Regression Models Course Project - Motor Trend Data Analysis Report

Khoo Jia Jun

Executive Summary

In this report, we analyze `mtcars` data set and explore relationship between a set of variables and miles per gallon (MPG). Data was extracted from 1974 *Motor Trend* US magazine and contains fuel consumption and 10 aspects of automobile design and performance for 32 automobiles (1973 and 1974 models). Regression models and exploratory data analysis are used to explore how **automatic** (am = 0) and **manual** (am = 1) transmissions features affect **MPG** feature. T-test done on performance difference between cars with automatic and manual transmission revealed that there is about 7 MPG more for cars with manual transmission than those with automatic transmission. We then fit several linear regression models and select the one with highest adjusted R-squared value. By keeping both weight and 1/4 mile time constant, manual transmitted cars have on average of [14.079 + (-4.141) * weight] more MPG than automatic transmitted cars. Hence, lighter manual transmission cars have higher MPG values than heavier automatic transmission cars.

Exploratory Data Analysis

Load `mtcars` data set and update selected variables from `numeric` class to `factor` class.

```
library(ggplot2)
data(mtcars)
mtcars[1:3, ] # Sample Data
```

```
##                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
```

```
dim(mtcars)
```

```
## [1] 32 11
```

```
mtcars$cyl <- as.factor(mtcars$cyl)
mtcars$vs <- as.factor(mtcars$vs)
mtcars$am <- factor(mtcars$am)
mtcars$gear <- factor(mtcars$gear)
mtcars$carb <- factor(mtcars$carb)
attach(mtcars)
```

```
## The following object is masked from package:ggplot2:
##
##     mpg
```

Perform exploratory data analysis on data. See **Appendix: Figures** section for plots. Box plot revealed that manual transmission generally yields higher values of MPG. Pair graph revealed that variables "wt", "disp", "cyl", "hp" are highly correlated.

Inference

Assume MPG variable follows a normal distribution. We take null hypothesis that MPG of both automatic and manual transmissions come from same population. We use two sample T-test to show it.

```
result <- t.test(mpg ~ am)
result$p.value
```

```
## [1] 0.001373638
```

```
result$estimate
```

```
## mean in group 0 mean in group 1
##        17.14737        24.39231
```

P-value is 0.00137. Hence, we reject null hypothesis. This implies automatic and manual transmissions are from different populations. Mean of MPG of manual transmitted cars is about 7 more than those of automatic transmitted cars.

Regression Analysis

We fit the full model as follows:

```
fullModel <- lm(mpg ~ ., data=mtcars)
summary(fullModel) # results hidden
```

This model has residual standard error of 2.833 on 15 degrees of freedom. Adjusted R-squared value is 0.779, which implies model can explain about 78% of the variance of MPG variable. However, no coefficients are significant at 0.05 significant level.
Backward selection is used to select statistically significant variables.

```
stepModel <- step(fullModel, k=log(nrow(mtcars)))
summary(stepModel) # results hidden
```

Model "mpg ~ wt + qsec + am" has residual standard error as 2.459 on 28 degrees of freedom. Adjusted R-squared value is 0.8336, which implies model can account for about 83% of variance of MPG variable. All coefficients are significant at 0.05 significant level.

See **Appendix: Figures** section for the plots. Scatter plot revealed there might be an interaction between "wt" variable and "am" variable, since automatic cars tend to be heavier than manual cars. Thus, the following model including the interaction term is used:

```
amIntWtModel<-lm(mpg ~ wt + qsec + am + wt:am, data=mtcars)
summary(amIntWtModel) # results hidden
```

This model has residual standard error as 2.084 on 27 degrees of freedom. Adjusted R-squared value is 0.8804, which implies model can explain about 88% of the variance of MPG variable. All coefficients are significant at 0.05 significant level.

