Regression Models Course Project

Steven Zhang

21 Nov, 2016

1 - Synopsis

This report is about vehicles, aims to explore the relationship between miles per gallon(MPG) and some other variables. Particularly focused on the following questions:

- "Is an automatic or manual transmission better for MPG?"
- "Quantify the MPG difference between automatic and manual transmissions."

The original question description can be found here.

From my analysis, the manual transmission car has a significant higher mpg than automatic's. In addition, from the generalized additive model, a manual transmission car has a fuel efficiency of 3.47 (MPG) higher than that of automatic transmission car.

2 - Exploratory Analysis

2.1 - Data Source Overview

The source dataset is embedded in R environment, named "mtcars", which was extracted from the 1974 *Motor Trend US* magazine, and comprises fuel consumption and 10 aspects of automobile design and performance for 32 automobiles (1973–74 models). This dataset has 32 observations on 11 variables as followed.

- mpg - Miles/(US) gallon
- cyl - Number of cylinders
- disp - Displacement (cu.in.)
- hp - Gross horsepower
- drat - Rear axle ratio
- wt - Weight (1000 lbs)
- qsec - 1/4 mile time
- vs - V/S
- am - Transmission (0 = automatic, 1 = manual)
- gear - Number of forward gears
- carb - Number of carburetors

2.2 - data loading and preprocessing

We load the data and convert the *am* variable to factor format.

```
df <- mtcars
summary(glm(mpg ~ ., data = df))$coefficients[,4]
```

```
## (Intercept)         cyl        disp          hp        drat          wt
##  0.51812440  0.91608738  0.46348865  0.33495531  0.63527790  0.06325215
##        qsec          vs          am        gear        carb
##  0.27394127  0.88142347  0.23398971  0.66520643  0.81217871
```

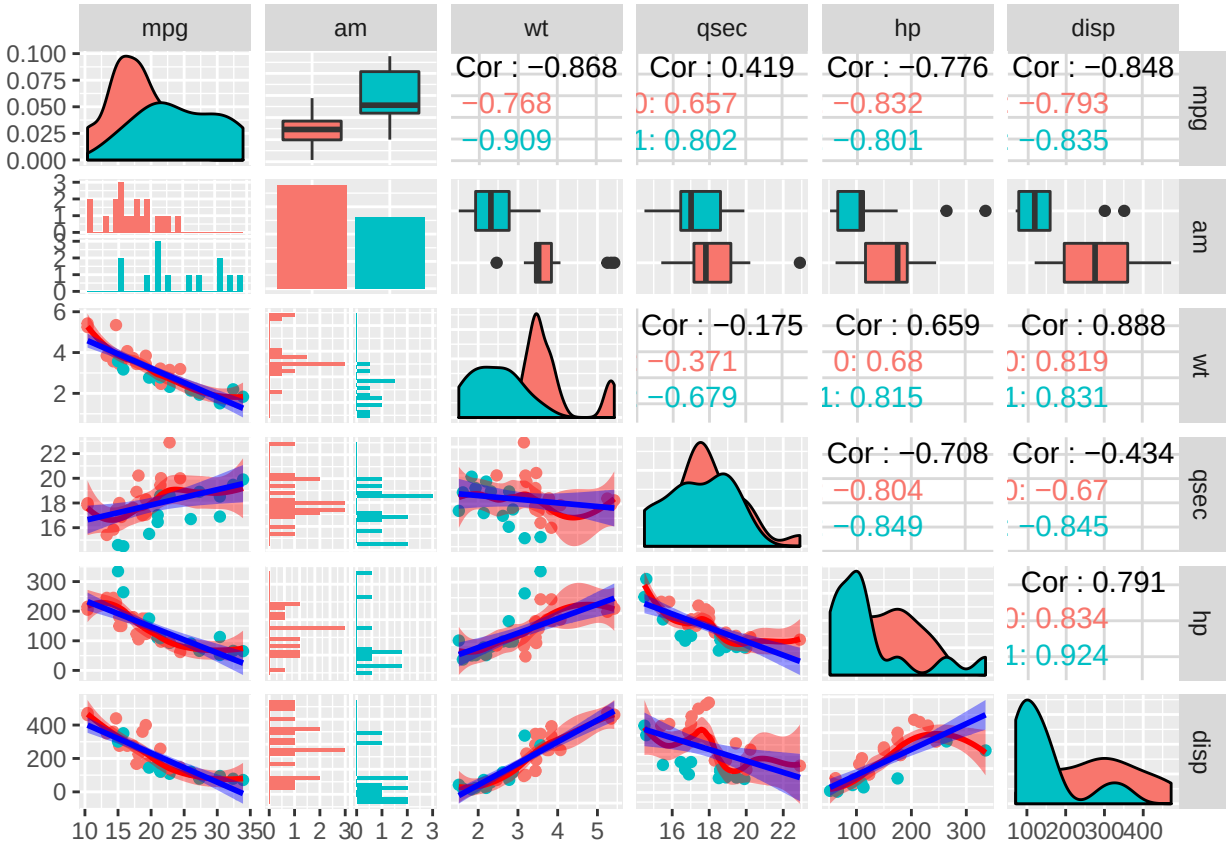
From a simple analysis, I choose 5 variables for next steps.

```
df <- select(df, c(mpg, am, wt, qsec, hp, disp))
df$am <- as.factor(df$am)
```

