Jucoen Yeater
MA598 – Data Analysis
Dr. Zubovic
Assignment 4 – Multiple Linear Regression

I began by selecting the numerical data and performing a correlation plot on them as shown in figure 1. The plot shows that rawhumidity has the smallest, as well as negative correlation with cnt. The variable rawwindspeed is the second smallest, but it looks to have a more significant negative correlation with cnt. The other two variables rawfeeltemp and raw temp look to have about the same correlation with cnt and they are both positive. These all make sense when they are thought about. If there is a high wind speed then people are less likely to want to be outside so a higher rawwindspeed will result in a lower cnt. If the blue variables show that a higher temperature will result in a higher cnt. People want to walk in warm weather. The final variable, rawhumidity, shows that humidity doesn't seem to be a large factor when people decide to go for a walk. I am going to start by fitting all of the numerical variables, but rawhumidity may end up being removed from the model since it has such a low correlation.

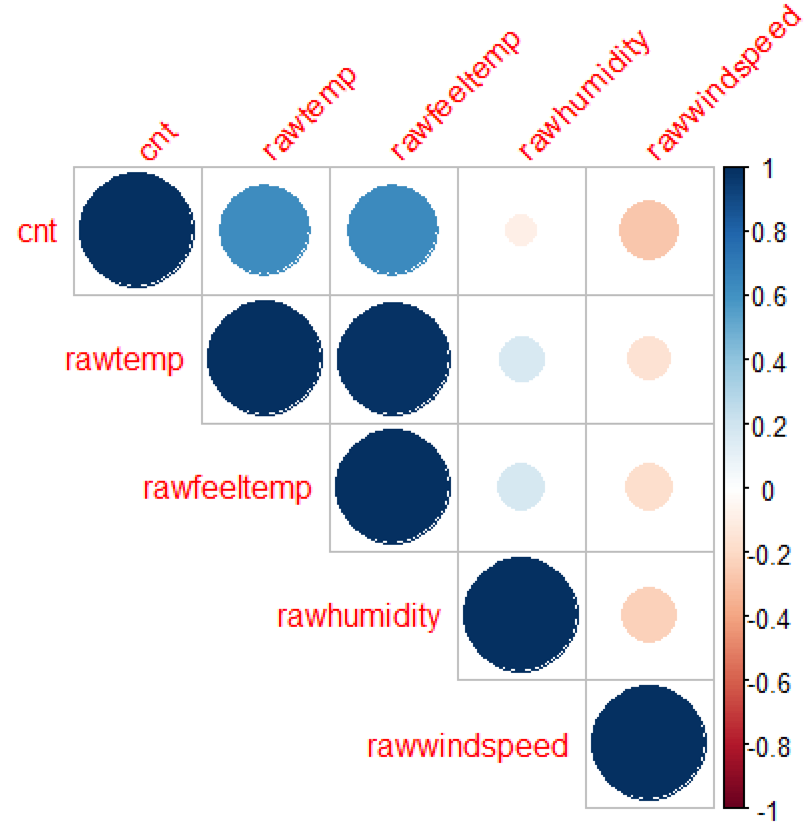


Figure 1: Correlation Plot for numerical variables

The first model fit used the following variables: cnt, season, workingday, weathersit, rawtemp, rawfeeltemp, rawhumidity, and rawwindspeed. The variables season, workingday, and weathersit are also included since they all would probably affect the count of people who walked. Running a summary of this model showed that workingday and weathermist have a p-value that is greater than 0.05 so those two variables are not statistically significant at 5%. It also showed that the adjusted R^2 value was

0.5686 which doesn't seem like a good value. Using the Anova table it shows that workingday is not statistically significant and can be removed from the model. The variable rawtemp was also removed because after running vif on the model it showed that they had greater than 10.

