Regression Model Project

Kun Zhu

Executive Summary

This report represents a empirical study on the relationship between miles per gallon (MPG) and type of transmission, i.e., automatic and manual. The studied dataset includes fuel consumption and 10 aspects of automobile design and performance for 32 automobiles.

By intuition, a manual transmission car outperforms a car with automatic transmission by offering better MPG. A naive regression study agrees with the above claim by suggesting the manual transmission car offers about 7 extra MPG. However, the detailed regression study leads to a different conclusion that weight is a cofounder that can also impact the MPG. Specially, the manual transmission can offer MPG benefit when the cars are light, e.g., below 3000 lbs. However, as the car gets heavier, weight becomes a more important factor than tranmission in terms of MPG benefit.

Exploratory data analysis

```
head(mtcars, 2)
```

```
##                mpg cyl disp  hp drat    wt  qsec vs am gear carb
## Mazda RX4       21   6  160 110  3.9 2.620 16.46  0  1    4    4
## Mazda RX4 Wag   21   6  160 110  3.9 2.875 17.02  0  1    4    4
```

Boxplot using ggplot package can be found in the Appendix. A t-test is performed with a null hypothesis that automatic and manual transmission cars offer the same MPG. The underlying assumption is that MPG is normally distributed.

```
meanDiff <- t.test(mtcars$mpg ~ mtcars$am)
meanDiff$p.value; meanDiff$estimate
```

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

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

As the p-value is smaller than 0.05, the null hypothesis above is rejected. Therefore, the inference study above suggests that cars with manual transmission in average offers extra, `24.39 - 17.15 = 7.14`, MPG than automatic tranmission cars.

Regression analysis

```
fitAm <- lm(mpg ~ am, data = mtcars)
summary(fitAm)$coefficients; summary(fitAm)$adj.r.squared
```

```
##              Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 17.147368   1.124603 15.247492 1.133983e-15
## am           7.244939   1.764422  4.106127 2.850207e-04

## [1] 0.3384589
```

For this model, the intercept is equal to the average MPG of automatic transmissions. Similarly, the sum of the intercept and slope coefficients is equal to the average MPG among manual transmissions. However, the adjusted r square suggests that this model can only explain 33% of the MPG variance.

The backward selection is used to select the model including statistically significant variables.

```
fitAll <- lm(mpg ~ ., data = mtcars)
stepModel <- step(fitAll, k=log(nrow(mtcars))); summary(stepModel)
```

The model with `mpg ~ wt + qsec + am` is chosen, as it returns the largest adjusted r square value, 0.8336. This means that the model can explain about 83% of the MPG variance. In addition, all coefficients are significant at 0.05 significant level.

Next, the interaction of the variable "wt", "qsec", "am" are also considered in the model. Specially, the model selection is made by examine the variance table of nested models, see details in Appendix. The study suggests the best option among the candidate models is:

```
fitAmIntWt <- lm(mpg ~ wt + qsec + am + am*wt, data=mtcars)
summary(fitAmIntWt)$coefficients[4:5, ]

##        Estimate Std. Error   t value     Pr(>|t|)
## am    14.079428   3.435251  4.098515 0.0003408693
## wt:am -4.141376   1.196812 -3.460340 0.0018085763
```

Manual transmission contributes to `14.079 - 4.141*wt` extra MPG on average than cars with automatic transmission, while hold "wt" (weight lb/1000) and "qsec" constant. The assumption that manual transmission car leads to advantage in MPG holds when the car is light. For example, a 1500 lb car with manual transmission can provide 7.87 extra MPG. However, the assumption no longer holds when the car is heavy. For example, a 3500 lb manual transmission car actually leads to 0.42 less MPG. The 95% confidence interval can be found in Appendix.

Residual Analysis and Diagnostics

The followings are shown in the residual plot, provided in Appendix:

- No correlation is observed from the Residuals/Fitted plot
- The QQ plot suggests that the residuals are normally distributed as the points reside along the line
- No patterns are observed from the Scale-Location.
- No outliers are observed from the Residuals/Leverage plot, as all values are within the 0.5 bands.