Here is the correlation of the variables. For the convenience of further analysis, I hold the *Transmission* (auto/manual) as colors.

```
my.ggpairs <- function(data, mapping, ...){
  p <- ggplot(data = data, mapping = mapping, bins = 30) +
    geom_point() +
    geom_smooth(method=loess, fill="red", color="red", ...) +
    geom_smooth(method=lm, fill="blue", color="blue", ...)
  p
}
ggpairs(df, lower = list(continuous = my.ggpairs), aes(color = am))
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



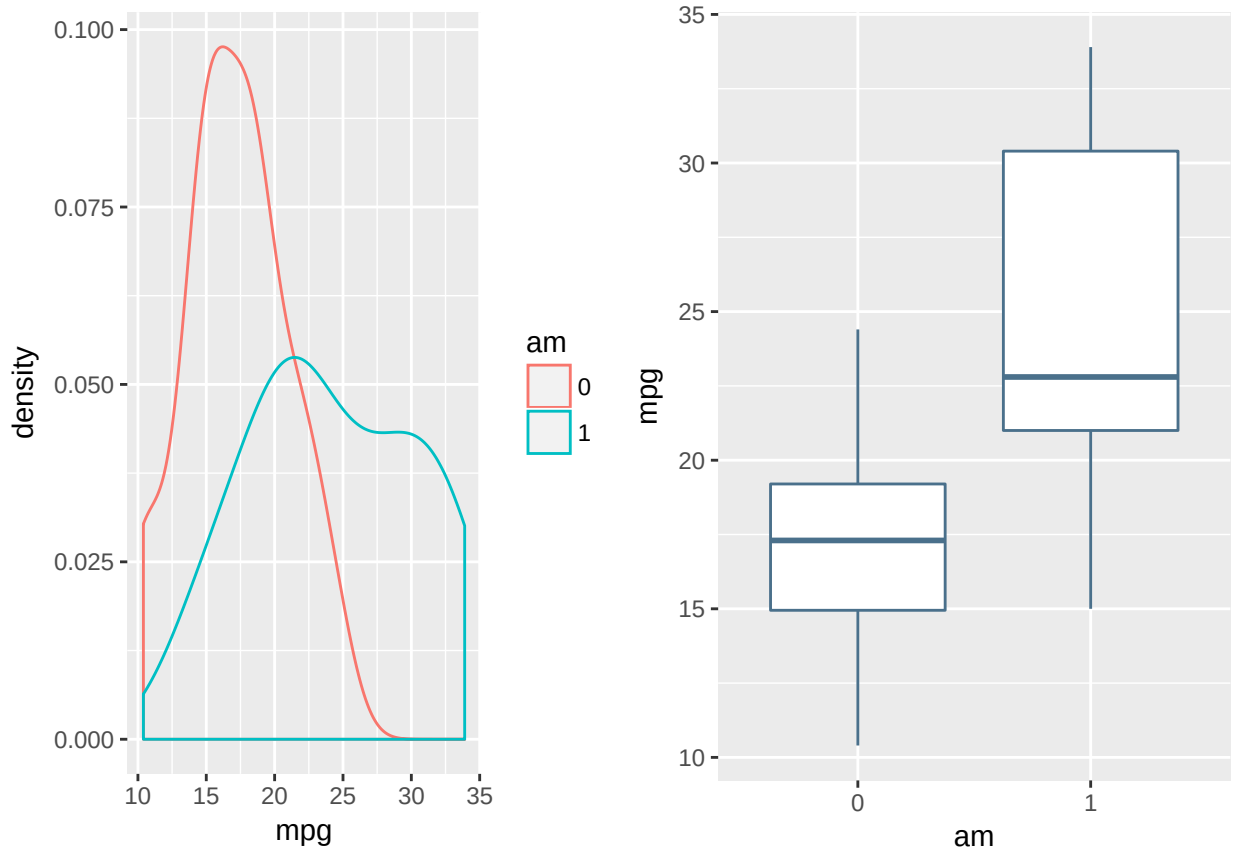
From the diagram above, we can see some clear relationships between mpg and other variables. More, we can see a clear difference different transmissions (red & blue).

3 - Analysis

3.1 - Question 1: Is an automatic or manual transmission better for MPG?

Now, we are going to explore the relationship between automatic and manual transmission.

