BSAN 450 Assignment 3

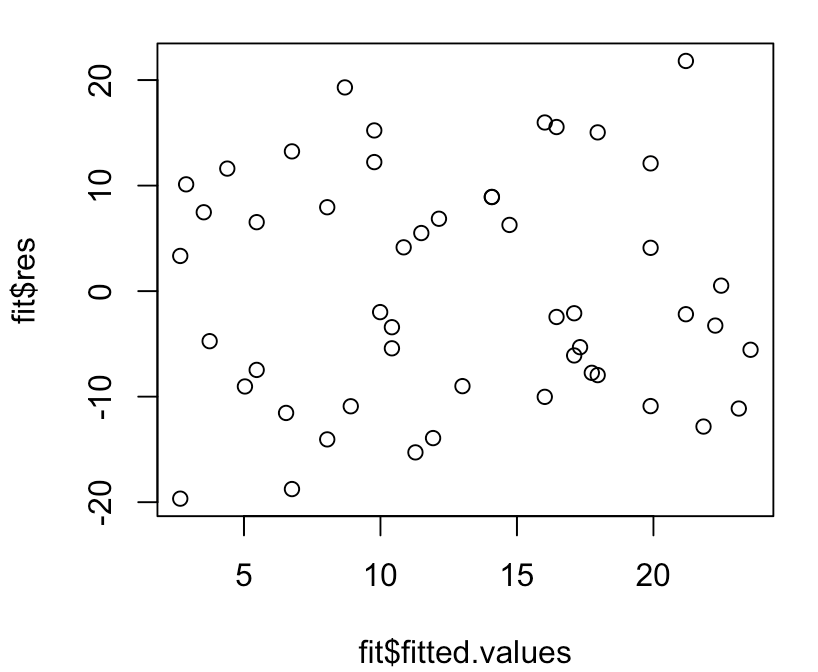
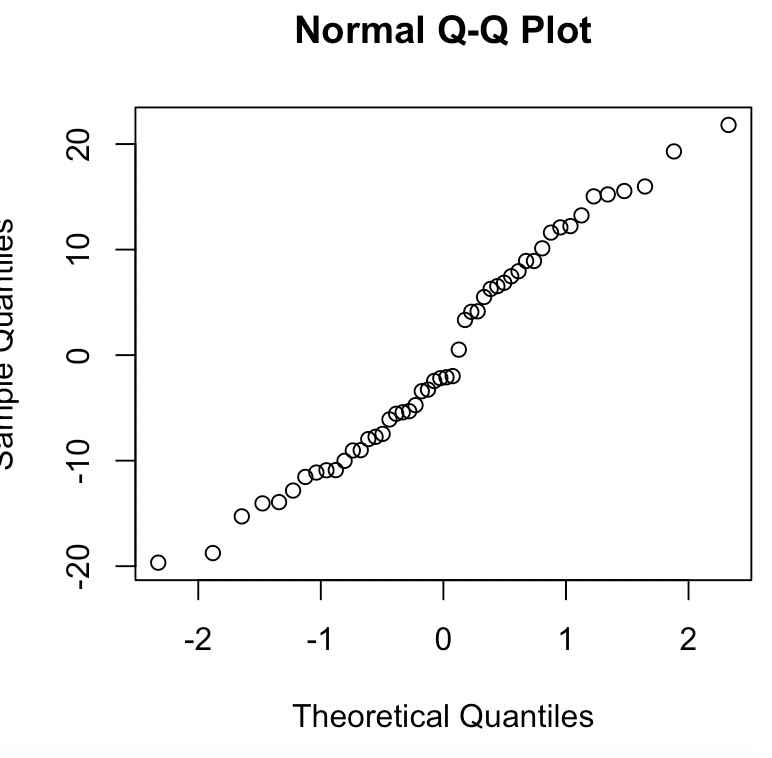
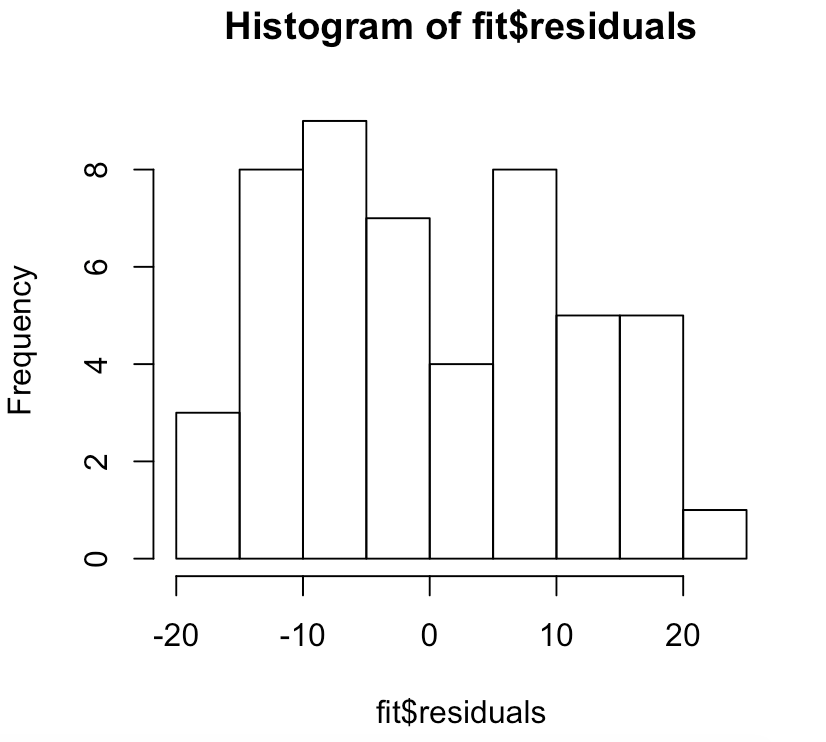
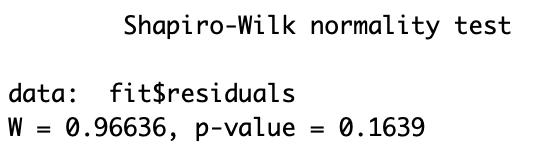
1) The order fulfillment process of a major distribution center is having trouble delivering orders on time. It has been suggested that the order volume has an impact upon the response time to fill the orders. The management of the distribution center is considering a new computer system to handle the increased volume of orders. The data is from 50 recent days in which the order volume each day and the average response time for each day. A negative value indicates an early delivery.

