# Regression Models Project — MPG Analysis Using Mtcars Data Set

*Feng Hong*

*5/9/2017*

# Summary

This analysis uses a data set of a collection of cars to explore the relationship between a set of car features and MPG. Specifically, it answers 2 questions concerning the influence of transmission type on MPG. Data are first divided into a subset of numeric variables and another of factor variables for rudimentary exploration. Subsequent feature selection is utilizes a stepwise algorithm based on AIC. Regression model comparison, coefficient interpretation and potential problems are presented in the third part of this analysis.

```
library(rmarkdown)
library(knitr)
library(dplyr)
library(ggplot2)
library(magrittr)
library(stargazer)
library(ggfortify)
```

# Exploratory Analysis

## Correlation Matrix of Numeric Variables

Check the correlation of each pair of the numeric variables. But keep in mind that variables that have a high correlation with MPG do not necessarily cause a high or low MPG.

```
data("mtcars")
kable(head(mtcars))
```

| | mpg | cyl | disp | hp | drat | wt | qsec | vs | am | gear | carb |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Mazda RX4 | 21.0 | 6 | 160 | 110 | 3.90 | 2.620 | 16.46 | 0 | 1 | 4 | 4 |
| Mazda RX4 Wag | 21.0 | 6 | 160 | 110 | 3.90 | 2.875 | 17.02 | 0 | 1 | 4 | 4 |
| Datsun 710 | 22.8 | 4 | 108 | 93 | 3.85 | 2.320 | 18.61 | 1 | 1 | 4 | 1 |
| Hornet 4 Drive | 21.4 | 6 | 258 | 110 | 3.08 | 3.215 | 19.44 | 1 | 0 | 3 | 1 |
| Hornet Sportabout | 18.7 | 8 | 360 | 175 | 3.15 | 3.440 | 17.02 | 0 | 0 | 3 | 2 |
| Valiant | 18.1 | 6 | 225 | 105 | 2.76 | 3.460 | 20.22 | 1 | 0 | 3 | 1 |

```
sapply(mtcars, class)
```

```
##       mpg       cyl      disp        hp      drat        wt      qsec 
## "numeric" "numeric" "numeric" "numeric" "numeric" "numeric" "numeric" 
##        vs        am      gear      carb 
## "numeric" "numeric" "numeric" "numeric"
```

```
mtcars[ , c('cyl', 'vs', 'am')] %<>% lapply(function(x) as.factor(x))
# Should not use sapply().
mtcars_numeric <- select_if(mtcars, is.numeric)
kable(cor(mtcars_numeric))
```

| | mpg | disp | hp | drat | wt | qsec | gear | carb |
|---|---|---|---|---|---|---|---|---|
| mpg | 1.0000000 | -0.8475514 | -0.7761684 | 0.6811719 | -0.8676594 | 0.4186840 | 0.4802848 | -0.5509251 |
| disp | -0.8475514 | 1.0000000 | 0.7909486 | -0.7102139 | 0.8879799 | -0.4336979 | -0.5555692 | 0.3949769 |
| hp | -0.7761684 | 0.7909486 | 1.0000000 | -0.4487591 | 0.6587479 | -0.7082234 | -0.1257043 | 0.7498125 |
| drat | 0.6811719 | -0.7102139 | -0.4487591 | 1.0000000 | -0.7124406 | 0.0912048 | 0.6996101 | -0.0907898 |
| wt | -0.8676594 | 0.8879799 | 0.6587479 | -0.7124406 | 1.0000000 | -0.1747159 | -0.5832870 | 0.4276059 |
| qsec | 0.4186840 | -0.4336979 | -0.7082234 | 0.0912048 | -0.1747159 | 1.0000000 | -0.2126822 | -0.6562492 |
| gear | 0.4802848 | -0.5555692 | -0.1257043 | 0.6996101 | -0.5832870 | -0.2126822 | 1.0000000 | 0.2740728 |
| carb | -0.5509251 | 0.3949769 | 0.7498125 | -0.0907898 | 0.4276059 | -0.6562492 | 0.2740728 | 1.0000000 |