```
density.plot <- ggplot(df, aes(mpg, color = am)) + geom_density()
box.plot <- ggplot(df, aes(am, mpg)) + geom_boxplot(col = "skyblue4")
grid.arrange(density.plot, box.plot, ncol = 2, nrow = 1)
```



We can see from the plot above that there are obvious correlation between mpg and transmission. However, we should run a hypothesis test to confirm it.

```
group.manual <- df[df$am == 1,]
group.auto <- df[df$am == 0,]
t.test(group.auto$mpg, group.manual$mpg, paired = FALSE, var.equal = FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  group.auto$mpg and group.manual$mpg
## t = -3.7671, df = 18.332, p-value = 0.001374
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -11.280194  -3.209684
## sample estimates:
## mean of x mean of y
```

```
##  17.14737  24.39231
```

From this T test, we can see the automatic transmission has a significant (p-value < 0.0015) less mpg value than manual's.

3.2 - Question 2: Quantify the MPG difference between automatic and manual transmissions

Based on the analysis above, we need to quantify the coefficients. I build three linear regression models with 1, 5 and 9 predictors and use the anova function to compare them.

```
fit1 <- lm(mpg ~ am, data = df)
fit2 <- lm(mpg ~ ., data = df)
mtcars$am <- as.factor(mtcars$am)
fit3 <- lm(mpg ~ ., data = mtcars)
anova(fit1, fit2, fit3)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ am + wt + qsec + hp + disp
## Model 3: mpg ~ cyl + disp + hp + drat + wt + qsec + vs + am + gear + carb
##   Res.Df    RSS Df Sum of Sq       F   Pr(>F)
## 1     30 720.90
## 2     26 153.44  4    567.46 20.1984 5.92e-07 ***
## 3     21 147.49  5      5.94  0.1692   0.9711
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From this comparance, we choose model 2 to do further test.

```
summary(fit2)
```

```
##
## Call:
## lm(formula = mpg ~ ., data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.5399 -1.7398 -0.3196  1.1676  4.5534
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 14.36190    9.74079   1.474  0.15238
## am1          3.47045    1.48578   2.336  0.02749 *
## wt          -4.08433    1.19410  -3.420  0.00208 **
## qsec         1.00690    0.47543   2.118  0.04391 *
## hp          -0.02117    0.01450  -1.460  0.15639
## disp         0.01124    0.01060   1.060  0.29897
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.429 on 26 degrees of freedom
## Multiple R-squared:  0.8637, Adjusted R-squared:  0.8375
## F-statistic: 32.96 on 5 and 26 DF,  p-value: 1.844e-10
```

Based on this generalized additive model, we can say that a manual transmission car has a fuel efficiency of 3.47 (MPG) higher than that of a automatic transmission car.

