Regression Models Course Project Report

This is a report on regression models course project, which will focus on the problem about the relationship between a set of variables and miles per gallon (MPG). At the beginning, the model was made just concerning the transmission as the predictor and the miles/gallon as output(`mpg~am`), which has great uncertainty with large t-value and p-value, and the diagnostics perform as unconvincing regression as well.To improve our model, another one and two factors are taken into consideration respectively. only 1/4 mile time(`mpg~am+qsec`), or both weight and number of forward carburetors(`mpg~am+wt+carbs`), can make adjusted model more convincible and precise. Through the comparison of `anova`, we accept the latter one(`mpg~am+qsec`) as the final model.

There are four coefficients in the model(including the intercept).

```
## (Intercept)          am          wt        carb
##   34.016296    2.526258   -3.633994   -1.159288
```

These coefficients of factors show the relationship of factors between the output(`mpg`). With the same weight and number of carburetors, for one gallon fuel, a manual transmission car will last 2.526258 more miles distance than automatic transmission one. **So munal transmission performs better than automatic transmission for MPG, and the difference between them is 2.526258.** Besides, heavier weight can reduce the MPG by 3.6339938 per 1000 pounds, and MPG will decrease by 1.1592877 when every single carburetor is added.

During the improvement, both common sense and experimental results are referenced. The chosen factors are intuitively related to MPG. As the transmission is binary, the single variable regression model cannot interpret the relationship forcefully, but it generally shows that the manual one have advantage on MPG. Considering no evidence for interaction, we add variables gradually to three with the experiments proving its confident.Finally, we find the relationship between `mpg` and those factors: `wt`, `am`, and `carb`.

Our steps of modelling can be found in the following description with some illustrative code and figures. Some analyses of data can also be found there. The experiments were done on Windows 8.1, with RStudio Version 0.99.332, and the report was written via knitr.

Firstly, when we simply consider the transmission as the only predictor to model:

```
fit_origin<-lm(mpg~am,data=mtcars)
fit_origin$coefficients
```

```
## (Intercept)          am
##   17.147368    7.244939
```

It simply means mpg for automatic transmission will be 17.1473684, while that for manual transmission will be 24.3923077.

```
summary(fit_origin)$coefficients
```

```
##              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
```

When looking at the coefficients, the problem appears. The intercept has a large t-value and the coefficient of `am` has a large p-value.In other world, for the intercept, the with of confident interval width is 30.4949843,which is terrible for regression. Considering the uncertainty, we fail to accept the model,and other evidences(regression and diagnostic figure) shows the same conclusion as following:

As the figures show, the model is not persuasive enough and other predictors should be taken into consideration. When choosing the second predictor, `qsec` can lead to relative lower t-value and p-value among all the factors except `am`,which does make sense in the real world.

```
fit1<-lm(mpg~am+qsec,data=mtcars)
summary(fit1)$coefficients
```

```
##               Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) -18.889281  6.5969729 -2.863326 7.710583e-03
## am            8.876331  1.2896638  6.882670 1.461462e-07
## qsec          1.981870  0.3601293  5.503218 6.270759e-06
```

If another two predictors included in the original model, `wt` and `carb` adjust the model better.

```
fit2<-lm(mpg~am+wt+carb,data=mtcars)
summary(fit2)$coefficients
```

```
##              Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 34.016296  2.9713276 11.448181 4.485516e-12
## am           2.526258  1.6478794  1.533036 1.364895e-01
## wt          -3.633994  0.9281206 -3.915433 5.269167e-04
## carb        -1.159288  0.4062840 -2.853393 8.046207e-03
```