## Factor Variable Exploration

Violin plots are created to show the MPG distribution dependent on transmission types (am). Furthermore, 2 addtional violin plots are drawn to reveal the abovementioned distribution when either engine type (vs) or number of cylinders (cyl) is taken into consideration.
(Please see “Violin Plot of MPG on Transmission Type (1) - (3)” in Appendix.)

```
mtcars_factor <- select(mtcars, mpg, cyl, vs, am)
levels(mtcars_factor$vs) <- c('V-shape', 'Straight')
levels(mtcars_factor$am) <- c('Automatic', 'Manual')

theme_format <- theme(plot.margin = unit(c(1,1,1,1), "cm"),
                      plot.title = element_text(hjust = 0.5, size = 16),
                      axis.text = element_text(size = 12),
                      axis.title = element_text(size = 14))

am_only <- ggplot(mtcars_factor, aes(x = am, y = mpg)) +
           geom_violin(trim = FALSE) + geom_boxplot(width = 0.2) +
           labs(x = "Transmission Type", y = "MPG", title = "Violin Plot of MPG on Transmission Type (1)") +
           theme_format

am_vs <- ggplot(mtcars_factor, aes(x = am, y = mpg, fill = vs)) + geom_violin(trim = FALSE) +
         labs(x = "Transmission Type", y = "MPG", title = "Violin Plot of MPG on Transmission Type (2)") +
         theme_format + scale_fill_manual(values = c("lightsteelblue1", "mistyrose"), name = "Engine Type")


am_cyl <- ggplot(mtcars_factor, aes(x = am, y = mpg, fill = cyl)) + geom_violin(trim = FALSE) +
          labs(x = "Transmission Type", y = "MPG", title = "Violin Plot of MPG on Transmission Type (3)") +
          theme_format + scale_fill_manual(values = c("thistle1", "lightsteelblue3", "navajowhite"),
                                           name = "Number of Cylinders")

am_only
am_vs
am_cyl
```

# Regression Models

## Model Selection

In this analysis, only OLS models will be built and compared.
First, select a formula-based model by AIC. This gives us a basic model that yields the best performance when interaction terms are not taken into consideration. Then we run two more regressions of MPG on the variables selected, but with different variable interactions added.

```
basic_model <- step(lm(mpg~., data = mtcars), trace = 0)
all.vars(formula(basic_model))
```

[1] "mpg" "wt" "qsec" "am"

```
inter_am_qsec <- lm(mpg ~ am * qsec + wt, data = mtcars)
inter_am_wt <- update(inter_am_qsec, mpg ~ am * wt + qsec)

stargazer(basic_model, inter_am_qsec, inter_am_wt,
          title = "Regression Results", align = TRUE, type = 'html', dep.var.labels = "MPG",
          covariate.labels = c('Weight', 'Manual Transmission * 1/4 Mile Time', '1/4 Mile Time',
                               'Manual Transmission * Weight', 'Manual Transmission'))
```

**Regression Results**

| | *Dependent variable:* | | |
|---|---|---|---|
| | MPG | | |
| | (1) | (2) | (3) |
| Weight | -3.917*** | -3.777*** | -2.937*** |
| | (0.711) | (0.671) | (0.666) |
| Manual Transmission * 1/4 Mile Time | | 1.060** | |
| | | (0.487) | |
| 1/4 Mile Time | 1.226*** | 0.817** | 1.017*** |
| | (0.289) | (0.330) | (0.252) |
| Manual Transmission * Weight | | | -4.141*** |
| | | | (1.197) |
| Manual Transmission | 2.936** | -15.614* | 14.079*** |
| | (1.411) | (8.624) | (3.435) |
| Constant | 9.618 | 16.529** | 9.723 |
| | (6.960) | (7.267) | (5.899) |
| Observations | 32 | 32 | 32 |
| $R^2$ | 0.850 | 0.872 | 0.896 |
| Adjusted $R^2$ | 0.834 | 0.853 | 0.880 |
| Residual Std. Error | 2.459 (df = 28) | 2.309 (df = 27) | 2.084 (df = 27) |
| F Statistic | 52.750*** (df = 3; 28) | 46.029*** (df = 4; 27) | 58.061*** (df = 4; 27) |
| *Note:* | | | *p<0.1;* ***p<0.05;*** p<0.01 |

