Regression_Assignment

treepruner

September 22, 2015

Executive Summary

The mtcars dataset was used determine a parsimonious model between the outcome variable, miles per gallon (MPG), and the other variables in the data set. Once a model was identified with an r squared above 80% and both regressors with p values less than 5%, the transmission variable was added in. The final model determined that a manual transmission is better for MPG holding weight and quarter second time constant. The model estimates a 2.935837 miles per gallon increase in manual transmissions.

Exploratory Data Analyses

My strategy for model selection was to first identify linear relationships between mpg and the other variables by running cor(mtcars), then plot the variables with a correlation > .75 using ggpairs.

Some of the variables are correlated to each other. Disp is higly correlated to cyl. Wt is correlated to disp and cyl. Hp is highly correlated to cyl and disp. See ggpairs output in appendix.

Fit Models

I ran a series of simple linear regression with each variable identified above.

```
wt <- lm(mpg ~ wt, data = mtcars)
cyl <- lm(mpg ~ cyl, data = mtcars)
disp <- lm(mpg ~ disp, data = mtcars)
hp <- lm(mpg ~ hp, data = mtcars)
drat <- lm(mpg ~ drat, data = mtcars)
qsec <- lm(mpg ~ qsec, data = mtcars)
vs <- lm(mpg ~ vs, data = mtcars)
am <- lm(mpg ~ am, data = mtcars)
gear <- lm(mpg ~ gear, data = mtcars)
carb <- lm(mpg ~ carb, data = mtcars)
```

The model with lowest p value was wt <- lm(mpg ~ wt, data = mtcars).

```
##      model r.squared adj.r.squared fstatistic p value
## [1,] "wt"  "0.7528"  "0.7446"      "91.38"    "0.0000000001293958701350553"
```

Next, new models were created with wt as the 1st regressor and each of the remaining variables was tested as the second regressor.

```
## Add regressors to wt
wt_drat <- lm(mpg ~ wt + drat, data = mtcars)
wt_qsec <- lm(mpg ~ wt + qsec, data = mtcars)
wt_vs <- lm(mpg ~ wt + vs, data = mtcars)
wt_am <- lm(mpg ~ wt + am, data = mtcars)
wt_gear <- lm(mpg ~ wt + gear, data = mtcars)
wt_carb <- lm(mpg ~ wt + carb, data = mtcars)
```

The model wt_qsec <- lm(mpg ~ wt + qsec, data = mtcars) had the lowest p value for the 2nd regressor.

```
##      model     r.squared adj.r.squared fstatistic p value
## [1,] "wt_qsec" "0.8264"  "0.8144"      "69.03"    "0.00000000003"
```

I ran an anova to compare the 2 models. The model with wt and qsec was a real improvement. Now that I had a model to predict mpg, I added in the transmission variable to differentiate the effect of transmission type.

Interpreting the Coefficients of the Final Model

```
##               Estimate Std. Error   t value       Pr(>|t|)
## (Intercept)  9.617781  6.9595930  1.381946 0.177915165459
## wt          -3.916504  0.7112016 -5.506882 0.000006952711
## qsec         1.225886  0.2886696  4.246676 0.000216173705
## factor(am)1  2.935837  1.4109045  2.080819 0.046715509919
```

As the weight goes up by a unit of 1 (which is lbs/1000), the mpg will decrease by 3.916504 miles per gallon. As the quarter mile time goes up by a quarter second, the mpg will increase by 1.225886 miles per gallon. All things being held equal, lighter weight, slower cars in the 1/4 mile, will have better MPG.

The intercept is what changes between the transmission types. The automatic transmission, am = 0, has an Intercept of 9.617781. The Intercept for a manual transmission is 9.617781 + 2.935837. The final models are:

```
# manual mpg = 9.617781 -3.916504 * wt + 1.225886 * qsec
# automatic mpg = (9.617781 + 2.935837) -3.916504 * wt + 1.225886 * qsec
```