Fit model using MPG as outcome variable and Transmission as predictor variable:

```
amModel<-lm(mpg ~ am, data=mtcars)
summary(amModel) # results hidden
```

On average, an automatic transmission car has 17.147 MPG. A manual transmission car has 7.245 more MPG. This model has residual standard error of 4.902 on 30 degrees of freedom. Adjusted R-squared value is 0.3385, which means model can explain about 34% of the variance of MPG variable. A low Adjusted R-squared value indicates other variables need to be added to the model.

Finally, we select the final model.

```
anova(amModel, stepModel, fullModel, amIntWtModel)
confint(amIntWtModel) # results hidden
```

We select model with highest Adjusted R-squared value, "mpg ~ wt + qsec + am + wt:am".

```
summary(amIntWtModel)$coef
```

```
##              Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)  9.723053  5.8990407  1.648243 0.1108925394
## wt          -2.936531  0.6660253 -4.409038 0.0001488947
## qsec         1.016974  0.2520152  4.035366 0.0004030165
## am1         14.079428  3.4352512  4.098515 0.0003408693
## wt:am1      -4.141376  1.1968119 -3.460340 0.0018085763
```

Result shows that when "wt" (weight lb/1000) and "qsec" (1/4 mile time) remain constant, cars with manual transmission add 14.079 + (-4.141) * wt more MPG (miles per gallon) on average than cars with automatic transmission. This means a manual transmitted car weighing 2000lbs have 5.797 more MPG than an automatic transmitted car with similar weight and 1/4 mile time.

Residual Analysis and Diagnostics

Refer to **Appendix: Figures** section for plots. According to residual plots, we can verify the following assumptions:
1. Residuals vs. Fitted plot shows no consistent pattern, supporting the validity of independence assumption.
2. Normal Q-Q plot indicates residuals are normally distributed because the points lie close to the line.
3. Scale-Location plot confirms constant variance assumption, as points are randomly distributed.
4. Residuals vs. Leverage shows that no outliers are present as all values fell within the 0.5 bands.

As for the measure of how much an observation has effected the estimate of a regression coefficient Dfbetas, we arrived at following result:

```
sum((abs(dfbetas(amIntWtModel)))>1)
```