Based on the above result, the third model (with interaction between transmission type and weight) is selected for MPG prediction because it has the largest F-stat as well as largest adjusted R-squared. Also we see that the regressors in this model all have significant coefficients.

# Residual Plotting

Four types of residual plots are created to show potential pitfalls in the model. But judging from the plots, the model does not appear to have significant problems such as heteroskedasticity or high leverage points.
(Please see "Residual Plots" in Appendix.)

```
autoplot(inter_am_wt, label.size = 3)
```

# Coefficient Interpretation & Potential Problems

The model fomula is:

$$\hat{MPG} = 9.723 + 14.079 \times Manual\_Transmission - 2.937 \times Weight + 1.017 \times (1/4\_Mile\_Time) - 4.141 \times Manual\_Transmission \times Weig$$

Holding other things constant, the predicted MPG for a car with automatic transmission is given by:

$$\hat{MPG}_{automatic} = 9.723 - 2.937 \times Weight + 1.017 \times (1/4\_Mile\_Time)$$

while that for a car with manual transmission is:

$$\hat{MPG}_{manual} = 23.802 - 7.078 \times Weight + 1.017 \times (1/4\_Mile\_Time)$$

**Question 1: Quantify the MPG difference between automatic and manual transmissions.**

The difference is measured by:
$\hat{MPG}_{automatic} - \hat{MPG}_{manual} = -14.079 + 4.141 \times Weight$

**Question 2: Is an automatic or manual transmission better for MPG?**

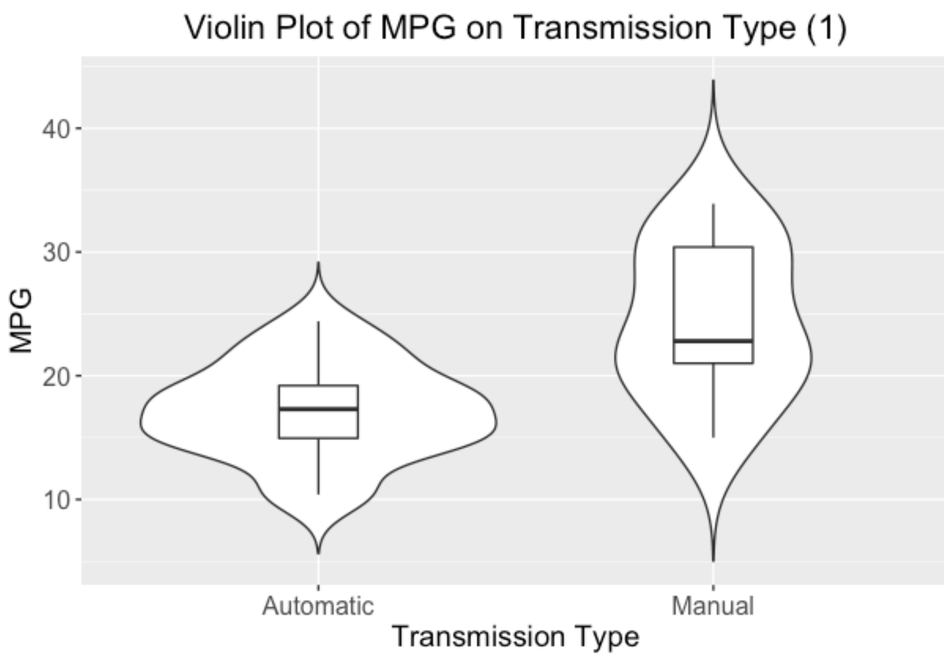
That depends. When a car is no heavier than $(14.079/4.141 \times 1000 \approx)$ 3,399.90 lbs, manual transmission has an edge in terms of MPG. If its weight exceeds 3,399.90 lbs, than automatic transmission is better for MPG.

