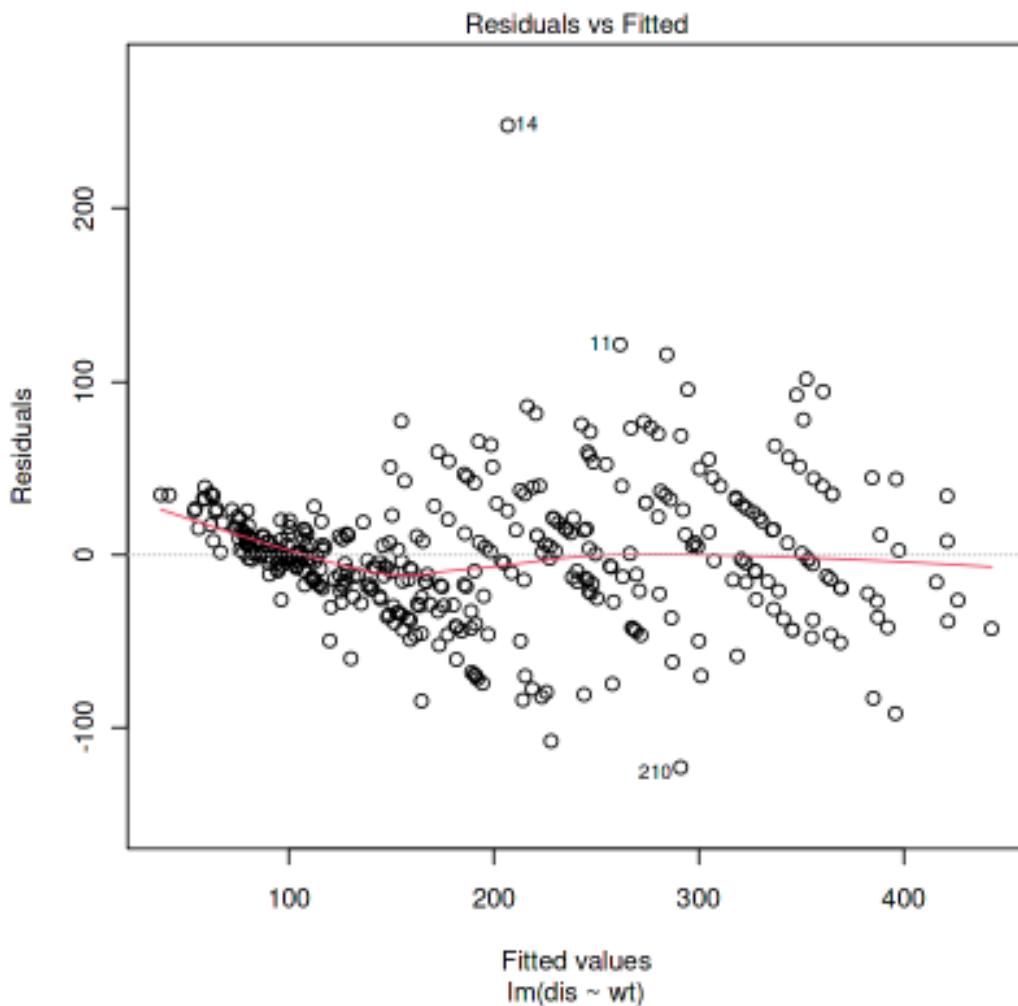
**a. View the residual plot for displacement as a function of weight. I called my model disAFwt (displacement as a function of weight) and used the command plot(disAFwt,1). The 1 after the comma tells R to plot only the residuals versus fitted values. Otherwise, the command plot(disAFwt) generates several different graphs (p. 113).**

#### **CODE:**

```
Auto=read.csv("Auto.csv",header=T,na.strings=?,stringsAsFactors=TRUE)
Auto=na.omit(Auto)
wt=Auto$weight
dis=Auto$displacement
disAFwt =lm(dis~wt)
plot(wt,dis)
abline(disAFwt,col="blue")
```



```
summary(disAFwt)
plot(disAFwt,1)
```