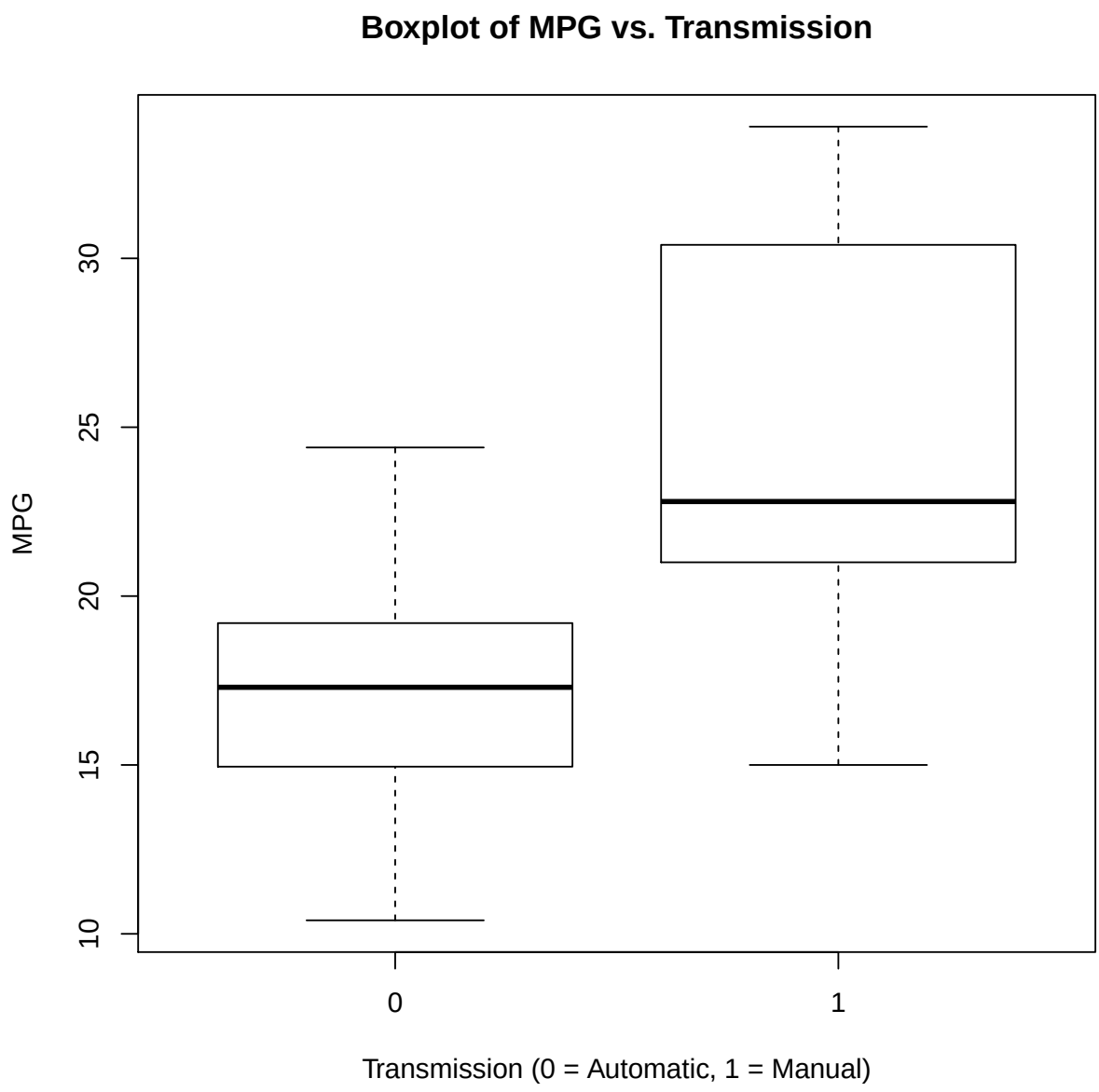
```
## [1] 0
```

We conclude that the above analysis meet all basic assumptions of linear regression and answers our questions.