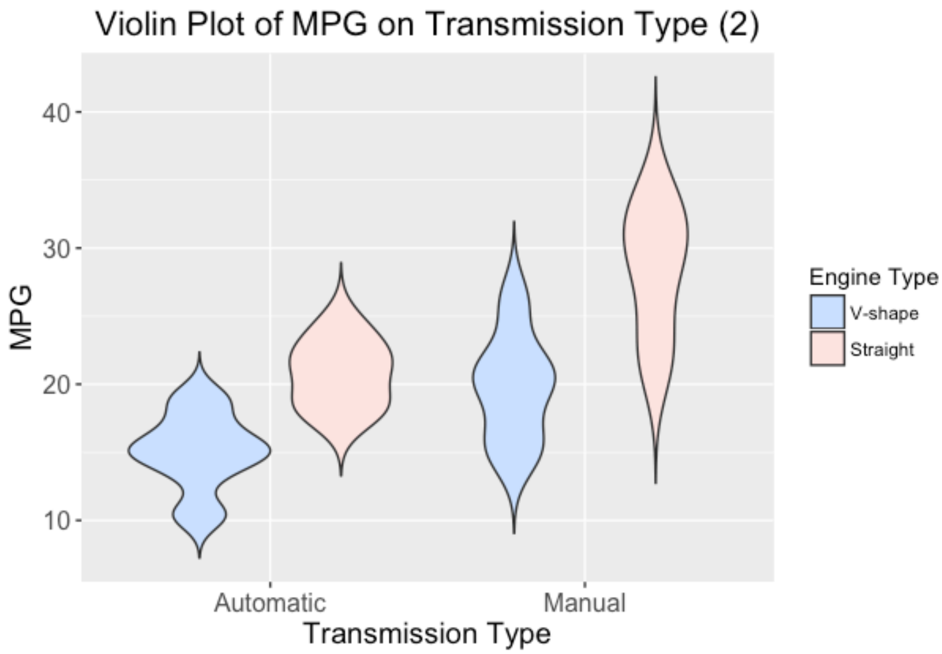
**Question 3: How about the uncertainties?**

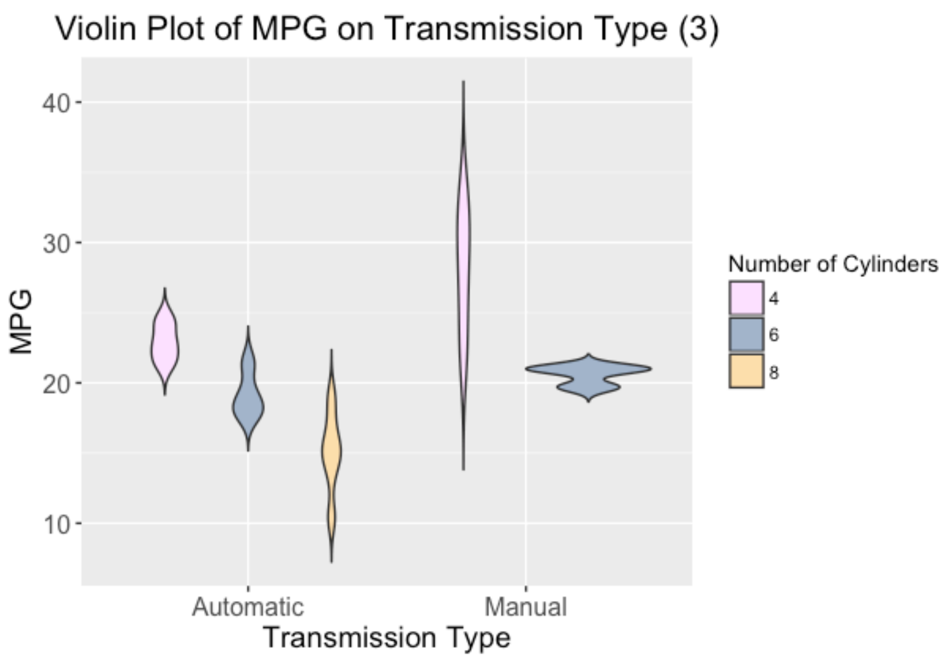
In this analysis, a potentially fatal problem is that the number of observations is way too small compared to the number of features. A small change in any observation of our data set can have a great influence in our model selection and robustness. Because of the lack of observations, we cannot determine with certainty whether the features not included in this analysis are actually influential or not.

