Essay on the the Cars Mileage

Oleg Krivosheev

Executive Summary

Dataset **mtcars** from the 1974 *Motor Trend US* magazine is used to evaluate the effect of transmission assembly on the car mileage, expressed in miles per gallon (MPG). For all control variables considered together, we discovered there is a significant effect, while for transmission regressor alone desired effect is not quite significant. Only with inclusion of the additional regressors like weight and quarter mile time, we were able to describe the mileage better than 88%. Holding quoter mile time and weight constant, cars with manual transmission will have intercept higher by 14.08 and additional weight slope of -4.141.

Exploratory data analysis

First, we load data and convert it to `data.table`.

```
data(mtcars); dt <- as.data.table(mtcars)
```

All relevant variables are made into factors. What we're interested in are `am` column, whcih indicates whether car has automatic (am=0) or manual (am=1) transmission. From the summary table for the whole dataset, one could notice that mean value of the `am` column is about 41%, which means we have roughly 60% cars with automatic transmission.

In **Appendix A** we display summary tables for all relevant variables in the *mtcars* dataset In **Appendix B** we plot the MPG vs weight dependency and MPG vs quater mile time dependency grouped by transmission value. We can see both plots indicates high level of dependencies.

Regression Models

We computed mean MPG values for the cars with automatic transmission and it is equal to 17.15, which is significantly lower than same value of 24.39 for a manual transmission cars. What factors affect such difference in MPG lead us to build, analyse and compare different models.

First, we build full regression model, where `MPG` depends on all regressors available to us.

```
fit.all <- lm(mpg ~ ., data = dt)
```

In this model, the produced summary tells us that we have adjusted R^2 equal to 0.78, thus the full model leave unexplained only 22% of the variation in the residuals. Residual standard error is equal to 2.833, with 15 degress of freedom. From the other hand, none of the regressors are marked by stars in the summary table, they are not significant at 5% significance level. Thus, we believe fitting so many regressors lead to multicollinearity and overfitting with inflated estimated standard error.

Baseline model to build is where MPG depends on transmission regressor only, and no other regressors are included.

```
fit.base <- lm(mpg ~ am, data = dt)
```

From fitting the baseline model we know that while we have highest significance indicator (three stars), the adjusted R^2 is only 0.34, thus implying transmission assembly factor alone has quite small explanatory power.

Residual standard error is also higher, at 4.9. We plot both all inclusive and baseline model in **Appendix C** and **D**.

We will use *Akaike's Information Criterion*, AIC. It is equal to the log of the likelyhood and proportional to the number of parameters.

$$AIC = -2 * log(L) + C * N_{par},$$

where N_{par} is number of parameters, L is likelyhood, and coefficient C is equal to 2 for classical AIC or $log(N_{obs})$ for BIC. We will use `step` function to get a clue about better model.

```
fit.step <- step(fit.all, k=log(nrow(dt)))
```

As a result of the AIC test, we could conclude that good linear mode shall include `wt`, `qsec` and `am` as regressors. We build and check this model.

```
fit.aic <- lm(mpg ~ wt + qsec + am, data = dt)
```

Indeed, we have now R^2 equal to 0.83, which left only 17% of the residual variance unexplained. Residual standard error is equal to 2.46. All regressors are significant at least 0.05 significance level. We believe, looking at scatter plots, there is a significant interaction between weight and transmission assembly. We would like to build model which reflects such interaction.

```
fit.best <- lm(mpg ~ qsec + wt * am, data = dt)
```

This model is clearly the best so far. All regressors including `qsec`, `wt`, `am` and `wt:am` are significant with level of significance at least 0.01. Adjusted R^2 is equal to 0.88, which means 88% of the variance of the residuals described by this model. Residual standard error is equal to 2.08 on 27 degress of freedom. Plots for AIC model and the best model could be found in **Appendix E** and **F**. Final MPG model best fit is

$$MPG = 9.723 + 1.017 * qsec - 2.937 * wt + 14.079 * am - 4.141 * wt * am$$

For `am` equal to 0 (auto) vs `am` equal to 1 (manual), holding weight and quater mile time constant, we got a lot higher intercept (+14.079) and additional negative weight slope of -4.141.

Analysis

Confidence interval, using t-test, is equal to

```
##        (Intercept)  qsec     wt    am1 wt:am1
## 2.5 %       -2.381 0.500 -4.303  7.031 -6.597
## 97.5 %      21.827 1.534 -1.570 21.128 -1.686
```

ANOVA test for all models is included into **Appendix G**, as well as confidence interval for the best model.