**b. Note that the horizontal axis has the fitted (output) values. The vertical axis shows the corresponding residuals. The red curve in the middle represents a local average of the residuals. Notice that for fitted values less than 100, the red curve goes positive. Now answer the following question: For cars that weigh less than 2000 pounds, does the model under-estimate or over-estimate the actual values? State your answer and explain how you can determine it from the residual plot.**

Based on the linear regression plot, it appears that for cars weighing 2000 pounds, the predicted displacement reaches a value of 100. However, when examining the residual plot for fitted values less than 100, it becomes evident that the red curve ascends, indicating a consistent trend of overestimation by the model. This suggests that the model tends to predict higher displacement values than the actual observed values for cars with weights below 2000 pounds, pointing to a potential limitation or bias in the model's predictive accuracy in this weight range.

**c. Then complete the more general statement: When the red curve is positive, for the corresponding predictor values, the model (over/under) estimates the actual values. When the red curve is negative, for the corresponding predictor values, the model (over/under)-estimates the actual values.**

When the red curve is positive, for the corresponding predictor values, the model **overestimates** the actual values. When the red curve is negative, for the corresponding predictor values, the model **underestimates** the actual values.

**3. (4 points) Outliers.** Let's continue with the residual plot from Question 2 (displacement as a function of weight). In the plot, we see one outlier whose residual is more than 200. The plot also tells us that the outlier is the 14th data point (the Buick Estate Wagon). Let's remove this one point and find a linear model for this updated data set. Use the command `Auto1=Auto[-14,]`. This creates a new data set with the Buick wagon removed. Find a new model (I called mine `disAFwt1`) and view the output of the summary command. Then complete the table below:

**Code:**

```
Auto=read.csv("Auto.csv",header=T,na.strings=?,stringsAsFactors=TRUE)
Auto=na.omit(Auto)
Auto1=Auto[-14,]
dis1=Auto1$displacement
wg1=Auto1$weight
Model1=lm(dis1~wg1)
summary(Model1)
```

Data Set	RSE	SE(beta 0)	SE(beta 1)	Adjusted R^2
Auto	37.71	6.951	0.002245	0.8701
Auto1	35.59	6.561	0.002119	0.8827

**Based on what you see in the table, do you agree with the book (p. 97) that outliers can have a significant effect on confidence intervals and measures of fit? Explain.**

After looking at the data table, I agree with the book's conclusion that outliers really affect confidence intervals and how well our model fits the data. When we took out data point 14, we saw some big improvements. The errors in our model got smaller, which means our model fits better. Also, the adjusted R-squared value got bigger, showing that our model can predict things more accurately. This shows that outliers are important and can change our results a lot.

**4. (4 points) Other diagnostic plots can also help us spot points of concern (outliers, high leverage points, etc.). These plots usually label points that could cause trouble. Using the Auto data set, find a linear model for mpg as a function of weight, year, and origin (we did this in class). Call your model `mpgMod`. Now view the following plots:**