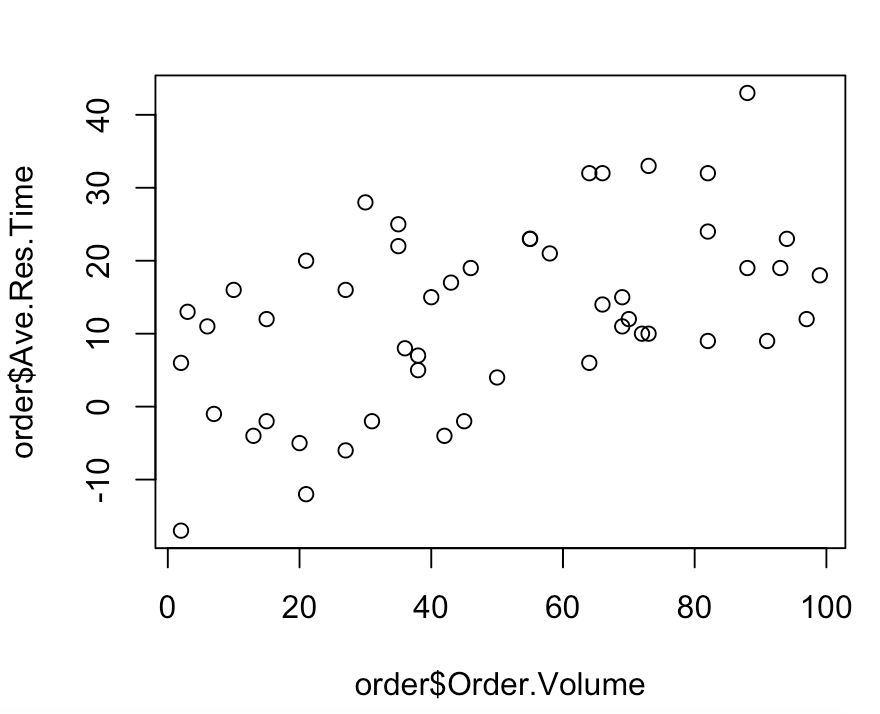
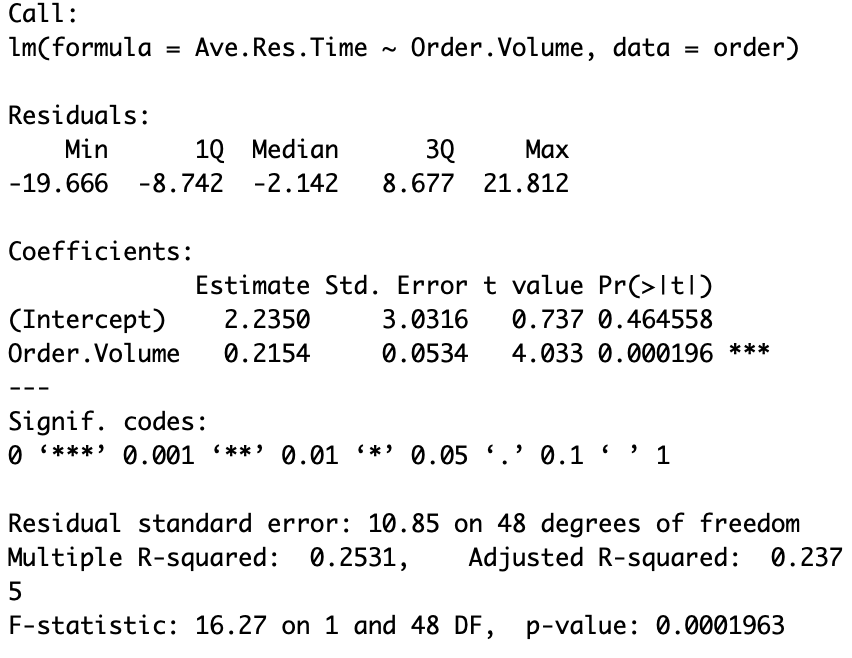
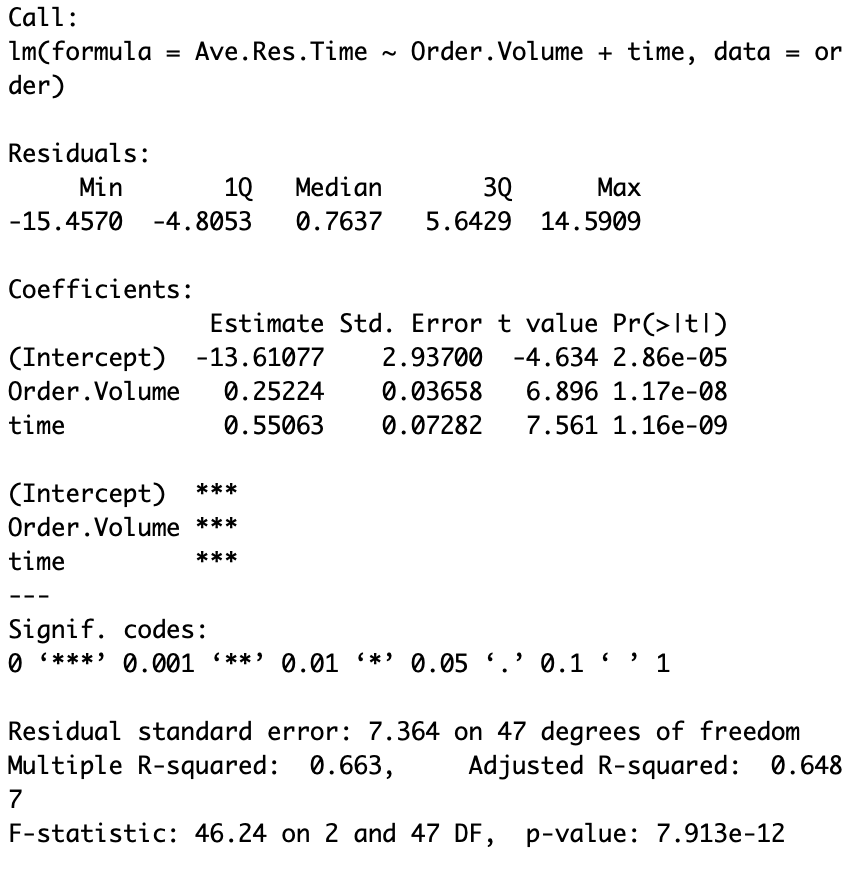
The data is in a file named “orderfulfillment.csv”. The names of the two variables are: Ave.Res.Time and Order.Volume.