Appendix: Figures

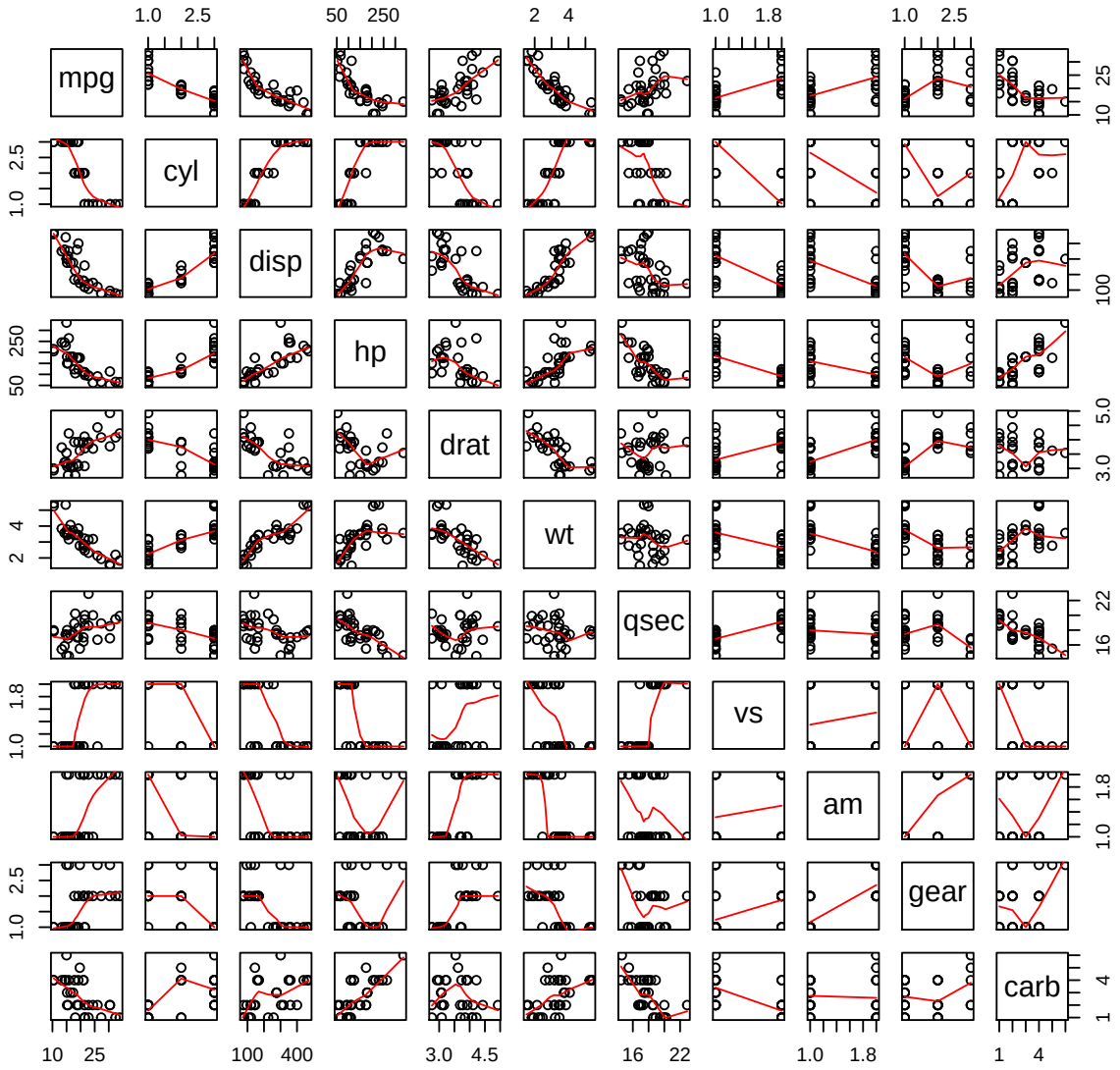
1. Boxplot of MPG vs. Transmission

```r
boxplot(mpg ~ am, xlab="Transmission (0 = Automatic, 1 = Manual)", ylab="MPG",
        main="Boxplot of MPG vs. Transmission")
```