**CODE:**

```
Auto=read.csv("Auto.csv",header=T,na.strings=?,stringsAsFactors=TRUE)
Auto=na.omit(Auto)
mpg=Auto$mpg
```

```
wt=Auto$weight
yr=Auto$year
origin=Auto$origin
mpgMod=lm(mpg ~ wt + yr + origin)
summary(mpgMod)

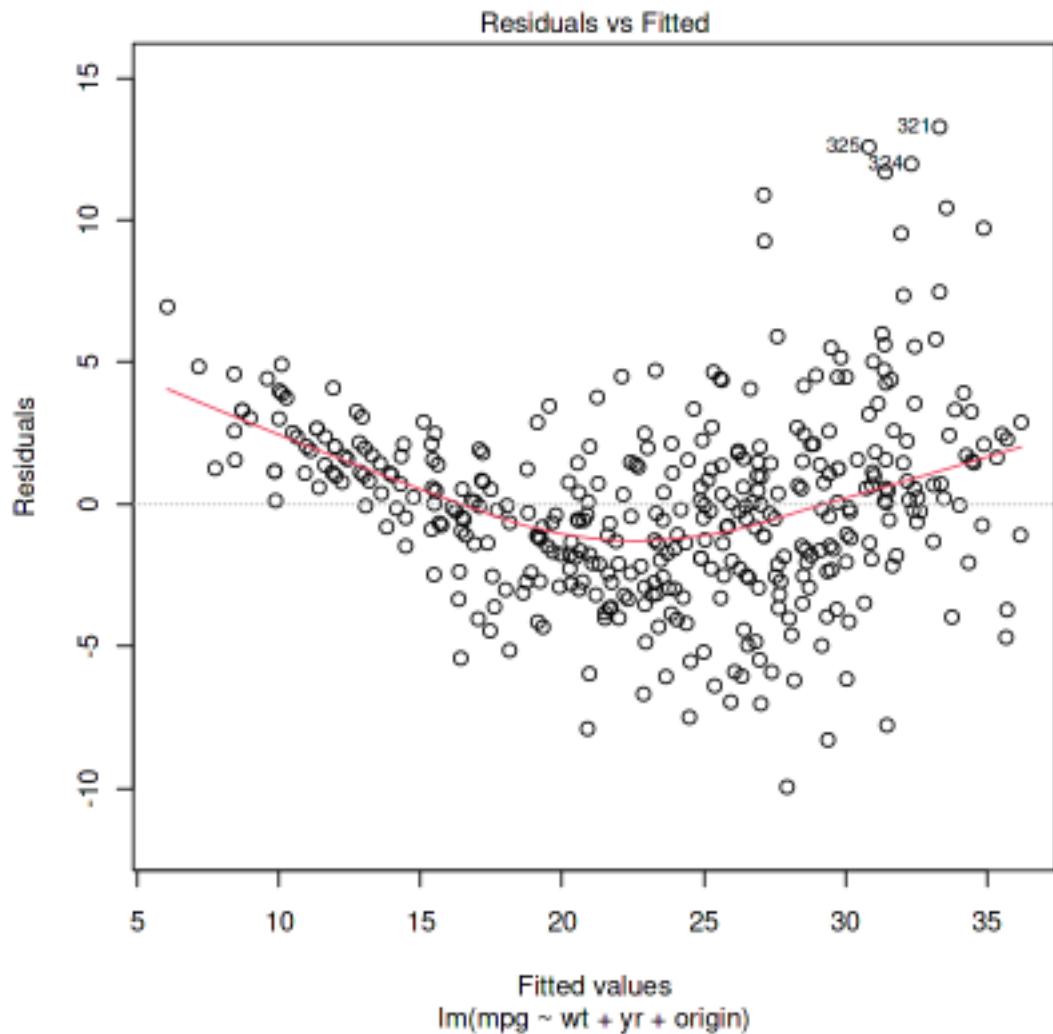
Call:
lm(formula = mpg ~ wt + yr + origin)

Residuals:
    Min      1Q  Median      3Q     Max 
-9.9440 -2.0948 -0.0389  1.7255 13.2722 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -1.805e+01  4.001e+00 -4.510 8.60e-06 ***