Read the data into R Studio using the following command.

order=read.csv("orderfulfillment.csv")

a)

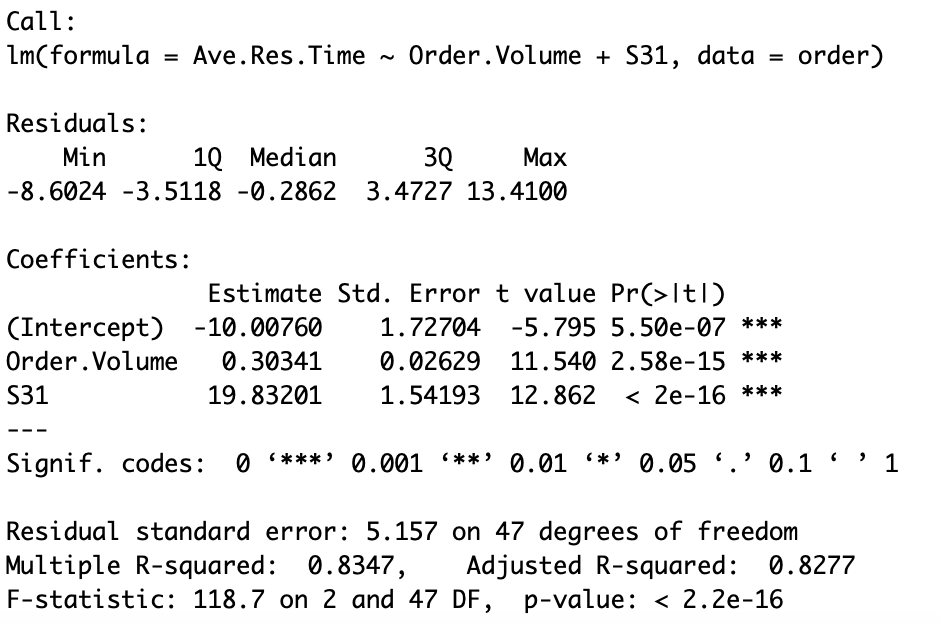
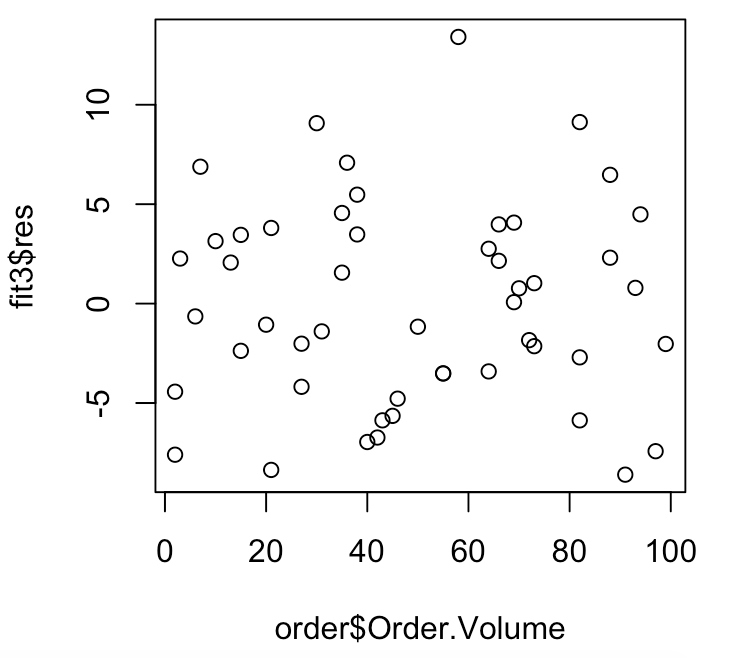
    