A summary was run on the new model with the two mentioned variables removed. Performing a summary (which is shown in figure 5 as I didn't think to include this until after the other figures) on the updated model returned a p-value that was less than 0.05 and the regression is statistically significant. However, the adjusted R^2 value is 0.5604 which means that the model can predict 56% of the variability in the count of rentals. The coefficients for each variable are interesting. All are negative except for rawfeeltemp. From the coefficients it is more likely that Spring results in the highest cnt and Winter would result in the lowest cnt of the seasons. Now for an analysis of the model.

The Residuals vs Fitted values plot is shown in Figure 2. It shows that the residuals look more to the right side of the graph with the larger fitted values. This means that the model has non-constant variance. Also, to note from the plot is that the residuals look to be cone shaped heading to the right which shows that the model is roughly linear. Figure 3 shows the histogram of the residuals. The histogram does not look to normally distributed. The qqplot shown in figure 4 agrees with that and shows that the residuals deviate from the line near the endpoints as well as a small part in the middle. I think that this represents that its not normally distributed more near the endpoints. A Shapiro-Wilk test on the model returns a p-value that is less than 0.05 which agrees that the variance is non-constant. This all together shows that this is not a great model for the data. Testing described after the plots.

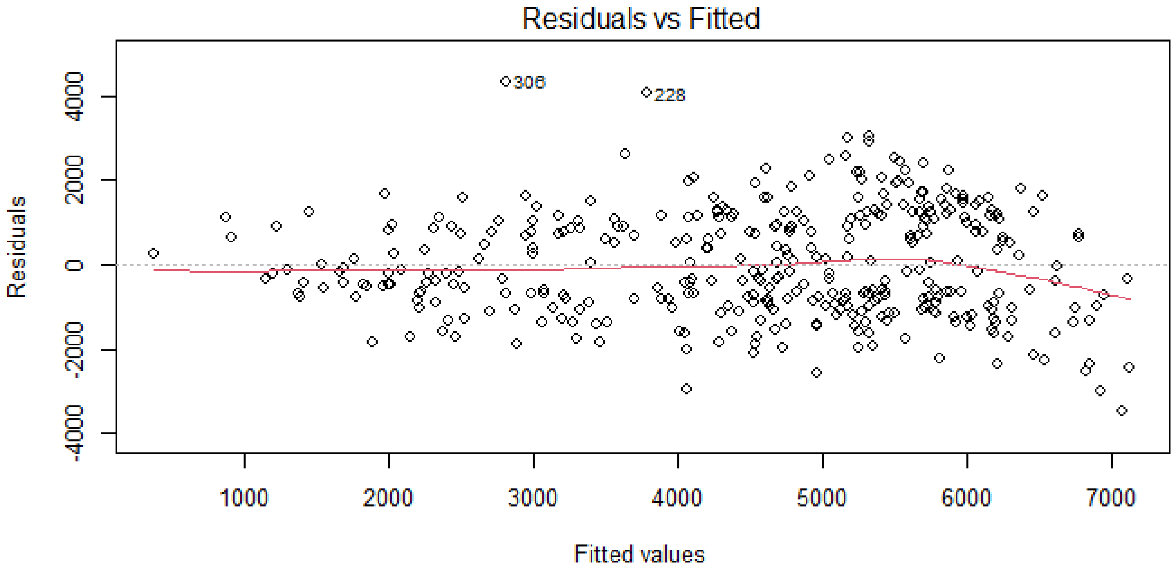


Figure 2: Plot of Residuals

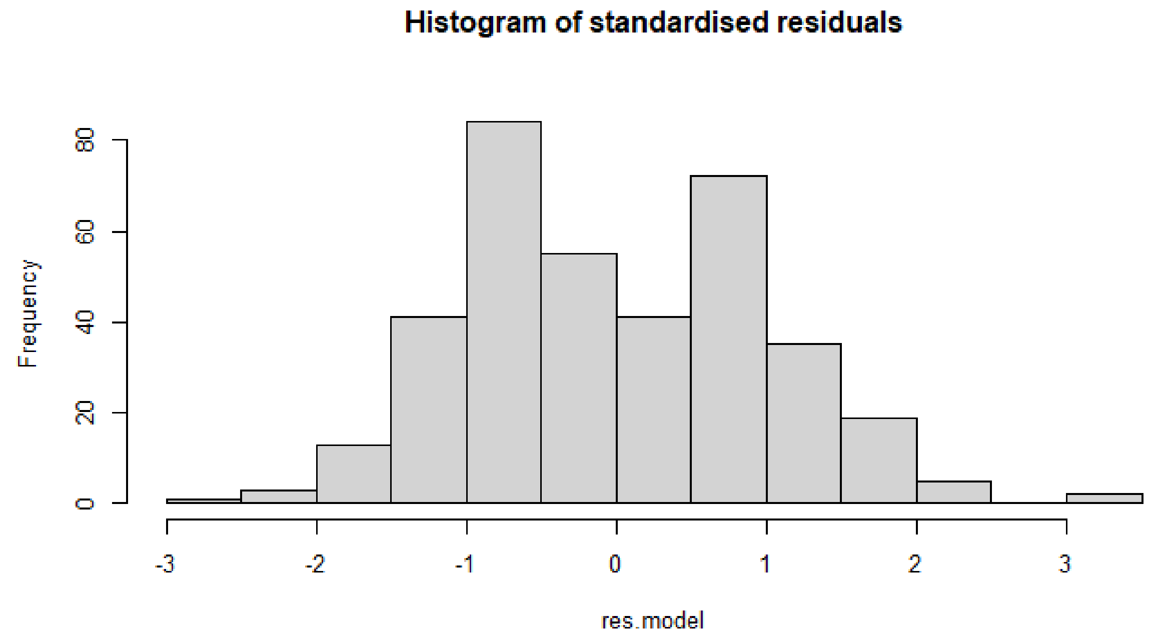


Figure 3: Histogram of Residuals

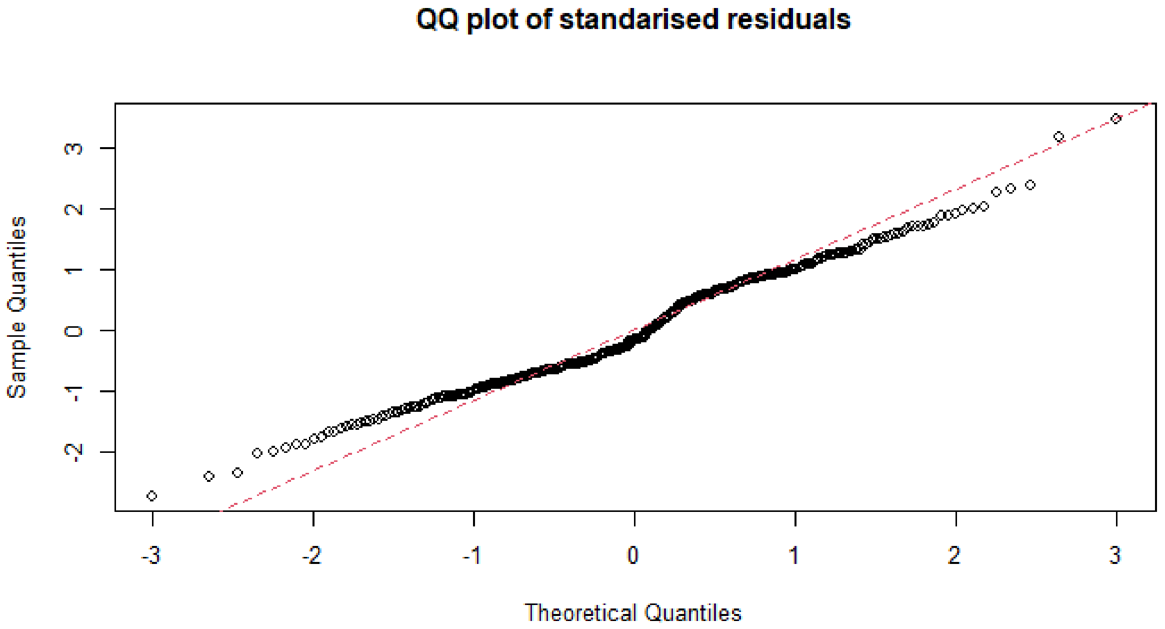
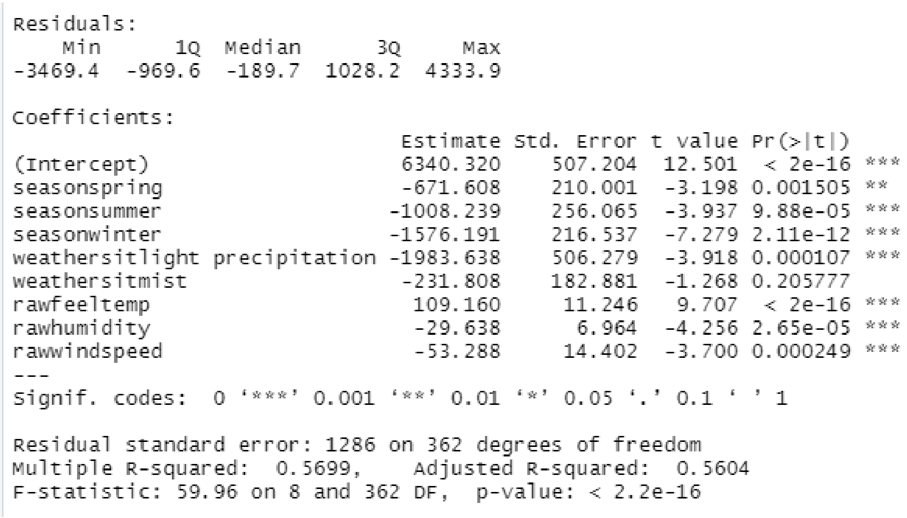


Figure 4: QQ plot of residuals



```
Residuals:
    Min      1Q  Median      3Q     Max 
-3469.4  -969.6  -189.7  1028.2  4333.9 

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)    
(Intercept)                   6340.320    507.204  12.501  < 2e-16 ***
seasonspring                  -671.608    210.001  -3.198 0.001505 ** 
seasonsummer                 -1008.239    256.065  -3.937 9.88e-05 ***
seasonwinter                 -1576.191    216.537  -7.279 2.11e-12 ***
weathersitlight precipitation -1983.638    506.279  -3.918 0.000107 ***
weathersitmist                -231.808    182.881  -1.268 0.205777    
rawfeeltemp                    109.160     11.246   9.707  < 2e-16 ***
rawhumidity                    -29.638      6.964  -4.256 2.65e-05 ***
rawwindspeed                   -53.288     14.402  -3.700 0.000249 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1286 on 362 degrees of freedom
Multiple R-squared:  0.5699,	Adjusted R-squared:  0.5604 
F-statistic: 59.96 on 8 and 362 DF,  p-value: < 2.2e-16
```

Figure 5: Summary of the final model

The model can now be tested on the test data provided. This was done by using the predict function in R and then computing the RMSE (Root Mean Squared Error) which gives the average difference between the cnt variable and the predicted cnt value from the model. The R-square was also looked at since it represents the correlation between cnt and the predicted value for cnt. A small RMSE (compared to the mean of the cnt) and a large R-square value are wanted. The mean of cnt is 4455.99 and the RMSE is 1328.446 which is 30% of the mean of cnt which is quite large and shows that there is a high average difference between cnt and the predicted value for cnt. The R-square value is 0.5359 which is similar to the adjusted R-square of the model. Both of these R-square values seem to be too low for the model to be accurate. This shows that the model does not fit the data well and is not a good predictor for cnt.