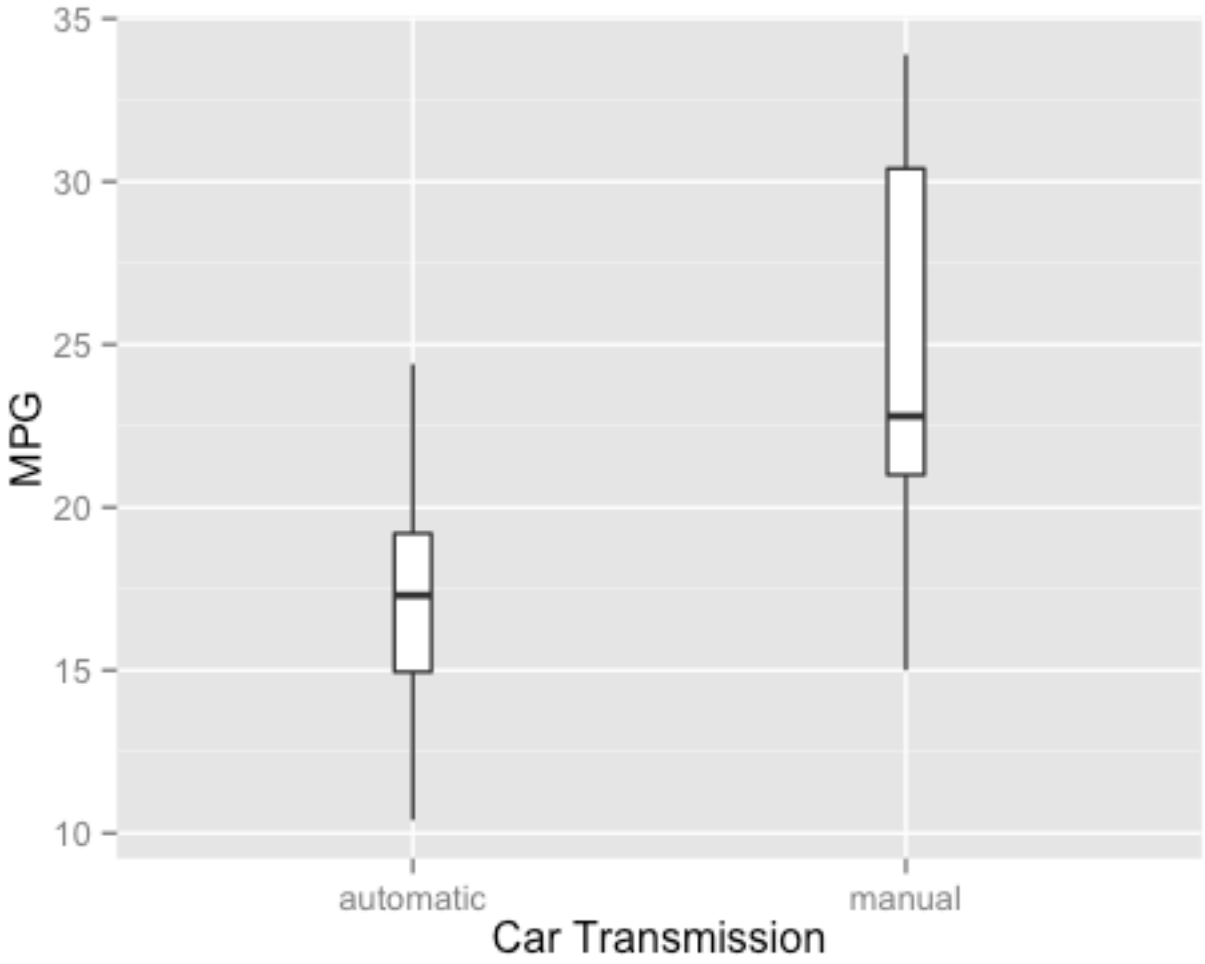
Appendix

Exploratory data analysis, boxplot

```
library(ggplot2)

## Warning: package 'ggplot2' was built under R version 3.1.3

mtcars$am <- as.factor(mtcars$am)
levels(mtcars$am) <- c("automatic", "manual")
p <- ggplot(mtcars, aes(x = am, y = mpg))
p + geom_boxplot(width = .1, fill = "white", outlier.color=NA) +
  labs(x="Car Transmission",y="MPG")
```