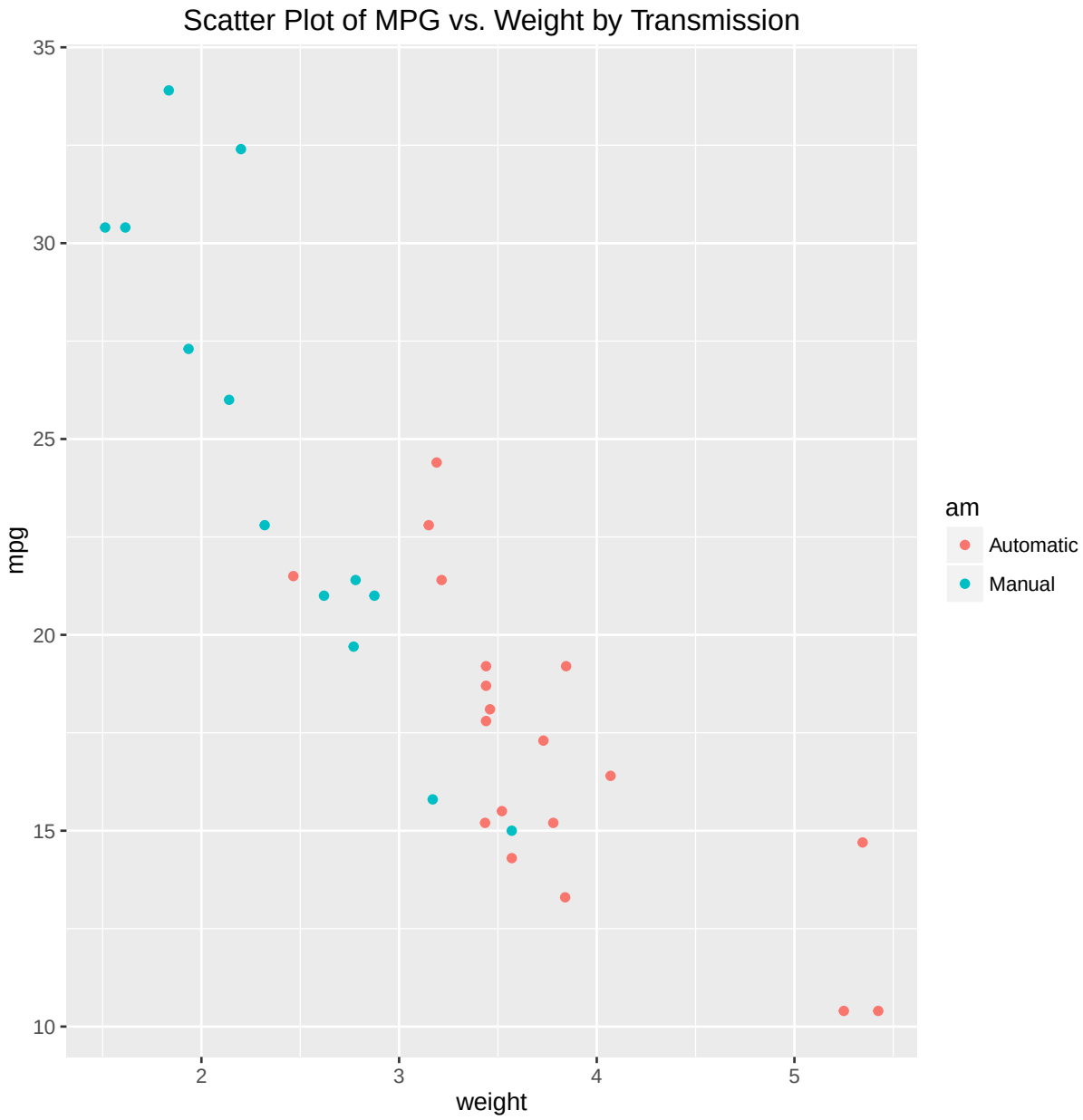
2. Pair Graph of Motor Trend Car Road Tests

```r
pairs(mtcars, panel=panel.smooth, main="Pair Graph of Motor Trend Car Road Tests")
```



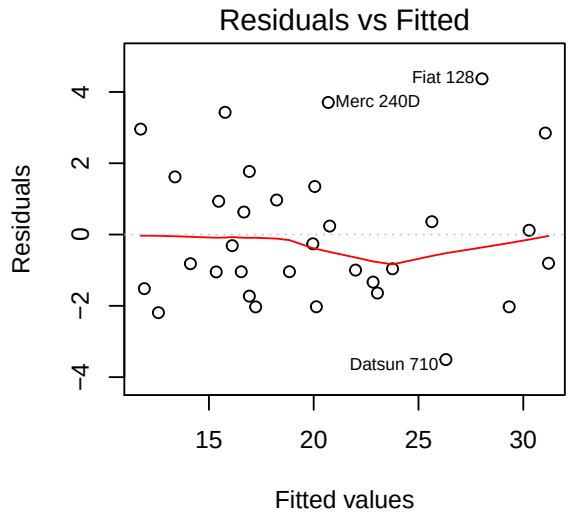
3. Scatter Plot of MPG vs. Weight by Transmission