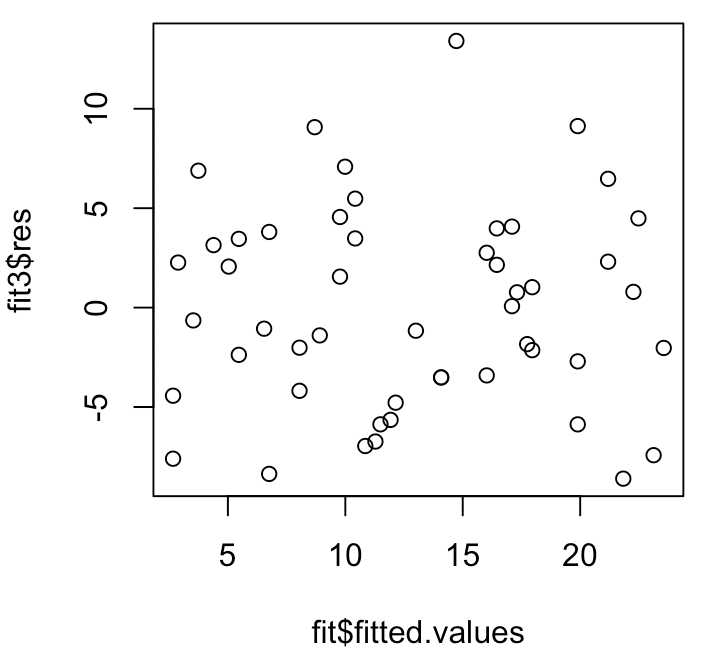
  

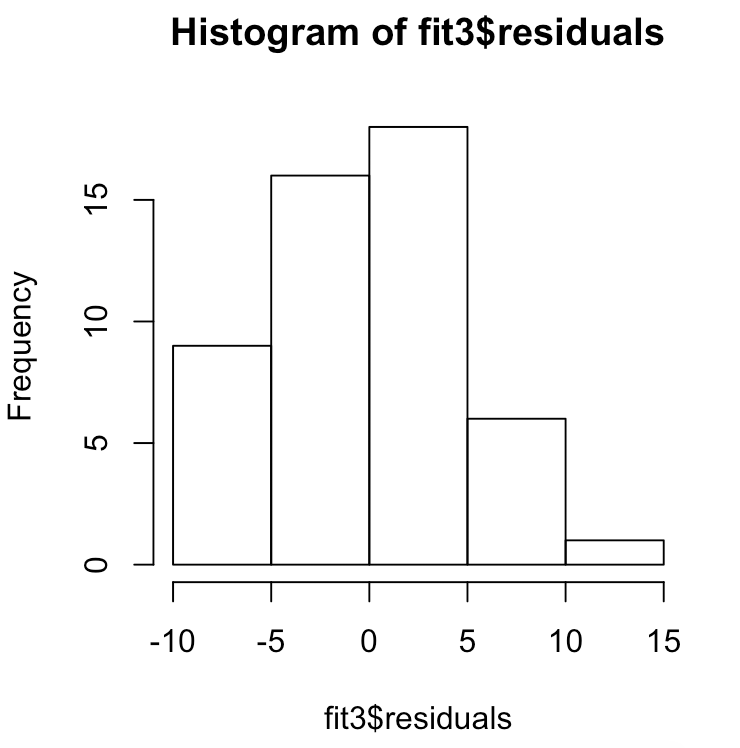
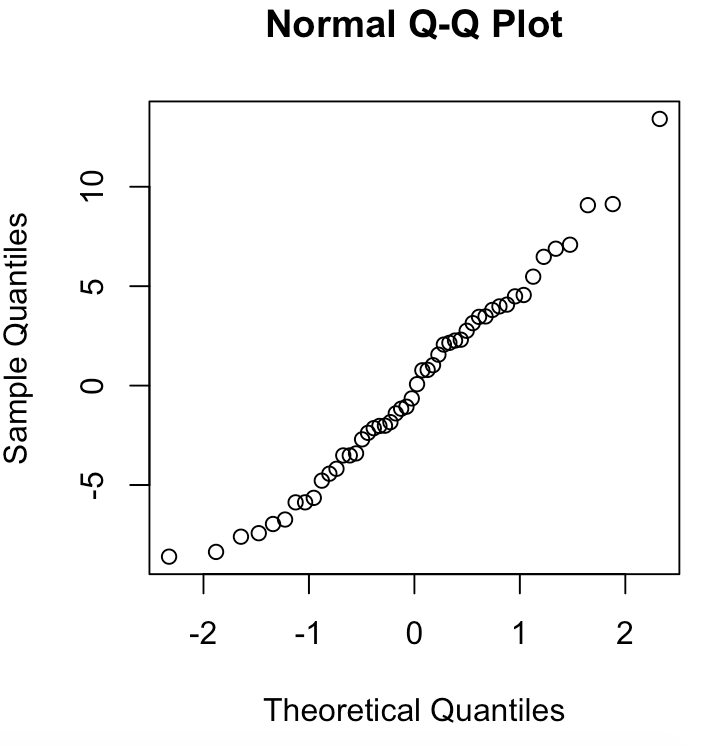
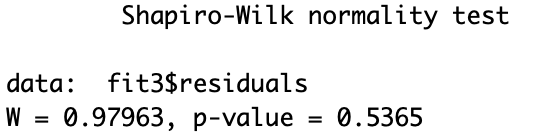
Are there any problems with this model?

When graphing the residual versus order time, there is an upward trend in the data which is a problem. There is also a shift after time 31. All the other diagnostics and graphs for this model look fine.

b)

The diagnostics for the model indicate no issue.

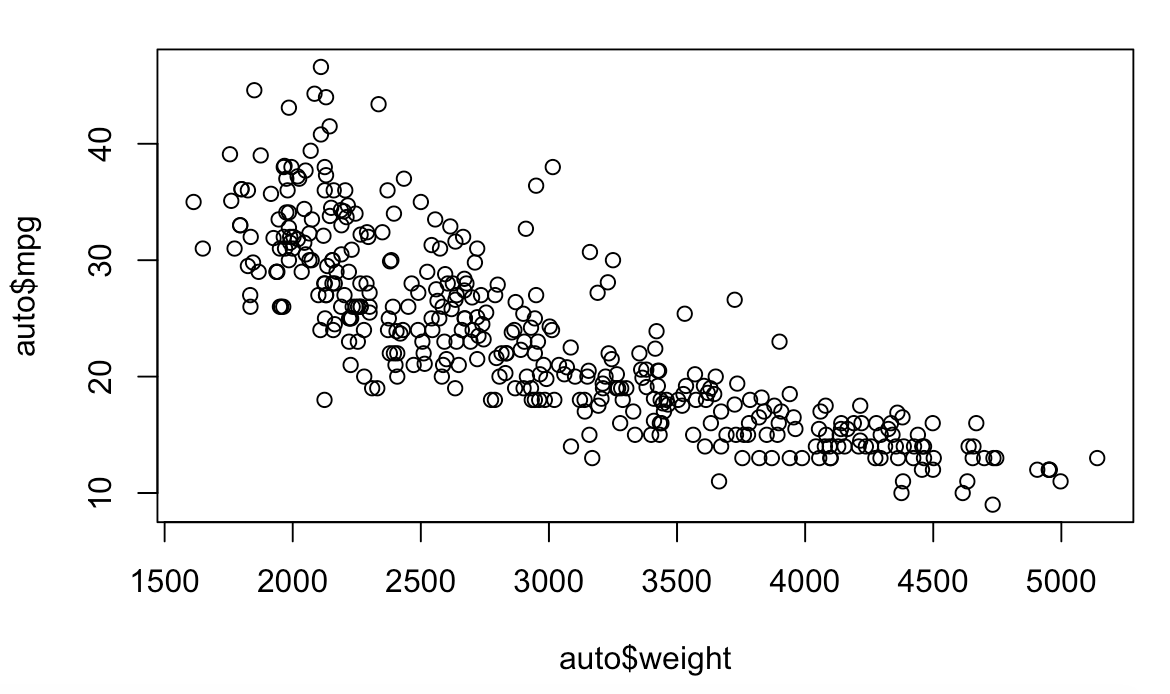
c) Compare the model fit in part a) with the model fit in part b). Which model fits better? Justify your answer.