References

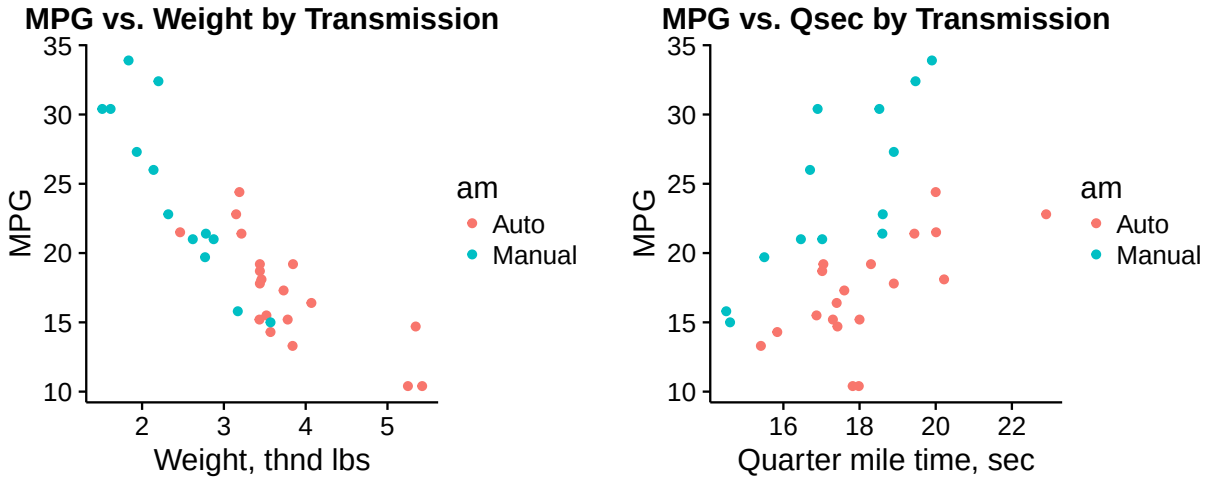
All project files could be downloaded from my github page. Cowplot package was taken from here.

Appendix A

mpg	cyl	disp	hp	drat
Min. :10.40	Min. :4.000	Min. : 71.1	Min. : 52.0	Min. :2.760
1st Qu.:15.43	1st Qu.:4.000	1st Qu.:120.8	1st Qu.: 96.5	1st Qu.:3.080
Median :19.20	Median :6.000	Median :196.3	Median :123.0	Median :3.695
Mean :20.09	Mean :6.188	Mean :230.7	Mean :146.7	Mean :3.597
3rd Qu.:22.80	3rd Qu.:8.000	3rd Qu.:326.0	3rd Qu.:180.0	3rd Qu.:3.920
Max. :33.90	Max. :8.000	Max. :472.0	Max. :335.0	Max. :4.930

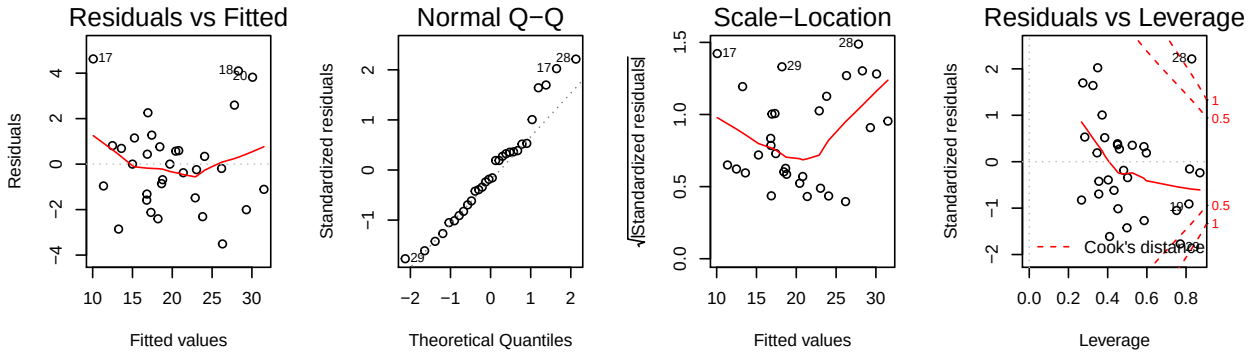
wt	qsec	vs	am	gear
Min. :1.513	Min. :14.50	Min. :0.0000	Min. :0.0000	Min. :3.000
1st Qu.:2.581	1st Qu.:16.89	1st Qu.:0.0000	1st Qu.:0.0000	1st Qu.:3.000
Median :3.325	Median :17.71	Median :0.0000	Median :0.0000	Median :4.000
Mean :3.217	Mean :17.85	Mean :0.4375	Mean :0.4062	Mean :3.688
3rd Qu.:3.610	3rd Qu.:18.90	3rd Qu.:1.0000	3rd Qu.:1.0000	3rd Qu.:4.000
Max. :5.424	Max. :22.90	Max. :1.0000	Max. :1.0000	Max. :5.000

