Week 6 Regression Report Sample

For this sample report, I’ll use a portion of R’s built-in data set mtcars. I created a data file with five of the variables from that set for the purposes of this sample report.

## Question

Can we predict the mpg of a car from its engine displacement, horsepower, weight, and number of gears?

## Understanding the Data

After loading the data, R created the following summary.

mpg disp hp wt   
 Min. :10.40 Min. : 71.1 Min. : 52.0 Min. :1.513   
 1st Qu.:15.43 1st Qu.:120.8 1st Qu.: 96.5 1st Qu.:2.581   
 Median :19.20 Median :196.3 Median :123.0 Median :3.325   
 Mean :20.09 Mean :230.7 Mean :146.7 Mean :3.217   
 3rd Qu.:22.80 3rd Qu.:326.0 3rd Qu.:180.0 3rd Qu.:3.610   
 Max. :33.90 Max. :472.0 Max. :335.0 Max. :5.424   
 gear   
 Min. :3.000   
 1st Qu.:3.000   
 Median :4.000   
 Mean :3.688   
 3rd Qu.:4.000   
 Max. :5.000

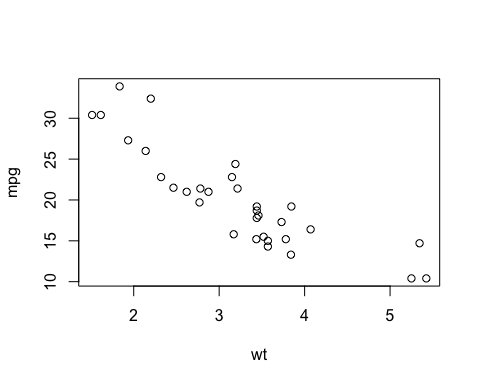
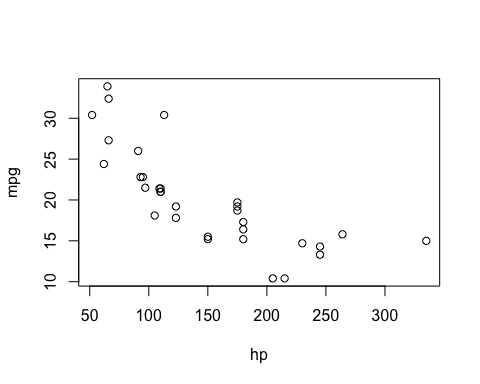
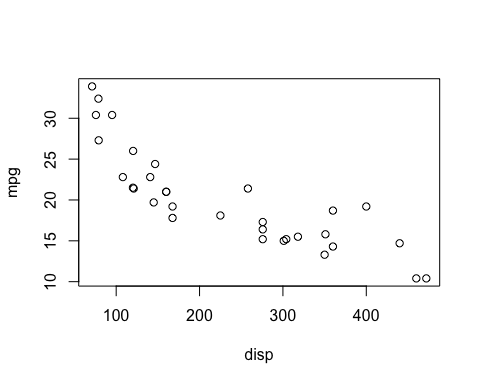
The variable gear only has three values which seems to suggest that we should treat it as a categorical variable, not numerical. We specified that gear is a factor in R.

We check the correlation between the other variables.

mpg disp hp wt  
mpg 1.0000000 -0.8475514 -0.7761684 -0.8676594  
disp -0.8475514 1.0000000 0.7909486 0.8879799  
hp -0.7761684 0.7909486 1.0000000 0.6587479  
wt -0.8676594 0.8879799 0.6587479 1.0000000

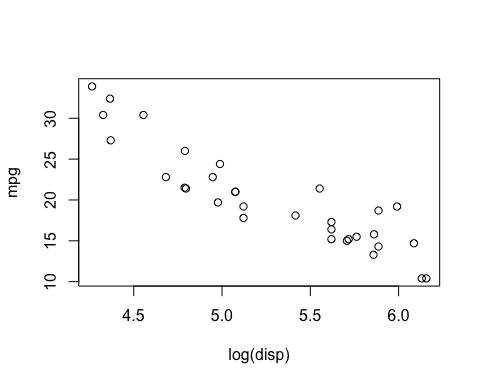
These all seem to have fairly strong linear relationships to mpg.

We will produce the scatterplots now.

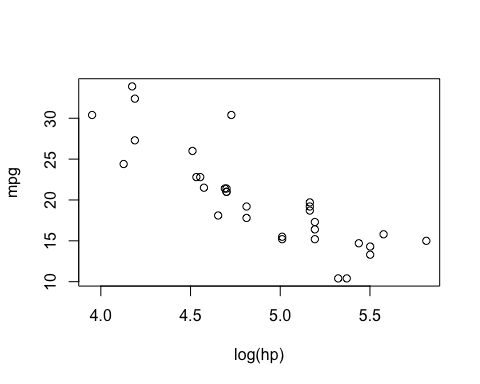


Displacement and horsepower both show some non-linearity. We will transform both and then attempt to build the model.

First we transform disp with a log function and store this in a new variable, log\_disp. The new scatterplot is below.



Next we transform hp with the log function and store this in a variable log\_hp. The new scatterplot is below.



These scatterplots show a much more linear relationship. We check the correlation coefficients to verify and see the following values:

mpg log\_disp log\_hp wt  
mpg 1.0000000 -0.9071119 -0.8487707 -0.8676594  
log\_disp -0.9071119 1.0000000 0.8617723 0.8845389  
log\_hp -0.8487707 0.8617723 1.0000000 0.7158277  
wt -0.8676594 0.8845389 0.7158277 1.0000000

The correlation coefficients for mpg with both log\_disp and log\_hp have increased.