The fit in part b is a better fit because in part b takes into account time and clearly be looking at the r-squared values the second model is a better fit.

2) We will consider again the Auto data in which we have the miles per gallon (mpg) on 392 vehicles. For this problem we will consider using the weight of the cars as an independent (X) variable. This data is stored in a file named “Auto.csv”.

a) Use the following command to read the data into R Studio. Plot a scatter plot of mpg vs. weight. Comment on this plot, is there a linear relationship between mpg and weight?

auto=read.csv("Auto.csv")



The relationship between weight and mpg looks to be slightly linear. However there is a slight curve in the data suggesting a potential issue.

b) Since the plot suggests a curved relationship between mpg and weight we want use a polynomial in the variable weight in a multiple regression model. You will use cross validation to help determine the degree of the polynomial. We will first consider 10 fold cross validation. You will fit successive polynomials of higher orders and use the cross validation mean squared error to choose the appropriate polynomial order (the lower the cross validation mean squared error, the better).

The following sequence of R commands fits a polynomial of degree 1 and computes the cross validation mean squared error for this model, then fits a polynomial of degree 2 and computes the cross validation mean squared error continuing up to a polynomial of degree 4. Execute the commands and interpret the results. What order of polynomial in the variable weight would you recommend? Justify your answer.

Note, it the boot package is not installed on your computer you will need to install it before executing the library command.

Degree 1 polynomial cross validation: 18.8638

Degree 2 polynomial cross validation: 17.50552

Degree 3 polynomial cross validation: 17.58858

Degree 4 polynomial cross validation: 17.86893

According to the cross validation, a polynomial of degree 2 is the best fit for the data.

c) Suppose that someone wanted you to use 5 fold cross validation rather than 10 fold cross validation. Note that in the R command above the K = 10 instructs R to use 10 fold cross validation. If the value 10 is changed to 5, then R will do 5 fold cross validation. If you wish to do leave one out cross validation you delete the expression K = 10

Repeat the analysis in part b) except use leave one out cross validation and use 5 fold cross validation. How do the results compare to what you obtained in part b)?

K=5 (5 Fold Cross Validation)

|  |  |
| --- | --- |
| Degree | Cross Validation Error |
| 1 | 19.10098 |
| 2 | 17.60524 |
| 3 | 17.58925 |
| 4 | 17.85854 |

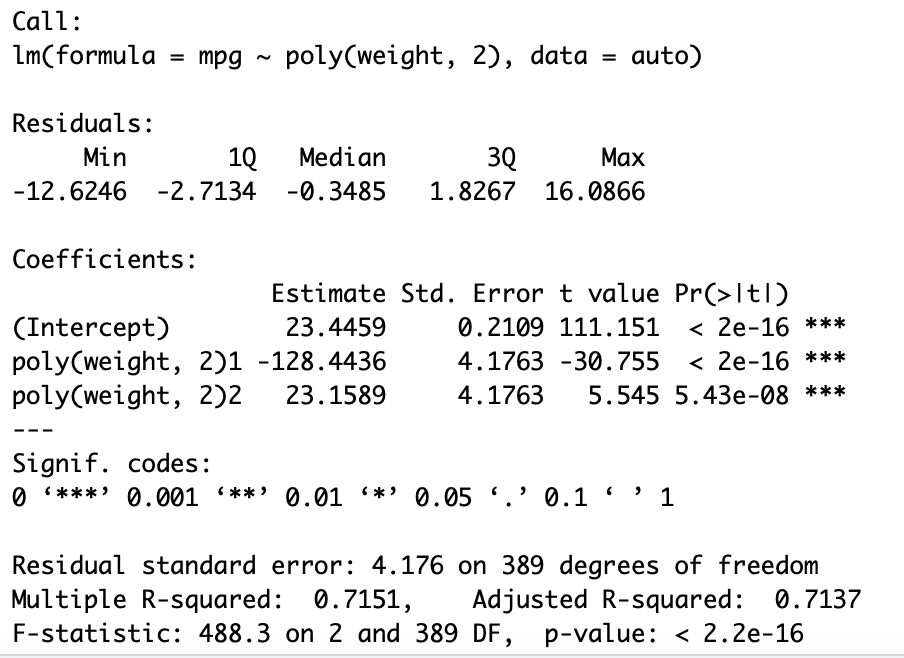
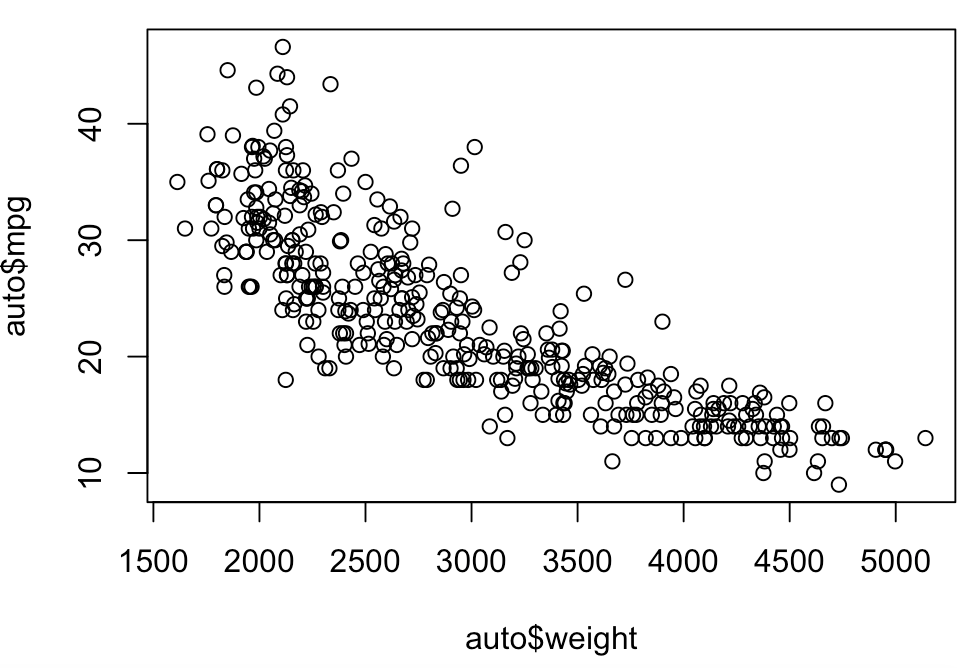
Doing 5 fold cross validation now indicates that the best fit model for the data would be a degree 3 polynomial and not a degree 2 polynomial like before.