Regression analysis, nested model variance table

```
fit <- lm(mpg ~ wt + qsec + am, mtcars)
fitAmIntWt<- lm(mpg ~ wt + qsec + am + am*wt, data=mtcars)
fitAmIntQsec <- lm(mpg ~ wt + qsec + am  + am*wt + am*qsec, data=mtcars)
fitAmIntQsecWt <- lm(mpg ~ wt + qsec + am + am*wt + am*qsec + wt*qsec,
```

```
data=mtcars)
anova(fit, fitAmIntWt, fitAmIntQsec, fitAmIntQsecWt)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ wt + qsec + am
## Model 2: mpg ~ wt + qsec + am + am * wt
## Model 3: mpg ~ wt + qsec + am + am * wt + am * qsec
## Model 4: mpg ~ wt + qsec + am + am * wt + am * qsec + wt * qsec
##   Res.Df    RSS Df Sum of Sq       F  Pr(>F)
## 1     28 169.29
## 2     27 117.28  1    52.010 11.9312 0.00198 **
## 3     26 116.47  1     0.802  0.1841 0.67159
## 4     25 108.98  1     7.496  1.7196 0.20166
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

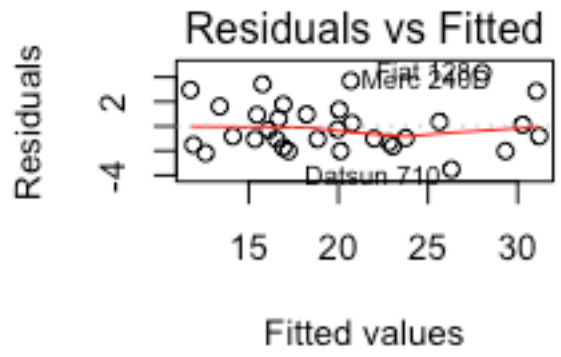
```
confint(fitAmIntWt)[4:5, ]
```

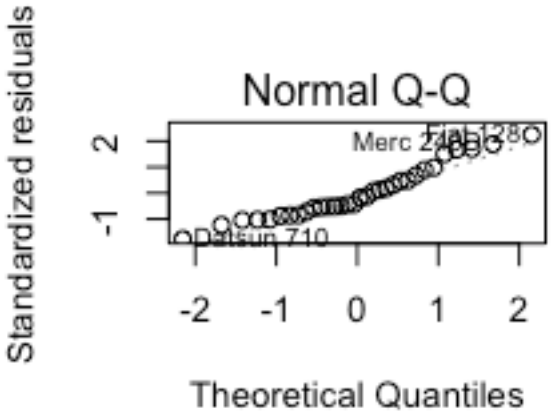
```
##                 2.5 %    97.5 %
## ammanual     7.030875 21.127981
## wt:ammanual -6.597032 -1.685721
```

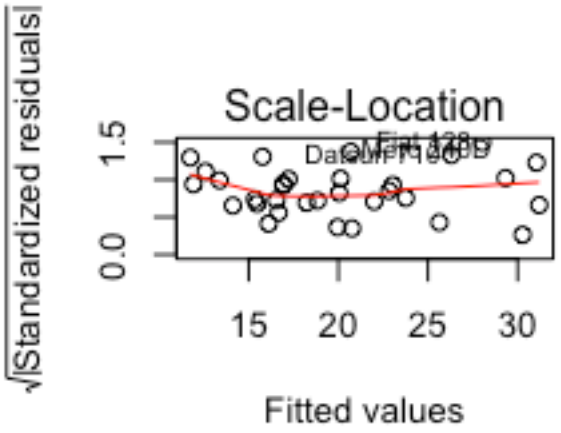
The p-value of model `fitAmIntWt` is smaller than 0.05 while p-values for `fitAmIntQsec` and `fitAmIntQsecWt` are greater than 0.05. This clearly shows that it is necessary to add term `am*wt` to the model while other interaction terms should be avoided.

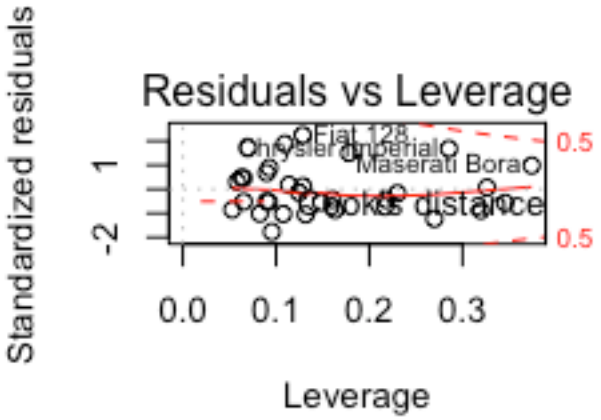
Residual Analysis and Diagnostics, residual plot

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


Residuals vs Fitted
Residuals
Fitted values
Merc 240D
Datsun 710
15
20
25
30
-4
2


Normal Q-Q
Standardized residuals
Theoretical Quantiles
Fiat 128
Datsun 710
-2
-1
0
1
2


Scale-Location
√|Standardized residuals|
Fitted values
Datsun 710
15
20
25
30
0.0
1.5


Residuals vs Leverage
Standardized residuals
Leverage
Fiat 128
Chrysler Imperial
Maserati Bora
Cook's distance
0.5
0.5
0.0
0.1
0.2
0.3
-2
1