Motor Trend Data Analysis

Regression Models Course Project

Ray Qiu

October 22, 2015

Executive Summary

This analysis uses some data science techniques to analyze the mtcars data set, and explore the relationship between a set of variables and miles per gallon (MPG) (outcome). The key findings are:

- Manual transmission is better than automatic transmission for MPG.
- The ratio between manual and automatic transmission for MPG is 1.806099, adjusted by cyl, disp, hp, and wt.

Load required R libraries

```
library(ggplot2)
library(gridExtra)
```

Load the mtcars data and perform some basic exploratory data analysis.

```
data(mtcars)
# Convert cyl, am and gear to factors
mtcars$cyl <- as.factor(mtcars$cyl)
mtcars$am <- factor(mtcars$am, labels = c("automatic", "manual"))
mtcars$gear <- as.factor(mtcars$gear)
```

Based on Plot #1 in Appendix, we can conclude the following: Manual transmission is better than automatic transmission for MPG.

Let's try to see what other variables should be included in the model

```
anova(lm(mpg ~ ., data = mtcars))
```

```
## Analysis of Variance Table
##
## Response: mpg
##           Df Sum Sq Mean Sq F value    Pr(>F)
## cyl        2 824.78  412.39 60.2490 5.953e-09 ***
## disp       1  57.64   57.64  8.4214  0.009139 **
## hp         1  18.50   18.50  2.7031  0.116598
```

```
## drat       1  11.91   11.91  1.7407  0.202734
## wt         1  55.79   55.79  8.1503  0.010134 *
## qsec       1   1.52    1.52  0.2227  0.642342
## vs         1   0.30    0.30  0.0441  0.835841
## am         1  16.57   16.57  2.4203  0.136271
## gear       2   5.02    2.51  0.3668  0.697741
## carb       1   3.95    3.95  0.5771  0.456767
## Residuals 19 130.05    6.84
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We can pick the variables that has a p-value close to 0.05, which are: cyl, disp, hp, wt, and am.