# Appendix

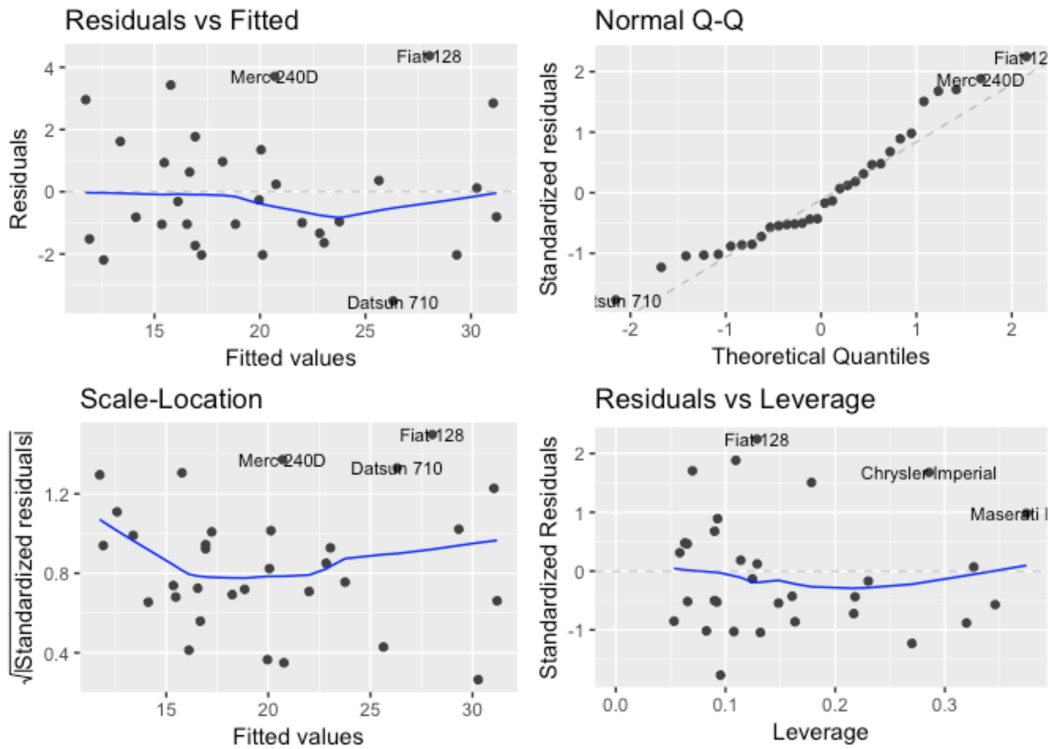
**MPG on Transmission:**






**Residual Plots:**


Residuals vs Fitted
Residuals
Fitted values
Fiat 128
Merc 240D
Datsun 710
Normal Q-Q
Standardized residuals
Theoretical Quantiles
Scale-Location
√|Standardized residuals|
Residuals vs Leverage
Standardized Residuals
Leverage
Chrysler Imperial