wt          -5.994e-03  2.541e-04 -23.588 < 2e-16 ***
yr           7.571e-01  4.832e-02  15.668 < 2e-16 ***
origin       1.150e+00  2.591e-01   4.439 1.18e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

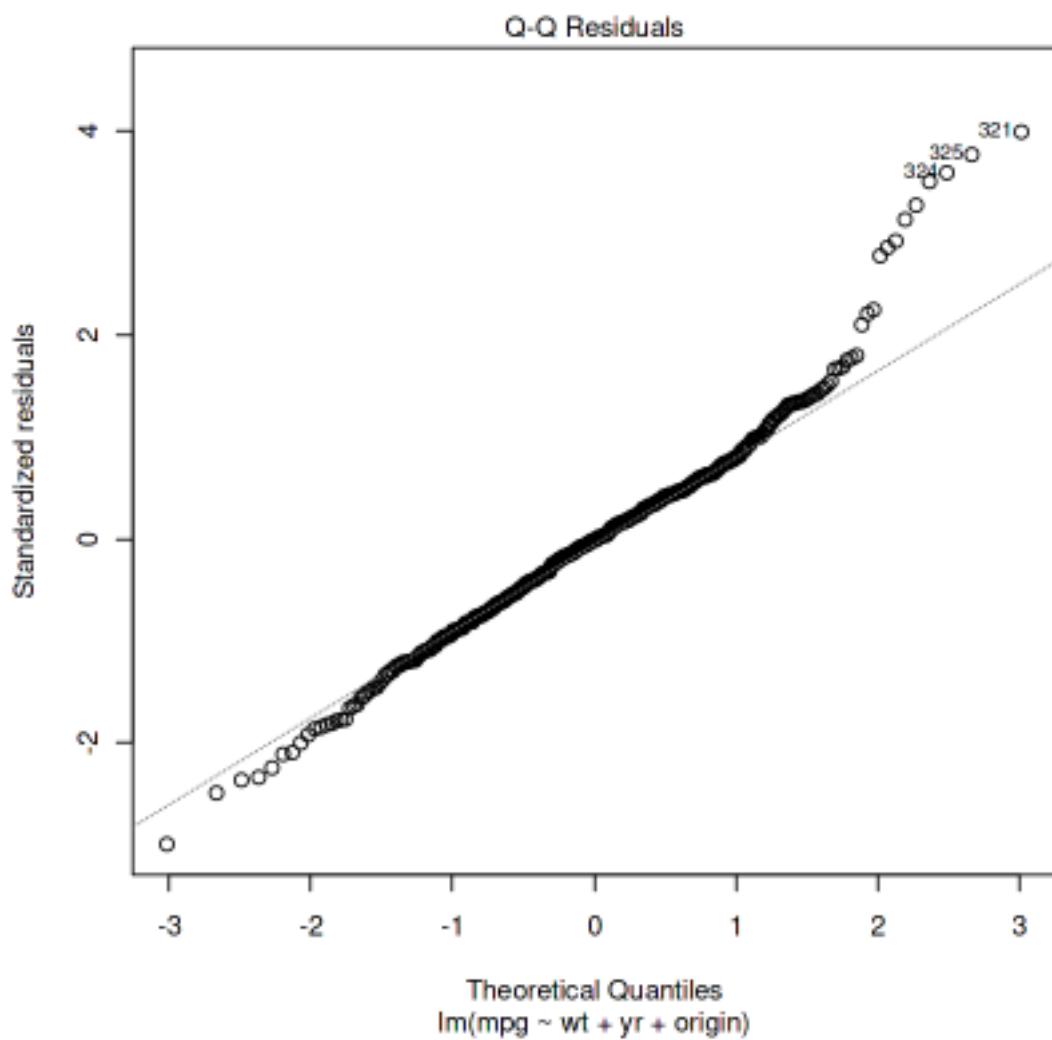
Residual standard error: 3.348 on 388 degrees of freedom
Multiple R-squared:  0.8175,    Adjusted R-squared:  0.816 
F-statistic: 579.2 on 3 and 388 DF,  p-value: < 2.2e-16
```