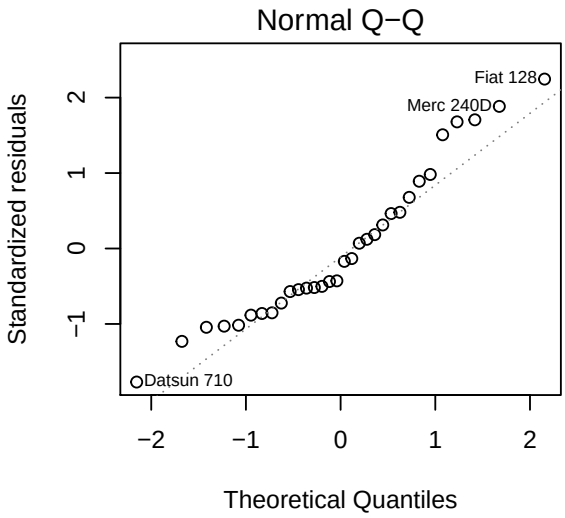
```
ggplot(mtcars, aes(x=wt, y=mpg, group=am, color=am, height=3, width=3)) + geom_point() +
scale_colour_discrete(labels=c("Automatic", "Manual")) +
xlab("weight") + ggtitle("Scatter Plot of MPG vs. Weight by Transmission")
```

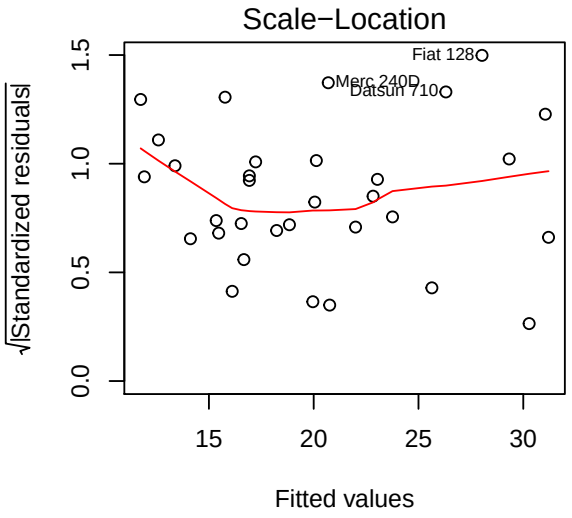


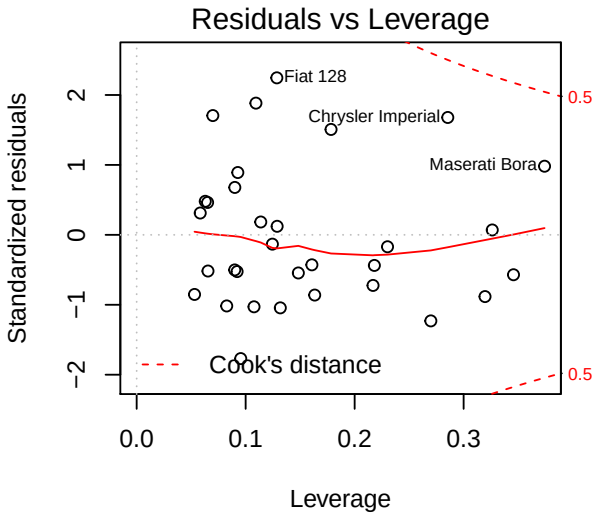
4. Residual Plots

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


Residuals vs Fitted
Residuals
Fitted values
Fiat 128
Merc 240D
Datsun 710


Normal Q−Q
Standardized residuals
Theoretical Quantiles
Fiat 128
Merc 240D
Datsun 710


Scale−Location
√|Standardized residuals|
Fitted values
Fiat 128
Merc 240D
Datsun 710


Residuals vs Leverage
Standardized residuals
Leverage
Fiat 128
Chrysler Imperial
Maserati Bora
Cook's distance
0.5
0.5