And then variable selection can be made. Let’s take a close look at `anova`.

```
anova(fit_origin,fit1)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ am + qsec
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     30 720.90
## 2     29 352.63  1    368.26 30.285 6.271e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(fit_origin,fit2)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ am + wt + carb
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     30 720.90
## 2     28 215.62  2    505.28 32.807 4.586e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The p-value is small, so both of them are acceptable. But the latter is more reasonable and has better robustness. Thus, finally, `fit2` is chosen.

Appendix

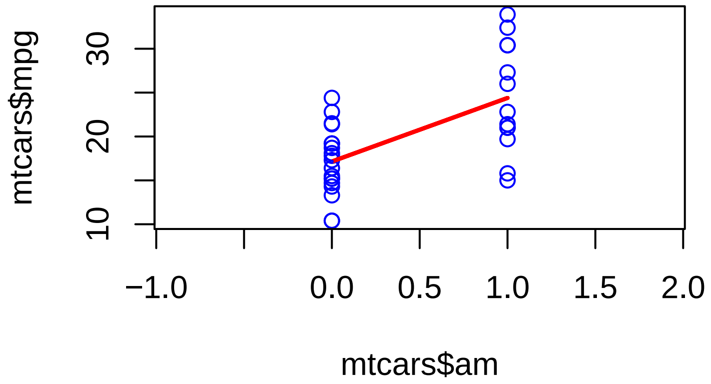

Figure 1: The regression of single factor `am`(`fit_origin`)

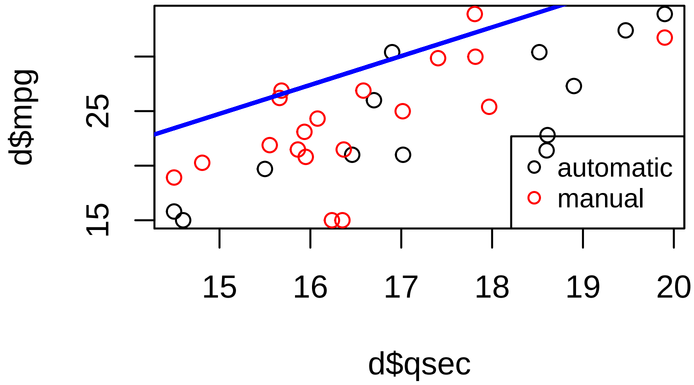

Figure 2: The relationship between `mpg` and `qsec` influenced by transimission

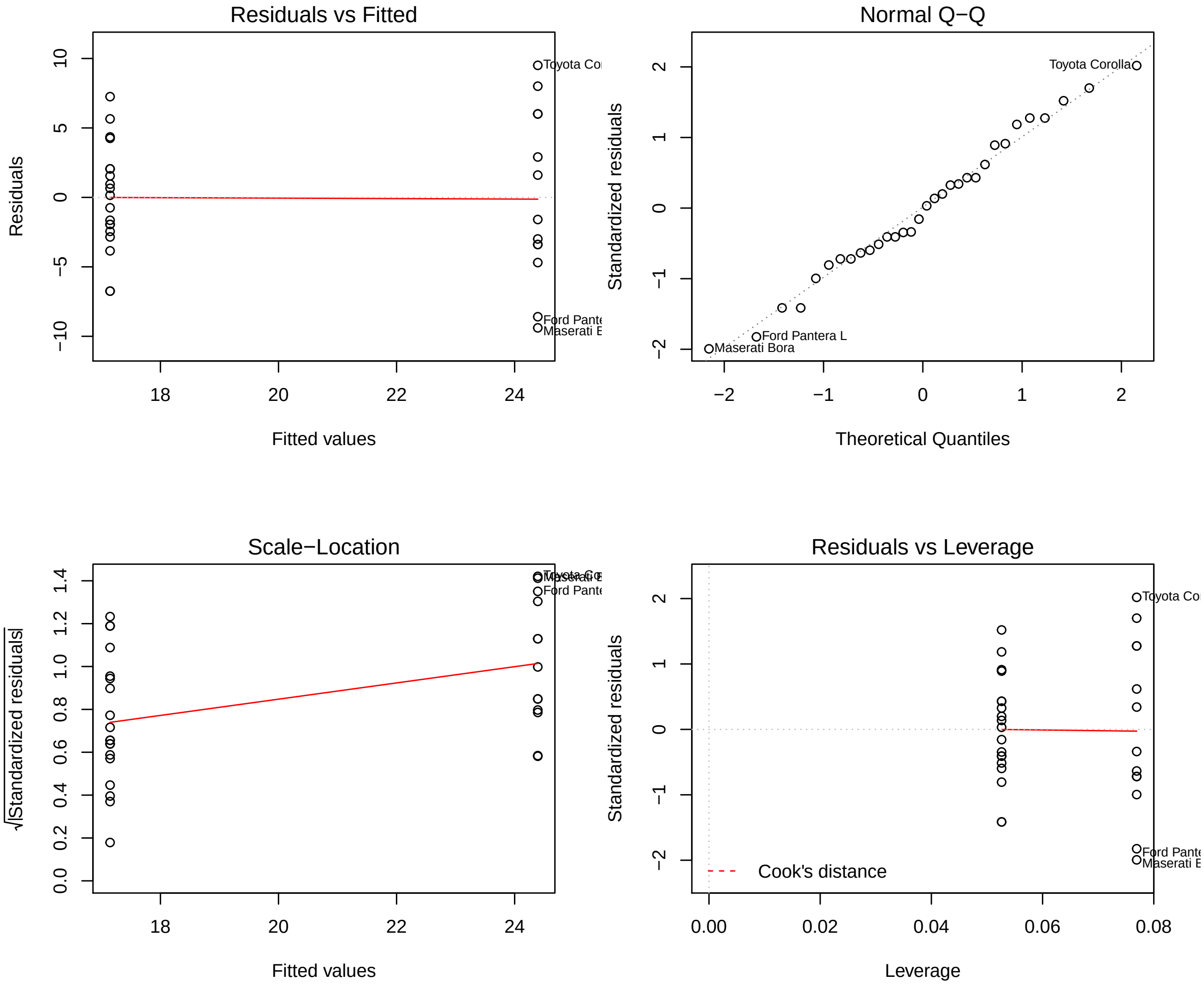

Figure 3: Diagnostics of `fit_origin`

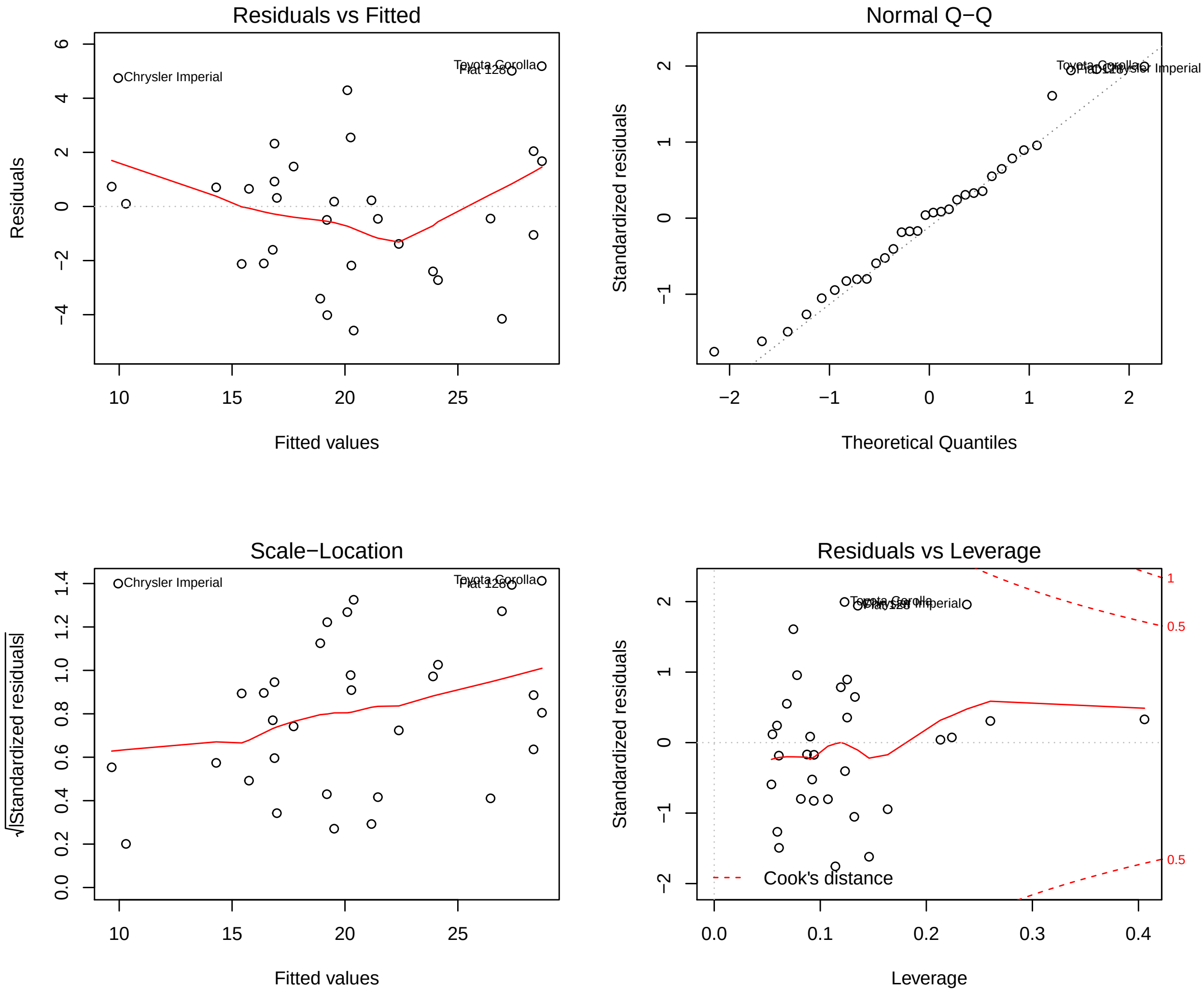

Figure 4: Diagnostics of `fit2`