**plot(mgpMod,1): This is the residual plot. You can see which points have the largest residuals (positive and negative).**

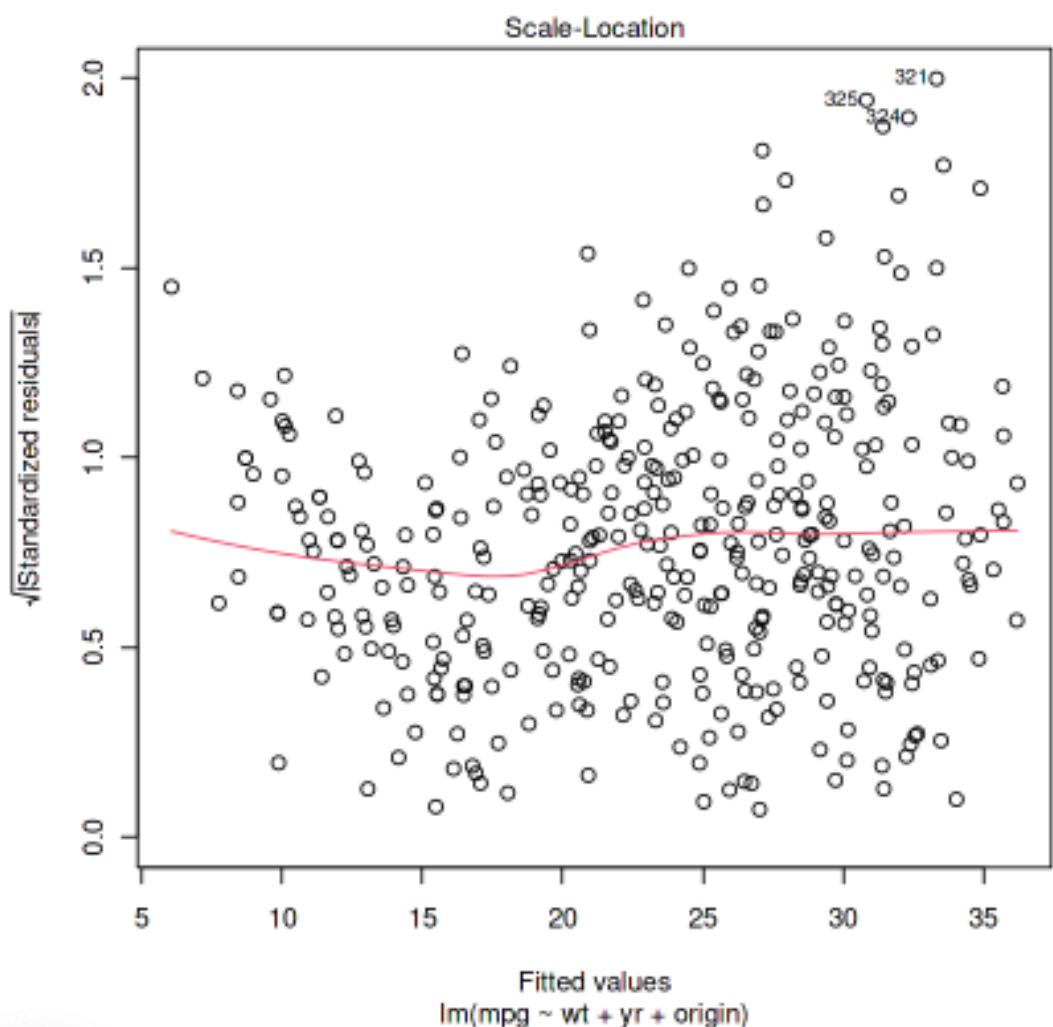


The points with the largest residuals are 321/324/325, all of them are positive

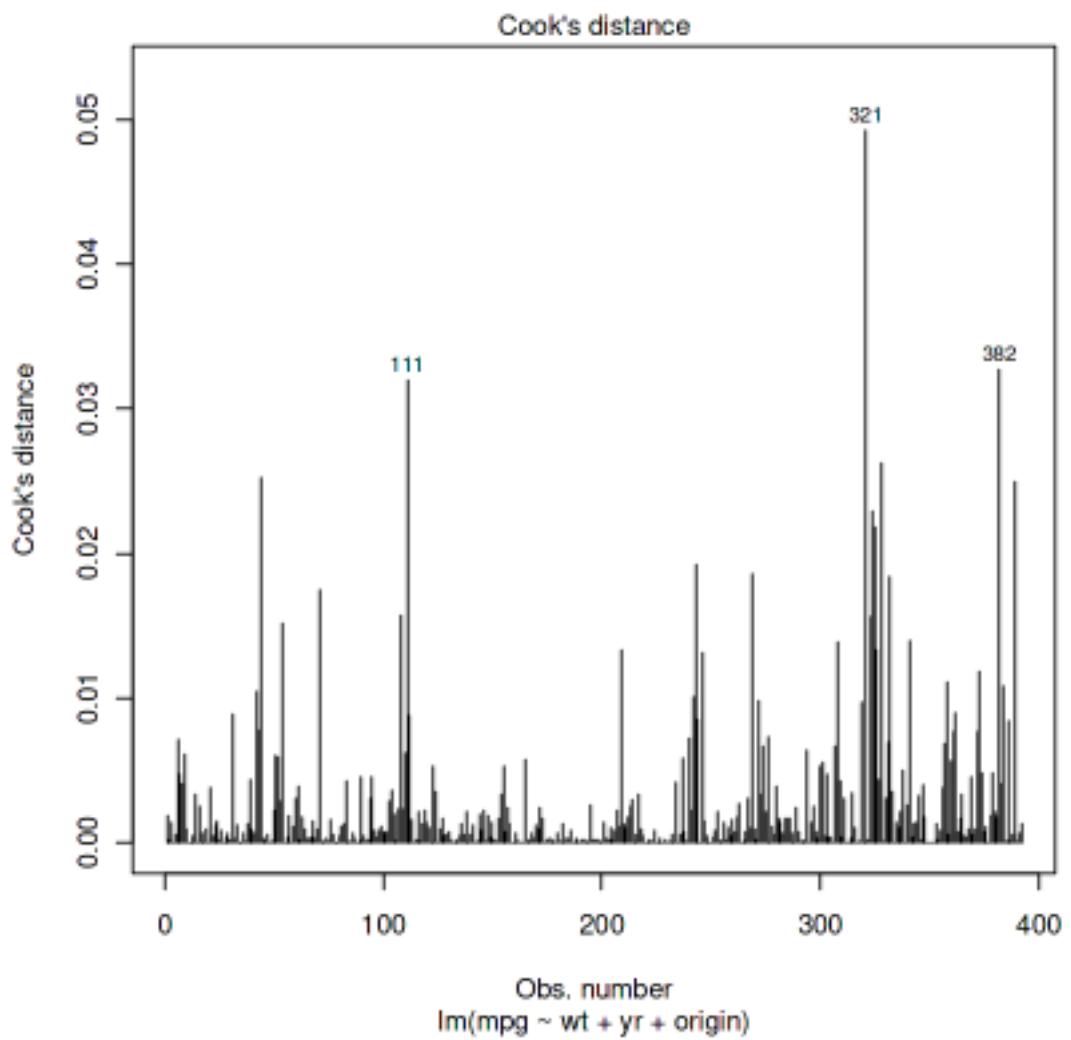
**plot(mpgMod,2): This gives what is called the Q-Q Plot. This helps us compare the distributions of the variables. When most of the points fall in a line, we can conclude that the variables are normally distributed (often what happens). Points that lie off the line could be outliers, etc.**