## Building the Model

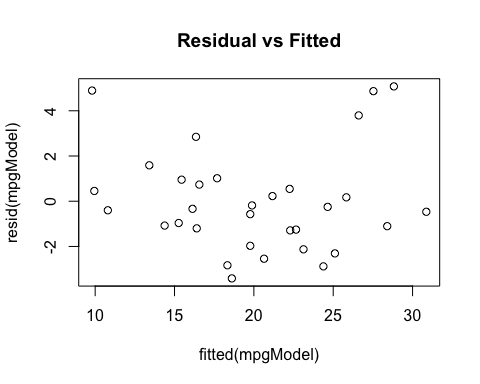
Finally we are ready to generate our model. We initially use all variables then use backward elimnation to remove unnecessary variables. Our process eliminates gear, then log\_disp, leaving us with the following model.

Call:  
lm(formula = mpg ~ log\_hp + wt)  
  
Residuals:  
 Min 1Q Median 3Q Max   
-3.4130 -1.2642 -0.3679 0.7902 5.0780   
  
Coefficients:  
 Estimate Std. Error t value Pr(>|t|)   
(Intercept) 59.5709 4.9769 11.970 9.64e-13 \*\*\*  
log\_hp -5.9218 1.2658 -4.678 6.20e-05 \*\*\*  
wt -3.2856 0.6148 -5.344 9.74e-06 \*\*\*  
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 2.339 on 29 degrees of freedom  
Multiple R-squared: 0.8591, Adjusted R-squared: 0.8494   
F-statistic: 88.44 on 2 and 29 DF, p-value: 4.542e-13

Both of the remaining variables and the model overall are significant.

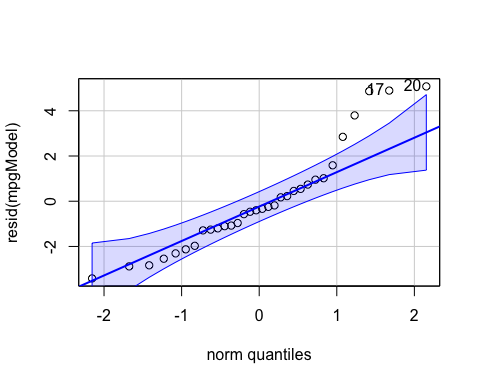
## Model Assumptions

We will check model assumptions to see if any more transformations are necessary.



There seems to be some nonlinearity, but the variance seems roughly equal. This is a somewhat worrisome plot.

The qqPlot below shows several points on the large end that leave the dashed lines. This combined with the residual plot above indicates a data transformation would be helpful.



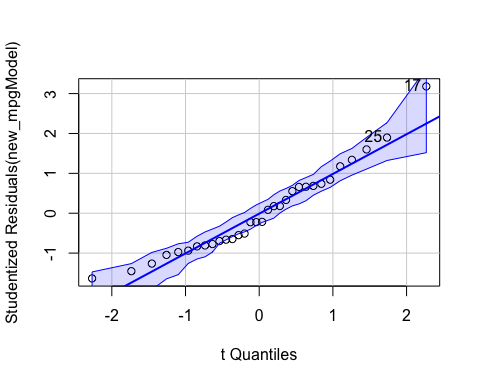
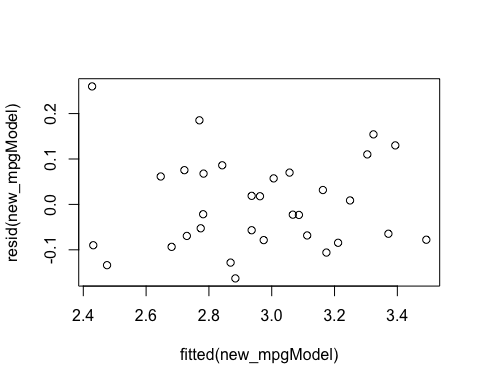
We will perform a log transformation on the response variable and create a new variable log\_mpg.

With this transformed variable, the linear model now looks like this:

Call:  
lm(formula = log\_mpg ~ log\_hp + wt)  
  
Residuals:  
 Min 1Q Median 3Q Max   
-0.16296 -0.07799 -0.02210 0.06837 0.25985   
  
Coefficients:  
 Estimate Std. Error t value Pr(>|t|)   
(Intercept) 4.83167 0.22198 21.766 < 2e-16 \*\*\*  
log\_hp -0.26566 0.05646 -4.706 5.75e-05 \*\*\*  
wt -0.17942 0.02742 -6.543 3.63e-07 \*\*\*  
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 0.1043 on 29 degrees of freedom  
Multiple R-squared: 0.8852, Adjusted R-squared: 0.8773   
F-statistic: 111.8 on 2 and 29 DF, p-value: 2.338e-14

So far the model looks better with a slightly larger adjusted R-squared and even smaller p-values for several of the variables.

We now check the model assumptions with residual plots.



These both seem improved. There is no more non-linearity in the residual plot. The normal-probability plot has most of the points very close to the straight line. This seems to be the best model we can construct from this dataset.

## Conclusion

Overall the model meets all assumptions and the R-squared value is rather high at 88%. We should be confident in using this model to predict mpg for cars based on their weight and horsepower.