Transmission System vs Fuel Efficiency

Simin Talebpour

Executive Summary

In this project I want to analyze **mtcars** dataset(a collection of cars), to explore the relationship between a set of variables and miles per gallon (MPG). I'm particularly interested in the following two questions:
-Is an automatic or manual transmission better for MPG?
-Quantify the MPG difference between automatic and manual transmissions?

Exploratory Data Analyses

```
## 'data.frame':    32 obs. of  11 variables:
##  $ mpg : num  21 21 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 ...
##  $ cyl : num  6 6 4 6 8 6 8 4 4 6 ...
##  $ disp: num  160 160 108 258 360 ...
##  $ hp  : num  110 110 93 110 175 105 245 62 95 123 ...
##  $ drat: num  3.9 3.9 3.85 3.08 3.15 2.76 3.21 3.69 3.92 3.92 ...
##  $ wt  : num  2.62 2.88 2.32 3.21 3.44 ...
##  $ qsec: num  16.5 17 18.6 19.4 17 ...
##  $ vs  : num  0 0 1 1 0 1 0 1 1 1 ...
##  $ am  : num  1 1 1 0 0 0 0 0 0 0 ...
##  $ gear: num  4 4 4 3 3 3 3 4 4 4 ...
##  $ carb: num  4 4 1 1 2 1 4 2 2 4 ...
```

The mtcars dataset consists of 32 observations of 11 (numeric) variables.
Looking at the box plot(Fig-1), I can say that manual cars are more gas efficient. I can verify that with a hypothesis test of simple linear regression.

Inference in Regression

I want to know if there is significant evidence that 'mph' depends on transmission system('am' variable)? Applying a simple regression, answers this question. Regression conducts a hypothesis test on the slope of the regression line using t-test methods to test the following hypothesis:
- Is the slope(coefficient for 'am') significantly different from zero?

```
fit <- lm(mpg ~ factor(am), data = mtcars)
summary(fit)$coef
```

```
##              Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 17.147368   1.124603 15.247492 1.133983e-15
## factor(am)1  7.244939   1.764422  4.106127 2.850207e-04
```

The P-value of 2.850207e-04 is significant(smaller than 0.05), so I reject the null hypothesis meaning that slope in my regression model is not zero, therefore it's statistically significant, so the result shows that the transmission system effects fuel consumption.

Model Selection

I want to find a model that includes all the important variables to predict the outcome. First I fit a model that includes all the variables.

```
fit_all <- lm(mpg ~ cyl+disp+hp+drat+wt+qsec+factor(vs)+factor(am)+gear+carb, data = mtcars)
```

Looking at the results in Table_1, I find none of the variables statistically significant(all p-vales are geater than 0.05). I need the best selection of variables. setp() function does it by implementing Stepwise regression.

```
library(MASS)
step <- step(fit_all, direction="both", trace=FALSE)
summary(step)$coeff
```

```
##              Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)  9.617781  6.9595930  1.381946 1.779152e-01
## wt          -3.916504  0.7112016 -5.506882 6.952711e-06
## qsec         1.225886  0.2886696  4.246676 2.161737e-04
## factor(am)1  2.935837  1.4109045  2.080819 4.671551e-02
```

```
sprintf('Adjusted R-squared: %f', summary(step)$adj.r.squared)
```

```
## [1] "Adjusted R-squared: 0.833556"
```

according to this result, the best fit is:
Bestfit <- lm(mpg ~ wt+qsec+factor(am), data = mtcars)

The best model says that fuel consumption in cars mostly depends on the car's weight(wt), quarter mile time(qsec) and transmission system(am), and. The adjusted R-squared is 83% which means that the model explains 83% of the variation in mpg, indicating it is a robust model.
Based on the best model results, the mean of 'mpg' for cars with manual transmission system is 2.94 more than automatic transmission cars. We found a differnt value in this model from the simple linear regression model(which was 7.24), and that's because I adjusted first fit with other varialbes(wt, qsec).

Diagnostics

These are the results of my diagnostic plots(Fig-2):
Residuals vs Fitted: Residuals are patternless, they are randomly distributed above and below zero. that is a good indication that I don't have non-linear relationships.
Normal Q-Q: This plot shows if residuals are normally distributed, and they follow a straight line. The points are normally distributed around the line which is good.
Scale-Location: This plot shows if residuals are spread equally along the ranges of predictors. It's good if you see a horizontal line, but unfortunatly it's not our case here.
Residulas vs Leverage: This plot helps us to find influential cases. We Look for cases outside of a dashed line, those are influential to the regression results. There is no influential case in this plot.