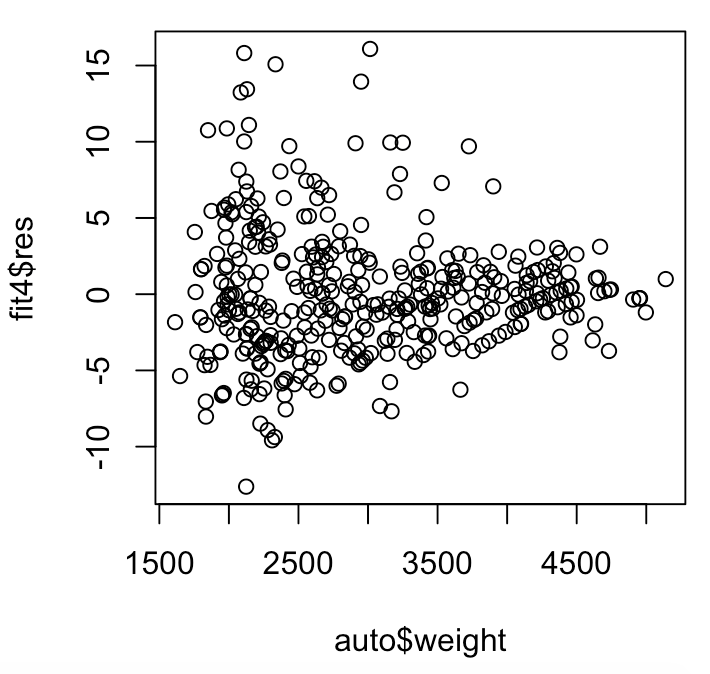
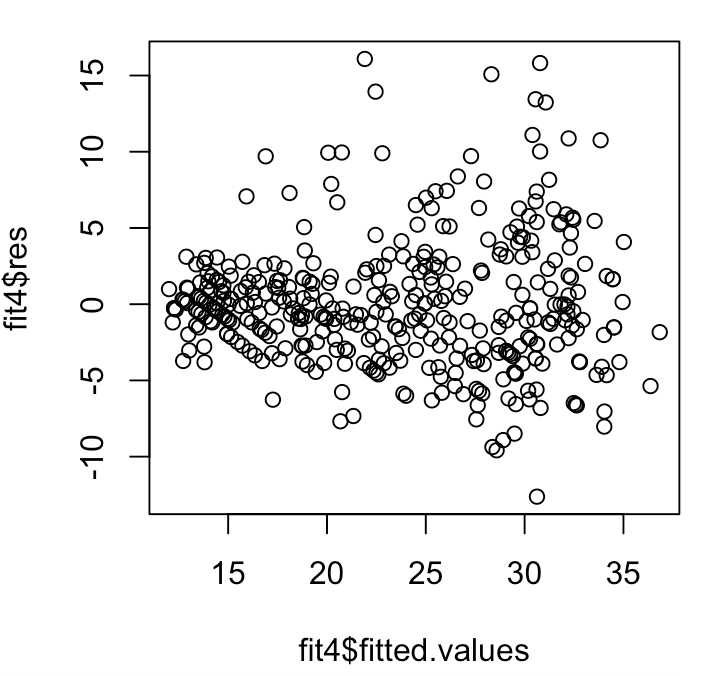
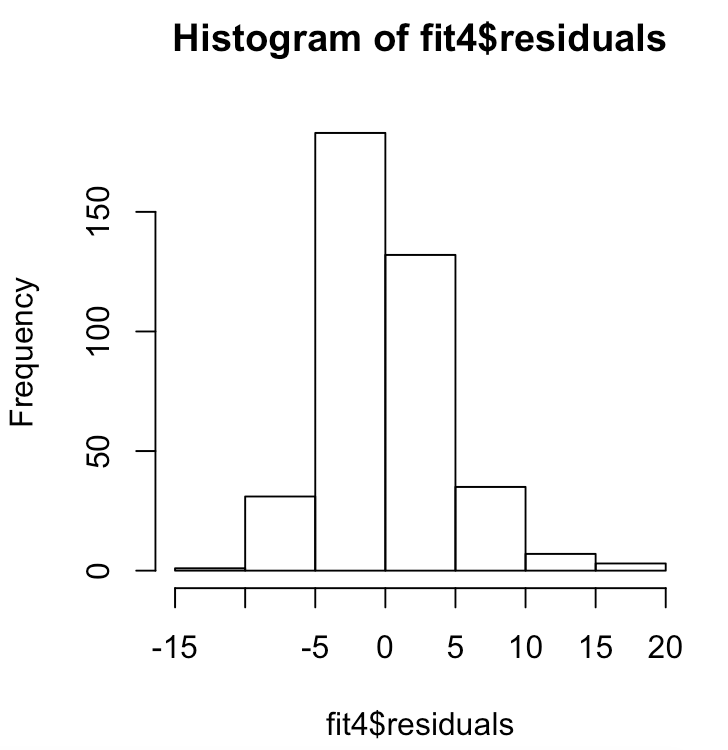
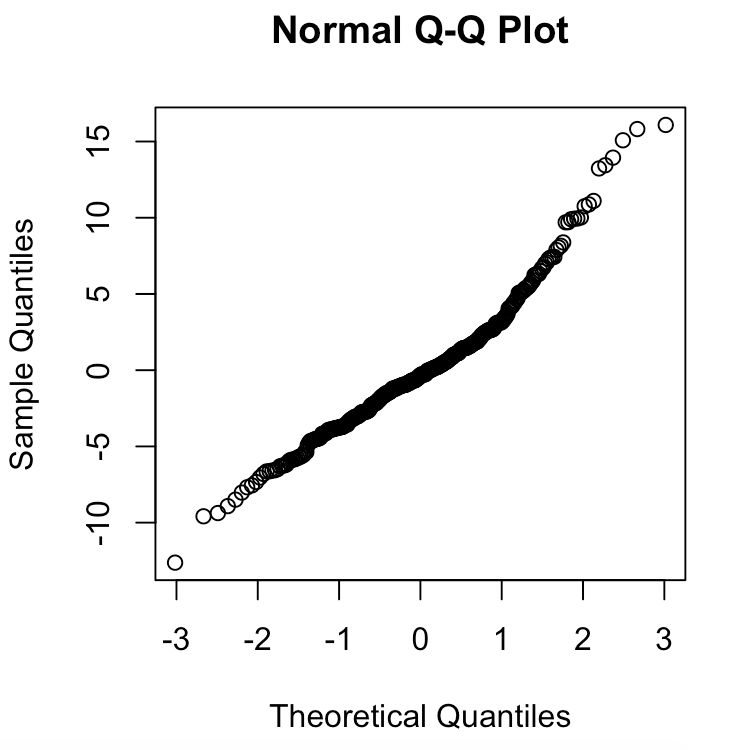
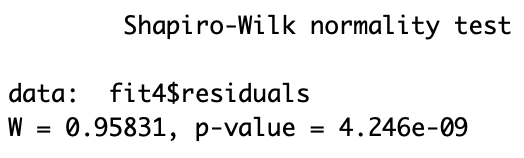
Leave one out Cross Validation