Evaluating the Model

The final model was NOT significantly better than the wt + qsec model, but the am variable is significant and does identify the effect of transmission type, the purpose of our analysis.

```
## Analysis of Variance Table
##
## Model 1: mpg ~ wt
## Model 2: mpg ~ wt + qsec
## Model 3: mpg ~ wt + qsec + factor(am)
##   Res.Df    RSS Df Sum of Sq       F    Pr(>F)
## 1     30 278.32
## 2     29 195.46  1    82.858 13.7048 0.0009286 ***
## 3     28 169.29  1    26.178  4.3298 0.0467155 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The sqrt(vif(fit)) is below 2, so the Variance Inflation Factor (VIF) VIF is ok and there doesn't appear to be an issue with multi-collinearity.

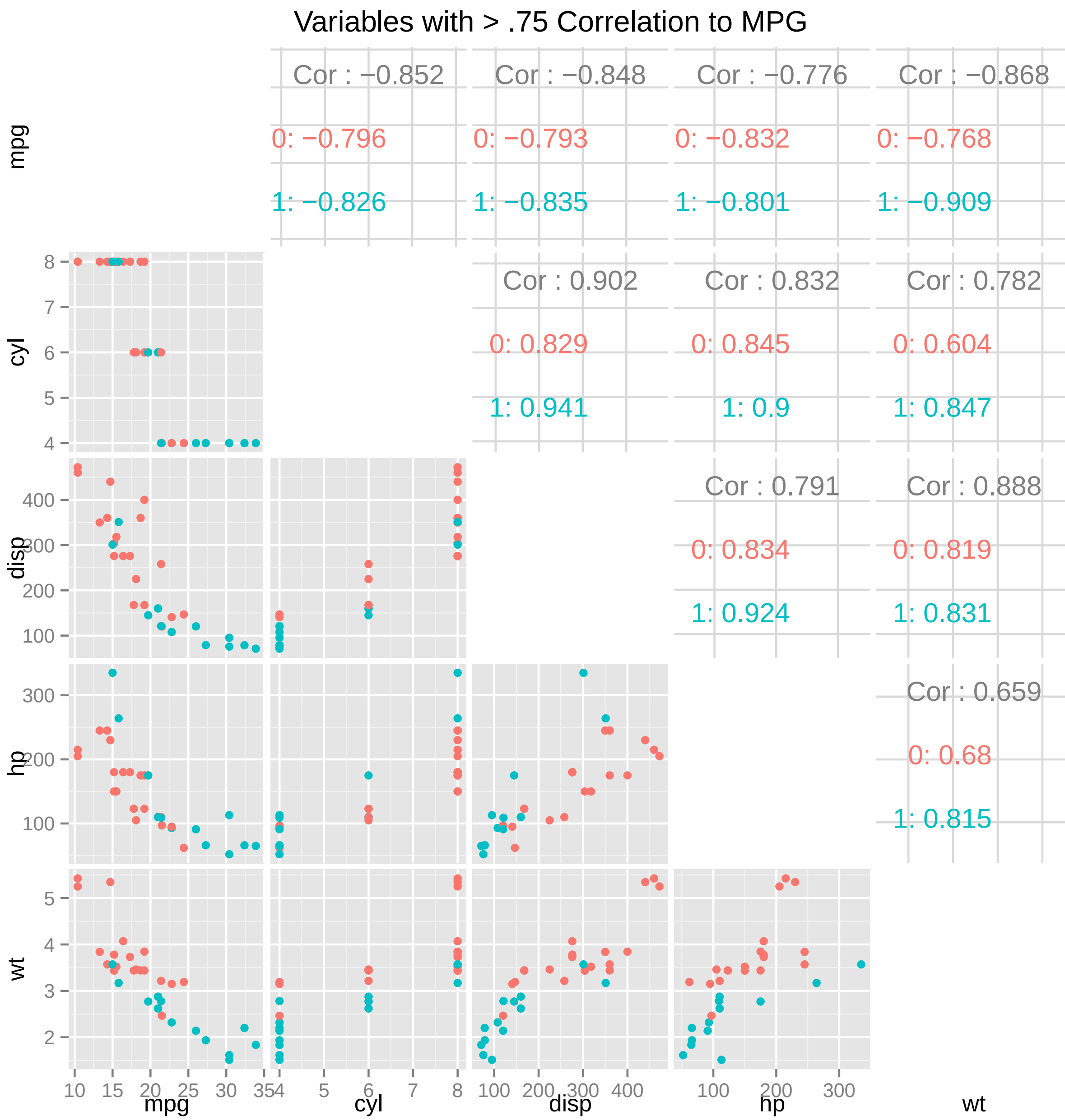
```
##         wt       qsec factor(am)
##   1.575738   1.168049   1.594189
```

The largest dfbetas value, 1.093842173234, is for Chrysler Imperial. The hatvalues went from Merc 450SLC at 0.05303857 to Merc 230 at 0.29704218.