Conclusion

Based on my Analysis, I can say that in terms of fuel efficiency, a manual transmission is a better option than automatic. Holding all other variables constant, on average, 'mph' in manual cars is 2.94 more than automatic cars. The best model I found explains 83% of the variability in the response variable(mpg).

Appendix

Fig_1

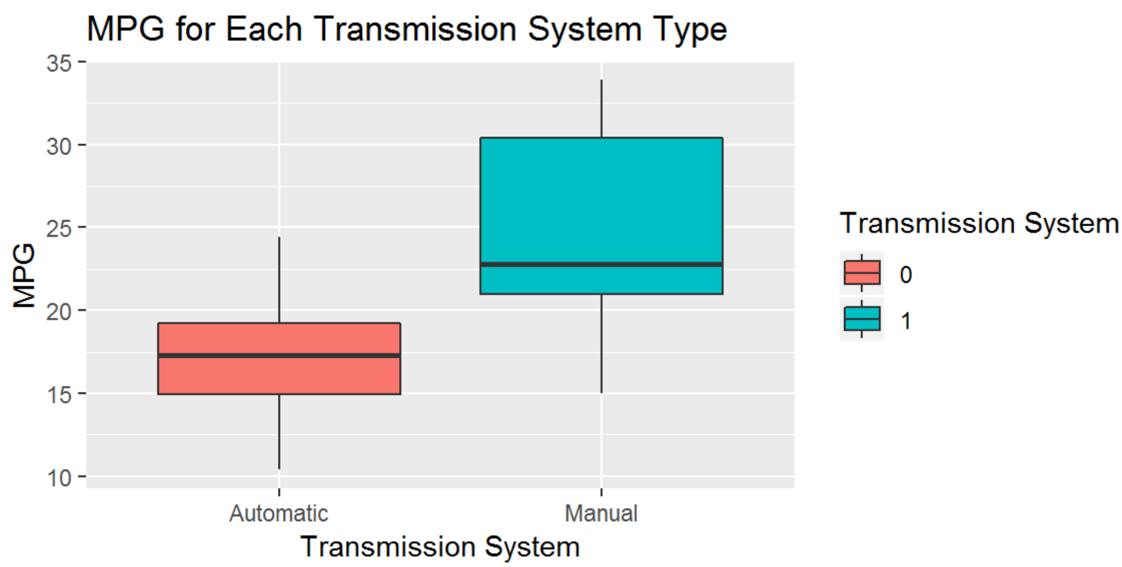


Table_1

```
summary(fit_all)$coeff
```

```
##                Estimate  Std. Error    t value   Pr(>|t|)
## (Intercept) 12.30337416 18.71788443  0.6573058 0.51812440
## cyl         -0.11144048  1.04502336 -0.1066392 0.91608738
## disp         0.01333524  0.01785750  0.7467585 0.46348865
## hp          -0.02148212  0.02176858 -0.9868407 0.33495531
## drat         0.78711097  1.63537307  0.4813036 0.63527790
## wt          -3.71530393  1.89441430 -1.9611887 0.06325215
## qsec         0.82104075  0.73084480  1.1234133 0.27394127
## factor(vs)1  0.31776281  2.10450861  0.1509915 0.88142347
## factor(am)1  2.52022689  2.05665055  1.2254035 0.23398971
## gear         0.65541302  1.49325996  0.4389142 0.66520643
## carb        -0.19941925  0.82875250 -0.2406258 0.81217871
```