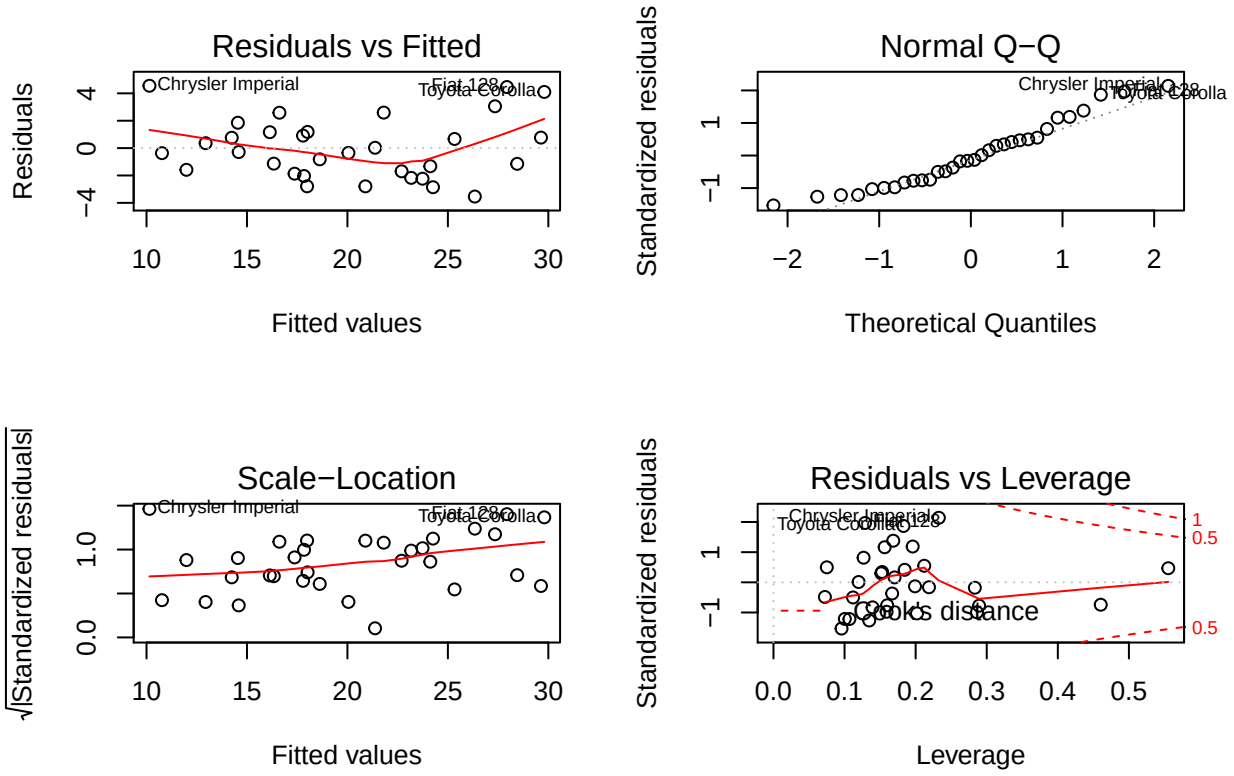
4 - Appendix

- [1] Henderson and Velleman (1981), Building multiple regression models interactively. *Biometrics*, 37, 391–411.

4.1 - Additional Plots

Here we list the plots of the regression.

```
par(mfrow = c(2, 2))
plot(fit2)
```



4.2 - Hardware & Software Environment

```
## R version 3.2.5 (2016-04-14)
## Platform: x86_64-w64-mingw32/x64 (64-bit)
## Running under: Windows 10 x64 (build 10586)
##
## locale:
## [1] LC_COLLATE=Chinese (Simplified)_China.936
```

```
## [2] LC_CTYPE=Chinese (Simplified)_China.936
## [3] LC_MONETARY=Chinese (Simplified)_China.936
## [4] LC_NUMERIC=C
## [5] LC_TIME=Chinese (Simplified)_China.936
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
## [1] gridExtra_2.2.1 GGally_1.3.0    ggplot2_2.2.0   dplyr_0.5.0
##
## loaded via a namespace (and not attached):
##  [1] Rcpp_0.12.7        knitr_1.14         magrittr_1.5
##  [4] munsell_0.4.3      colorspace_1.2-6   R6_2.2.0
##  [7] stringr_1.1.0      plyr_1.8.4         tools_3.2.5
## [10] grid_3.2.5         gtable_0.2.0       DBI_0.5-1
## [13] htmltools_0.3.5    yaml_2.1.13        lazyeval_0.2.0
## [16] assertthat_0.1     digest_0.6.10      tibble_1.2
## [19] reshape2_1.4.1     RColorBrewer_1.1-2 formatR_1.4
## [22] evaluate_0.9       rmarkdown_1.0      labeling_0.3
## [25] stringi_1.1.2      scales_0.4.1       reshape_0.8.5
```