Appendix B



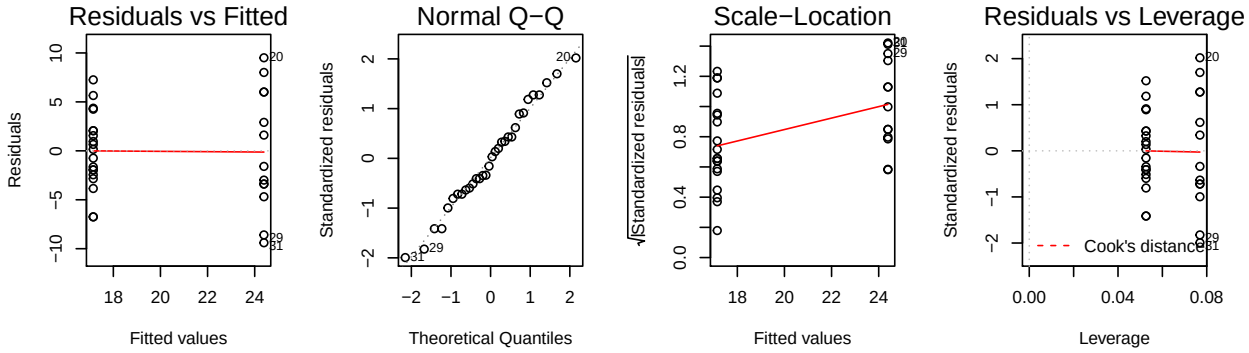
Appendix C

Plot fot all regressors included model



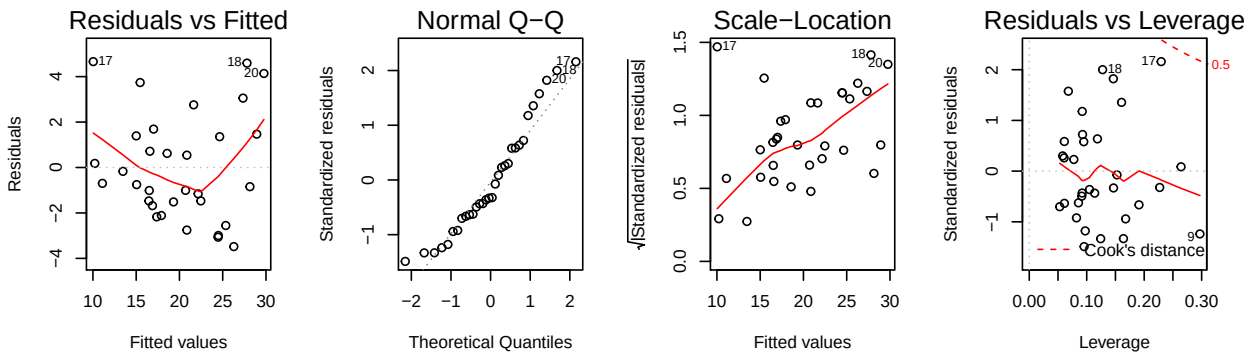
Appendix D

Plot for MPG vs transmission regressor model



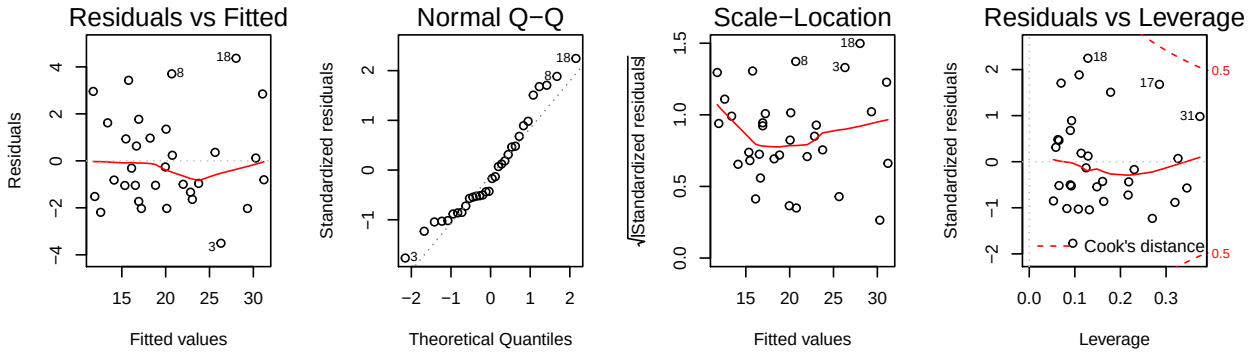
Appendix E

Plot for MPG vs weight + qsec + transmission regressors model



Appendix F

Plot for MPG vs qsec + weight*transmission regressors model



Appendix G

ANOVA test for all four models

```
anova(fit.all, fit.base, fit.aic, fit.best)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ cyl + disp + hp + drat + wt + qsec + vs + am + gear + carb
## Model 2: mpg ~ am
## Model 3: mpg ~ wt + qsec + am
## Model 4: mpg ~ qsec + wt * am
##   Res.Df    RSS  Df Sum of Sq       F    Pr(>F)
## 1     15 120.40
## 2     30 720.90 -15   -600.49  4.9874  0.001759 **
## 3     28 169.29   2    551.61 34.3604 2.509e-06 ***
## 4     27 117.28   1     52.01  6.4795  0.022398 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
confint(fit.best)
```

```
##                  2.5 %    97.5 %
## (Intercept) -2.3807791 21.826884
## qsec         0.4998811  1.534066
## wt          -4.3031019 -1.569960
## am1          7.0308746 21.127981
## wt:am1      -6.5970316 -1.685721
```