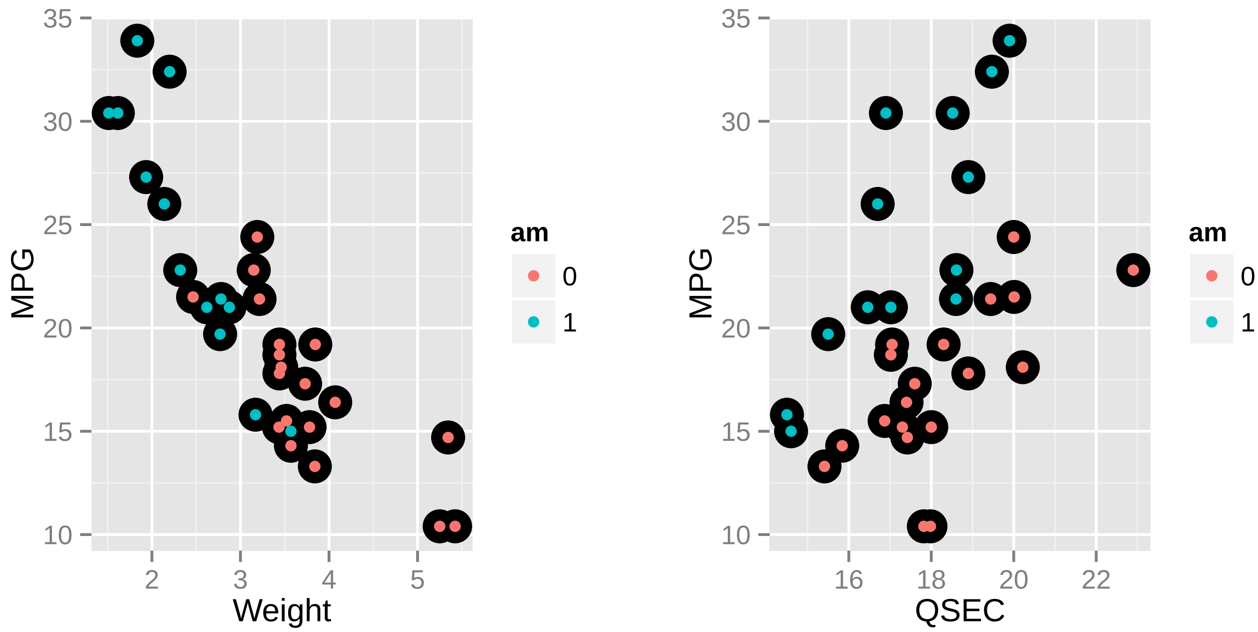
The Residuals vs Fitted Values plot didn't reveal a systematic pattern, which is good. The Normal Q-Q plot evaluates normality in the error terms and it looked ok.

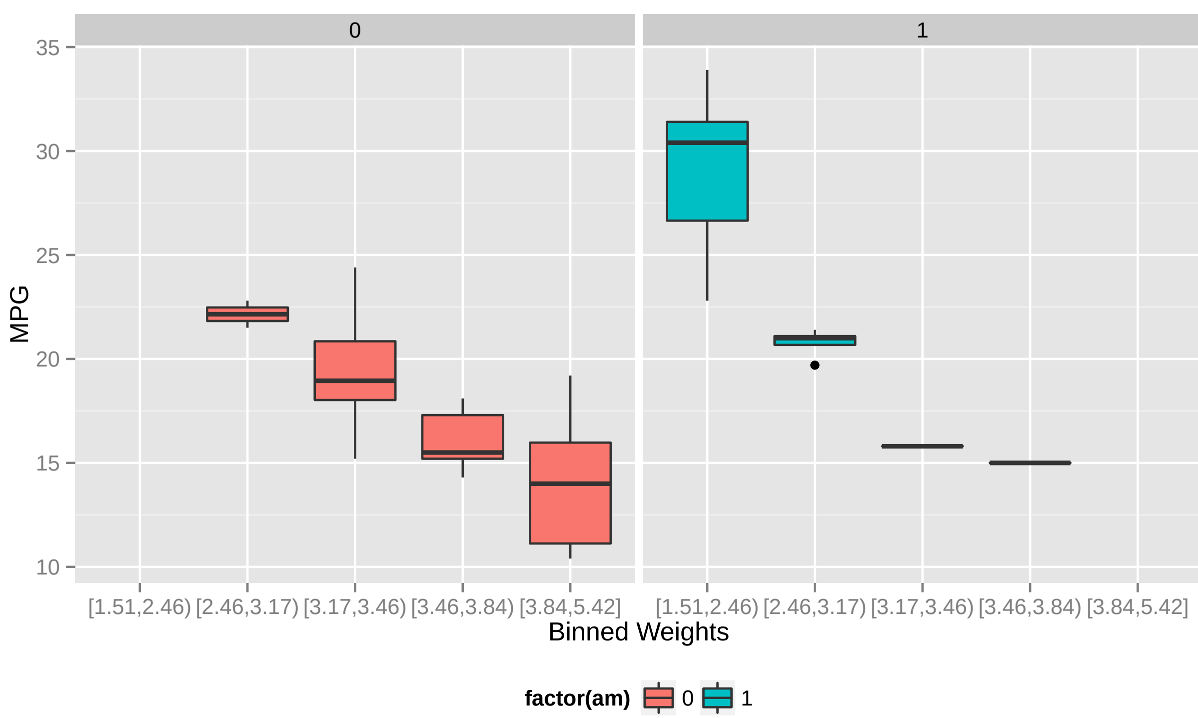
Appendix Area

Exploratory Figures

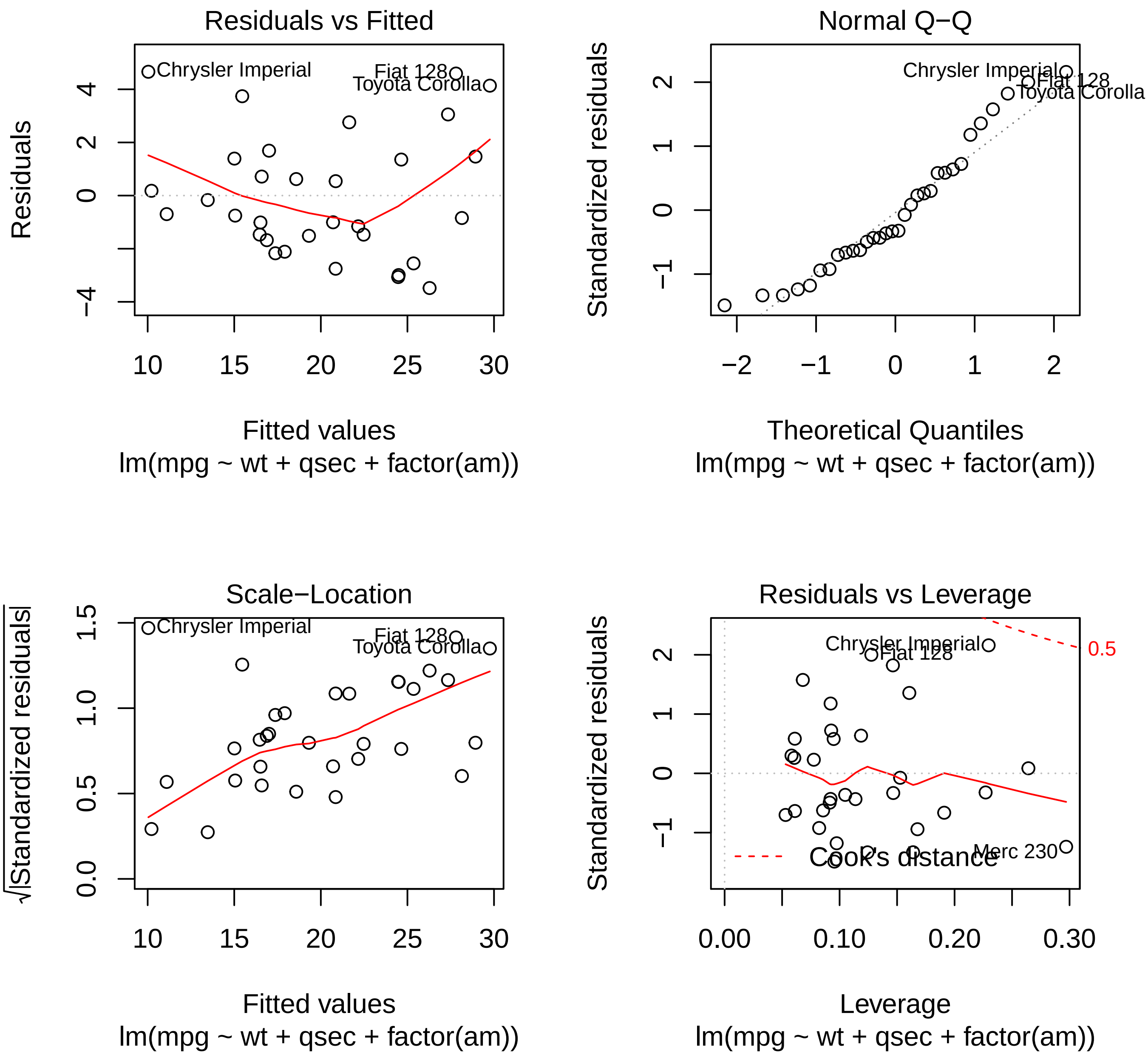


Key Variables vs MPG





Plot the Model



This PDF was created in Knitr