```
fit2 <- lm(mpg ~ cyl + disp + hp + wt + am, data = mtcars)
summary(fit2)
```

```
##
## Call:
## lm(formula = mpg ~ cyl + disp + hp + wt + am, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.9374 -1.3347 -0.3903  1.1910  5.0757
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 33.864276   2.695416  12.564 2.67e-12 ***
## cyl6        -3.136067   1.469090  -2.135   0.0428 *
## cyl8        -2.717781   2.898149  -0.938   0.3573
## disp         0.004088   0.012767   0.320   0.7515
## hp          -0.032480   0.013983  -2.323   0.0286 *
## wt          -2.738695   1.175978  -2.329   0.0282 *
## ammanual     1.806099   1.421079   1.271   0.2155
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.453 on 25 degrees of freedom
## Multiple R-squared:  0.8664, Adjusted R-squared:  0.8344
## F-statistic: 27.03 on 6 and 25 DF,  p-value: 8.861e-10
```

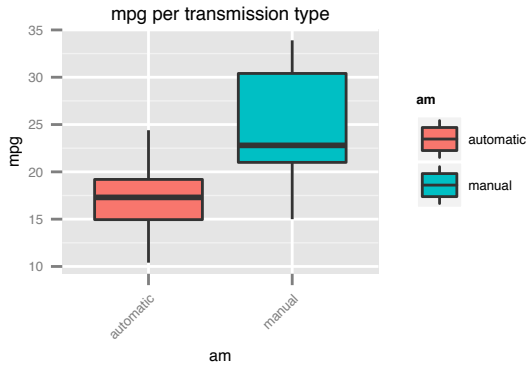
Now, the "Adjusted R-squared" is 0.8344, which we believe is a good fit model for the data.

Answer to Question #2: Cars with Manual transmission get better MPG than Automatic transmission, and the coefficient is 1.806099, adjusted by cyl, disp, hp, and wt.

Appendix: Plots for the model

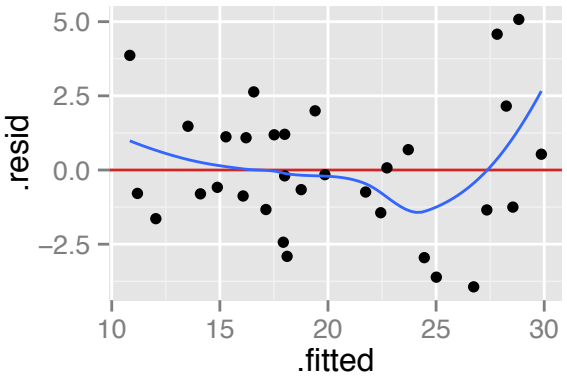
Plot to show the relationships between mpg and manual/automatic transmission.

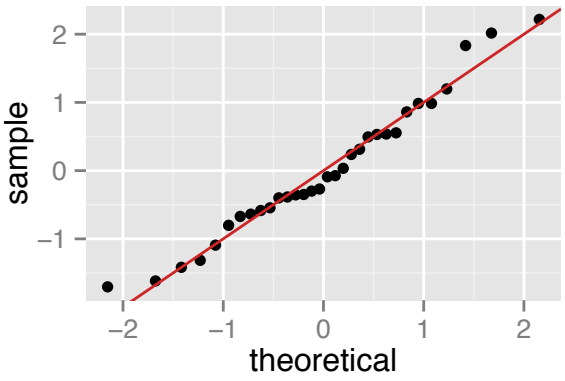
```
p <- ggplot(mtcars, aes(x = am, y = mpg)) +
    ggtitle("mpg per transmission type") +
    geom_boxplot(aes(fill = am)) +
    theme(text = element_text(size=6), axis.text.x = element_text(angle = 45, hjust = 1))
grid.arrange(p, ncol = 2, nrow = 2) # Use grid to make the plot smaller (1/4 size)
```

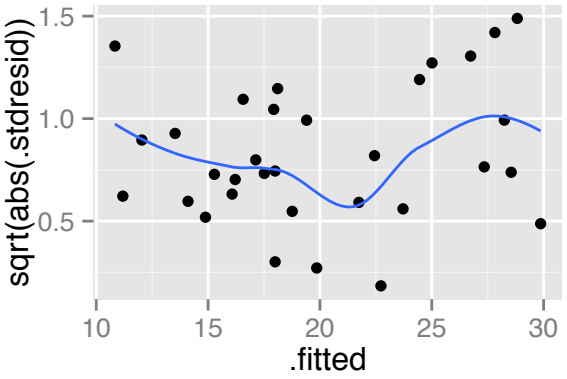


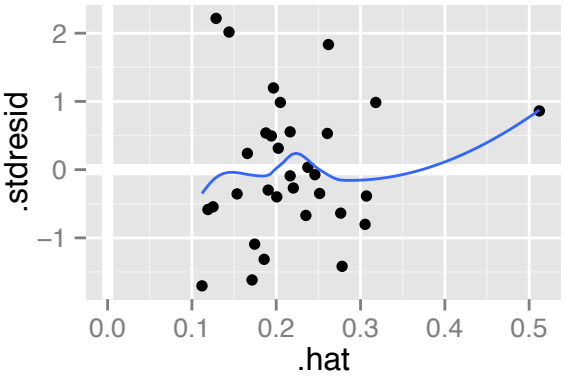
Plots for the model

```
mtcars <- fortify(fit2)
plot1 <- ggplot(data = mtcars, aes(x = .fitted, y = .resid)) +
    geom_hline(yintercept = 0, colour = "firebrick3") +
    geom_point() +
    geom_smooth(se = FALSE, method = loess)
plot2 <- ggplot(data = mtcars, aes(sample = .stdresid)) +
  stat_qq() +
  geom_abline(colour = "firebrick3")
plot3 <- ggplot(data = mtcars, aes(x = .fitted, y = sqrt(abs(.stdresid)))) +
    geom_point() +
    geom_smooth(se = FALSE, method = loess)
plot4 <- ggplot(data = mtcars, aes(.hat, .stdresid)) +
    geom_vline(size = 2, colour = "white", xintercept = 0) +
    geom_hline(size = 2, colour = "white", yintercept = 0) +
    geom_point() +
    geom_smooth(se = FALSE, method = loess)
grid.arrange(plot1, plot2, plot3, plot4, ncol = 2)
```


.resid
5.0
2.5
0.0
−2.5
10
15
20
25
30
.fitted


sample
2
1
0
−1
−2
−1
0
1
2
theoretical


sqrt(abs(.stdresid))
1.5
1.0
0.5
10
15
20
25
30
.fitted


.stdresid
2
1
0
−1
0.0
0.1
0.2
0.3
0.4
0.5
.hat