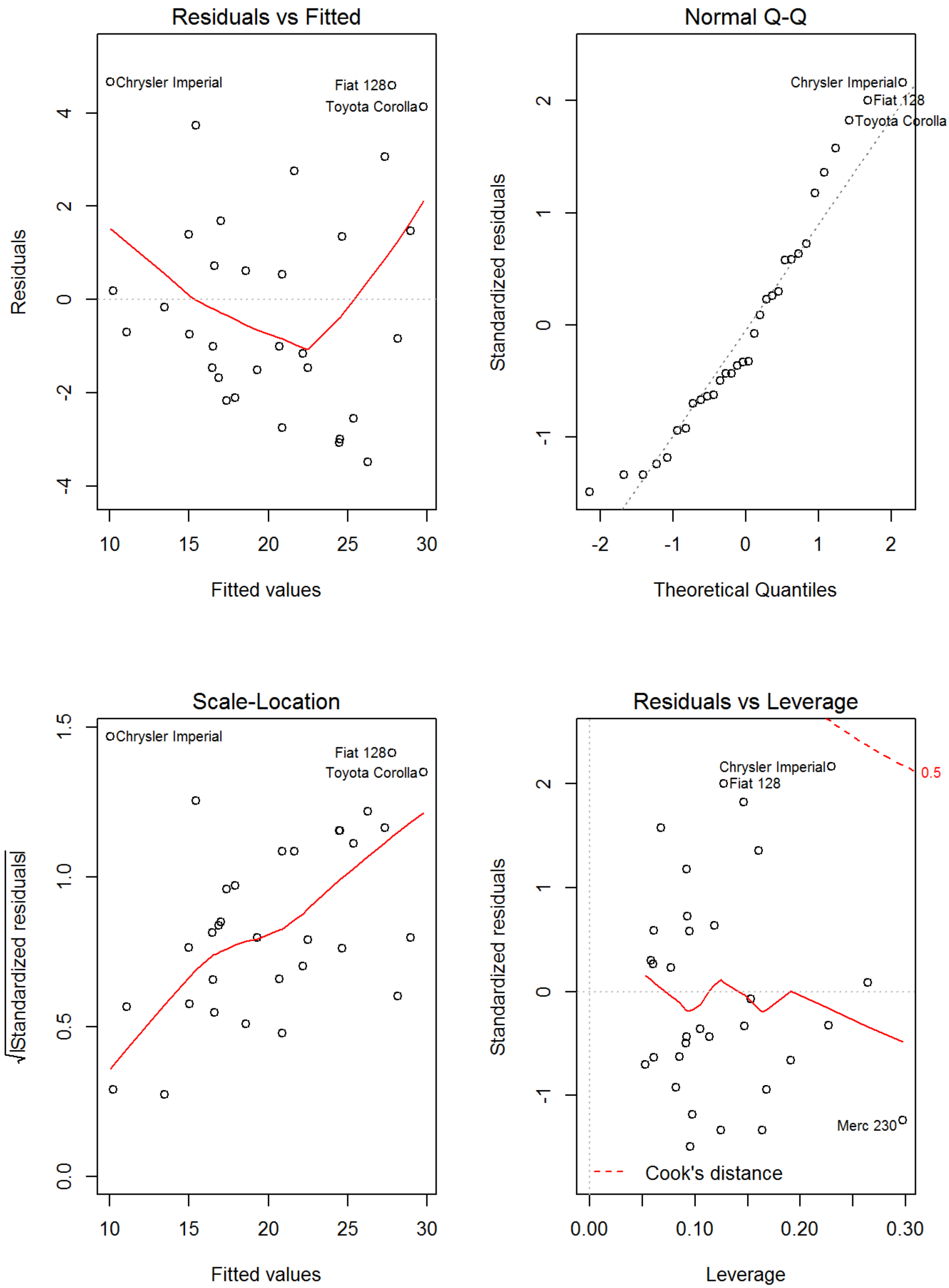
Stepwise Regression

The general idea behind the stepwise regression procedure is that we build our regression model from a set of candidate predictor variables by entering and removing predictors - in a stepwise manner - into our model until there is no justifiable reason to enter or remove any more.

Refrence: https://newonlinecourses.science.psu.edu/stat501/node/329/ (https://newonlinecourses.science.psu.edu/stat501/node/329/)

Fig-2

```
Bestfit <- lm(mpg ~ wt+qsec+factor(am), data = mtcars)
par(mfrow = c(2,2))
plot(Bestfit)
```


Residuals vs Fitted
Residuals
Fitted values
Chrysler Imperial
Fiat 128
Toyota Corolla
Normal Q-Q
Standardized residuals
Theoretical Quantiles
Scale-Location
√|Standardized residuals|
Residuals vs Leverage
Leverage
Merc 230
Cook's distance
0.5