|  |  |
| --- | --- |
| Degree | Cross Validation Error |
| 1 | 18.85161 |
| 2 | 17.52474 |
| 3 | 17.57811 |
| 4 | 17.62324 |

Doing Leave one out Cross Validation still indicates that the best fit for this data is a degree 2 polynomial.

d) For the model you picked in part b, fit this model, print out the summary, and do the diagnostic checks. Comment on the results and the diagnostic checks.

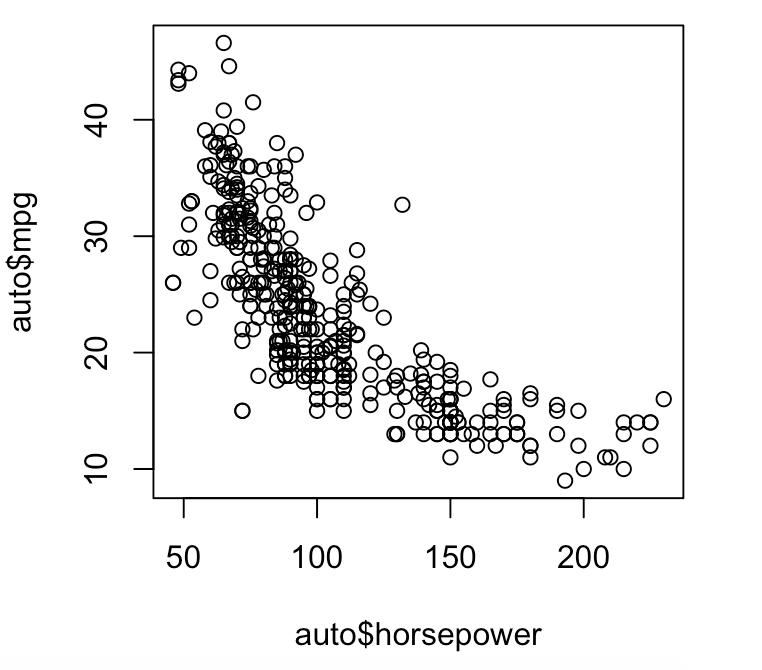
    

The dianogstics for this model look slightly problematic. On both of the residual graphs it appears to have a slightly increase or decrease respectively.

3) We will continue with the analysis of the auto data that was considered in problem 2. Now we will consider using the horsepower of the cars as an independent (X) variable. This data is stored in a file named “Auto.csv”.

a) Use the following command to read the data into R Studio. Plot a scatter plot of mpg vs. horsepower. Comment on this plot, is there a linear relationship between mpg and horsepower?

auto=read.csv("Auto.csv")



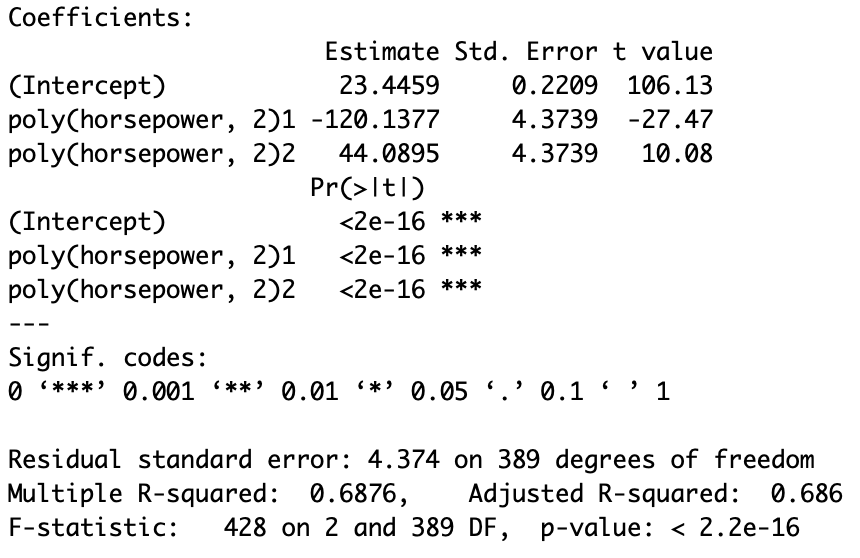
There appears to be a relationship between horsepower and mpg, however it does not look linear, but polynomial.