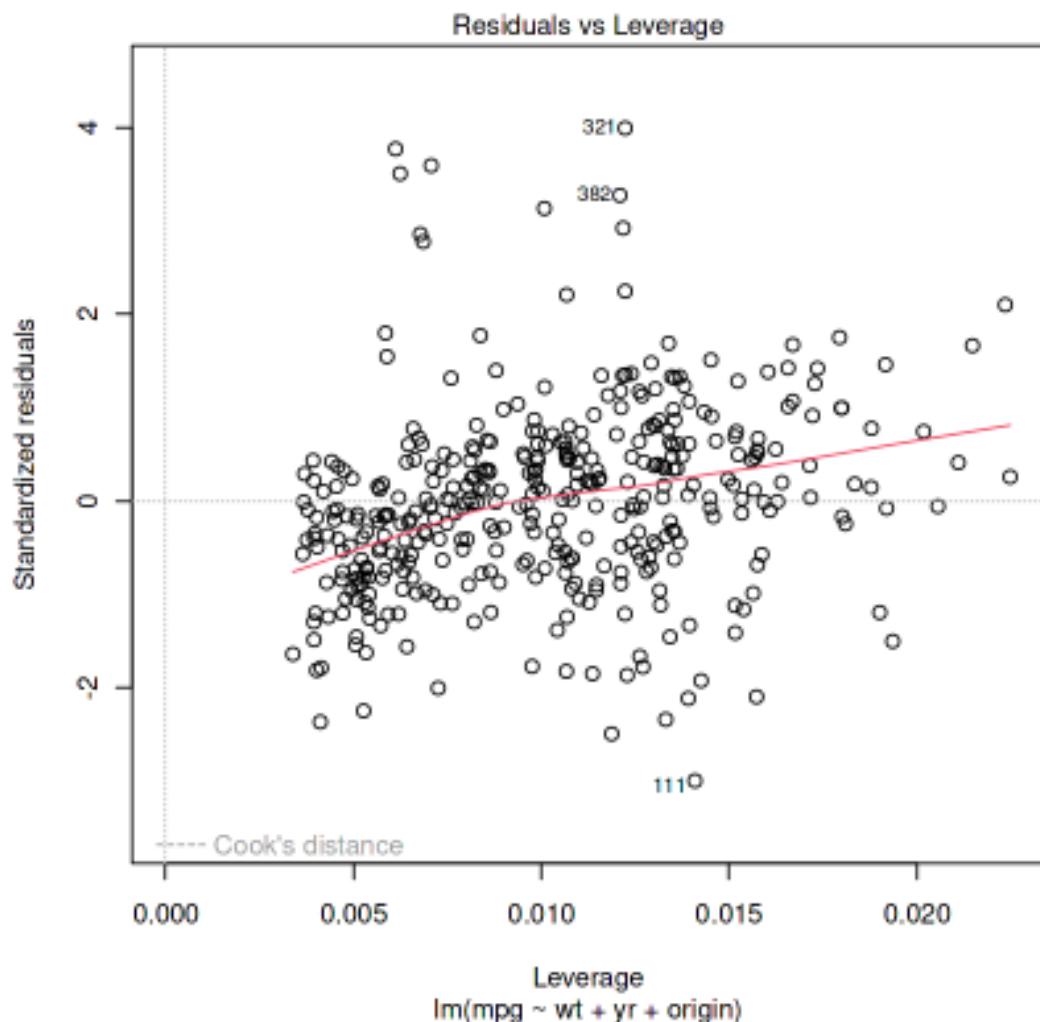
**plot(mpgMod,3): This gives a plot of standardized residuals. Note that the heights of the points are positive. The highest points are the ones to scrutinize.**