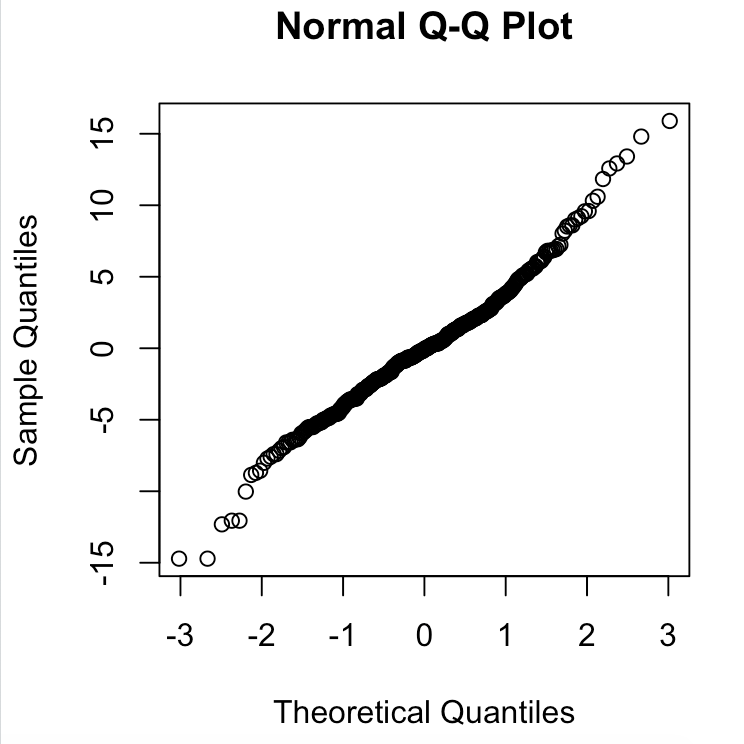
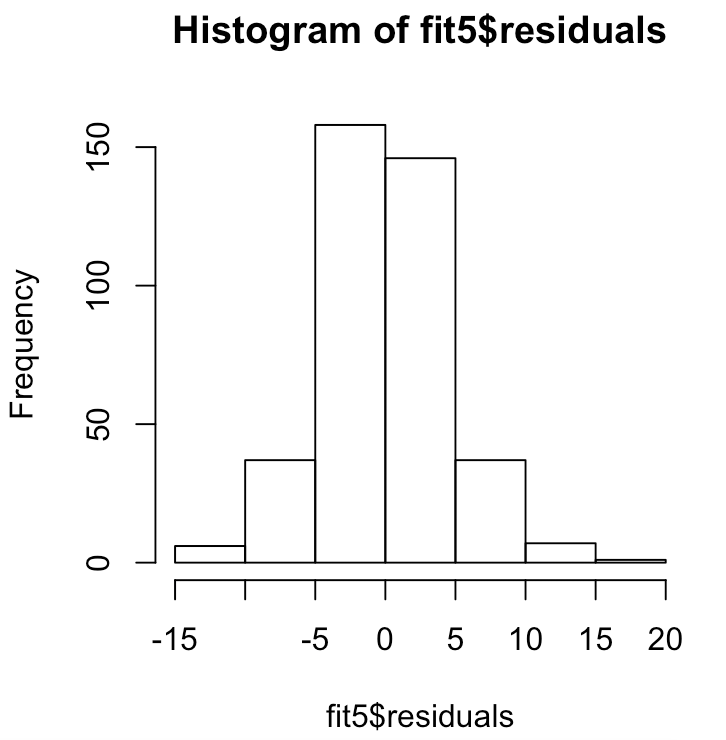
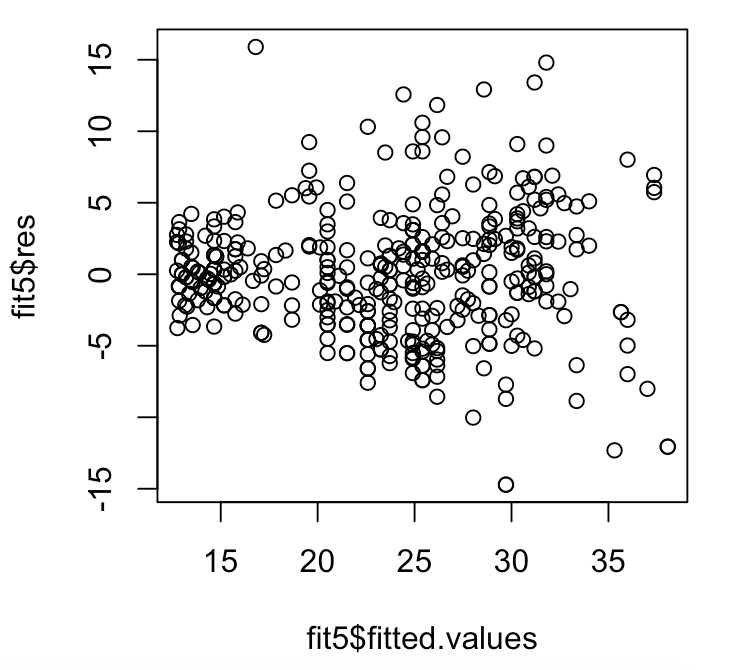
b) Since the plot suggests a curved relationship between mpg and weight we want use a polynomial in the variable weight in a multiple regression model. Use 10 fold cross validation to determine the degree of the polynomial that is needed to describe the relationship between mpg and horsepower.

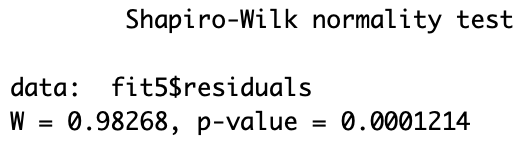
|  |  |
| --- | --- |
| Degree | Cross Validation Error |
| 1 | 24.25012 |
| 2 | 19.28027 |
| 3 | 19.43691 |
| 4 | 19.30127 |

Using 10 fold cross validation, it indicated that using a polynomial of degree 2 is the best fit.

c) For the model you picked in part b, fit this model, print out the summary, and do the diagnostic checks. Comment on the results and the diagnostic checks.







The graph of the fitted values and the residual looks problematic, because it appears as the variance might be increasing.