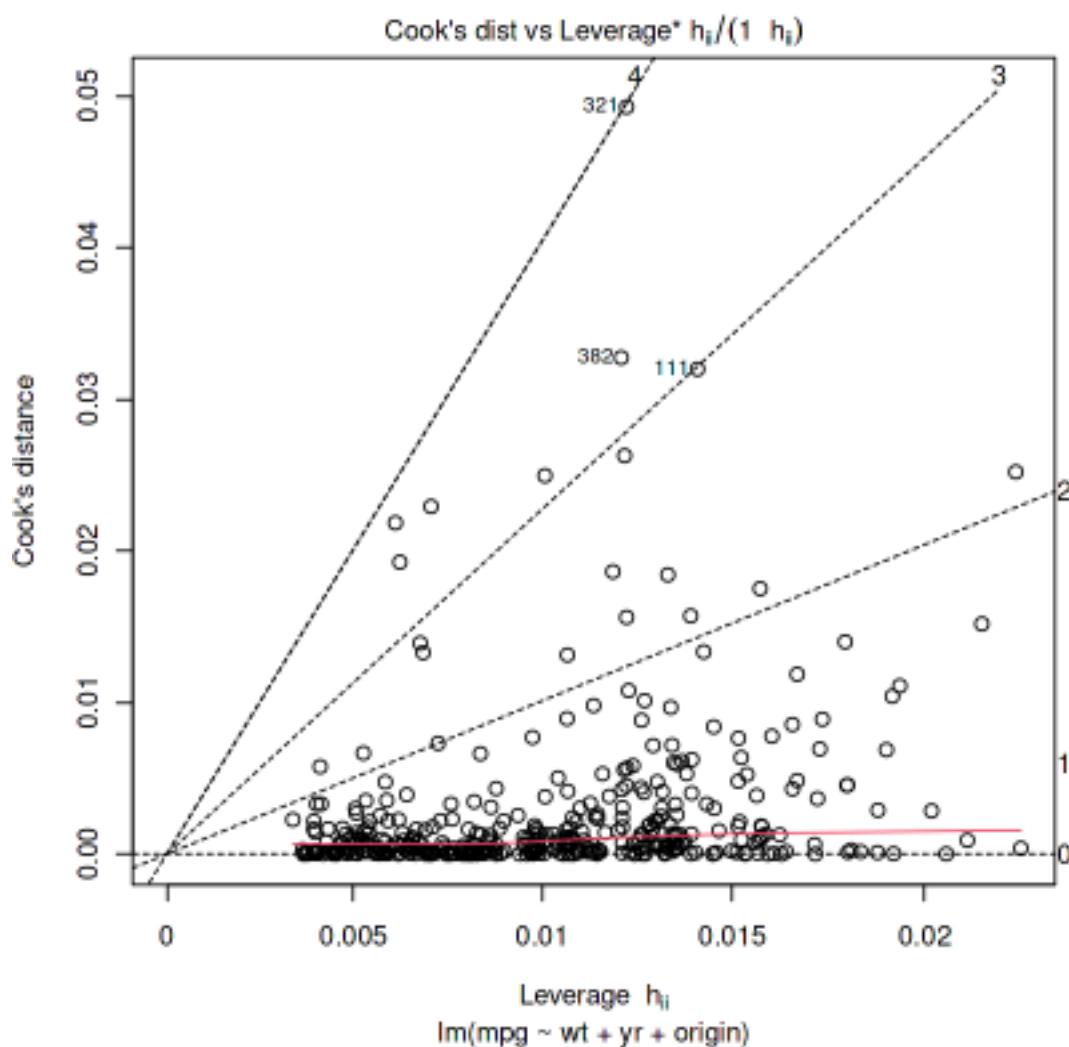
`plot(mpgMod,4):` This gives a graph of Cook's distance, which is yet another way to identify points that have significant influence on a linear regression model. The highest spikes reveal potential troublemakers.



`plot(mpgMod,5)`: This is a plot of residuals versus leverage. The highest and lowest points are the ones to scrutinize.



`plot(mpgMod,6)`: This plot compares leverage with Cook's distance. The dotted lines break the points into categories that look like slices of pizza. The further you rotate counterclockwise, the more influential the points.



After viewing the plots above, which one point seems to consistently be of concern?  
 What car is it? For practice, you may want to remove this one point and redo your model, but you do not need to include this work in your report.

After analyzing all graphs the point that seems to consistently be of concern is 321, which is "Mazda GLC"

**Be sure to include all 6 plots in your report**