Analysis of variables affecting a car's fuel efficiency

Robert Merriman

November 22, 2015

Executive Summary

I work for Motor Trend, a magazine about the automobile industry. I analyzed the mtcars car data to answer the following two questions:

1. Is an automatic or manual transmission better for MPG?
2. Quantify the MPG difference between automatic and manual transmissions?

For the data analyzed I concluded that a car with a manual tranmission is likely to have a higher fuel efficiency (MPG) than a car with an automatic transmission. However, upon building linear regressions, I concluded that the transmission type did not statistically significant affect MPG and the is confounded by other variables. As the data did not include any cars that were identical except for the transmission type, further analysis may be needed.

Methodology

Source Data

The source data is mtcars in the R datasets library. "The data was extracted from the 1974 Motor Trend US magazine, and comprises fuel consumption and 10 aspects of automobile design and performance for 32 automobiles (1973–74 models)."ˆReference 1, figure set 1ˆ

Data Manipulation

I converted the variables vs and am as factors and I relabled the am factors to Autmoatic and Manual to improve understanding the visuals. I did not convert the variables gear, cyl, or carb as factors as the variables are not dichotomous or ordinal. For example increasing the number of gears has enabled better engineering optimization of a car's fuel efficiencyˆReference 2ˆ.

Data Exploration

In exploring the data I concluded that it is reasonable to consider MPG approximately normally distributed. I noted that there are outliers at the right tail due to some cars with manual transmission having high MPG.ˆFigure set 2ˆ

t.test of MPG ~ AM

The mean mileage of cars with an automatic transmission is 17 mpg and for cars with a manual transmission is 23 mpg. Treating the MPG data as i.i.d. and normally distributed, by t.test comparing MPG to the transmission type, the 95% confidence interval of the difference in mean gas mileage is between 3.21 and 11.28 mpg. As the mean does not include 0 and the p.value 0.0014 is less than an alpha level of 0.05, the difference is statistically significant.

Model Selection for linear regerssion

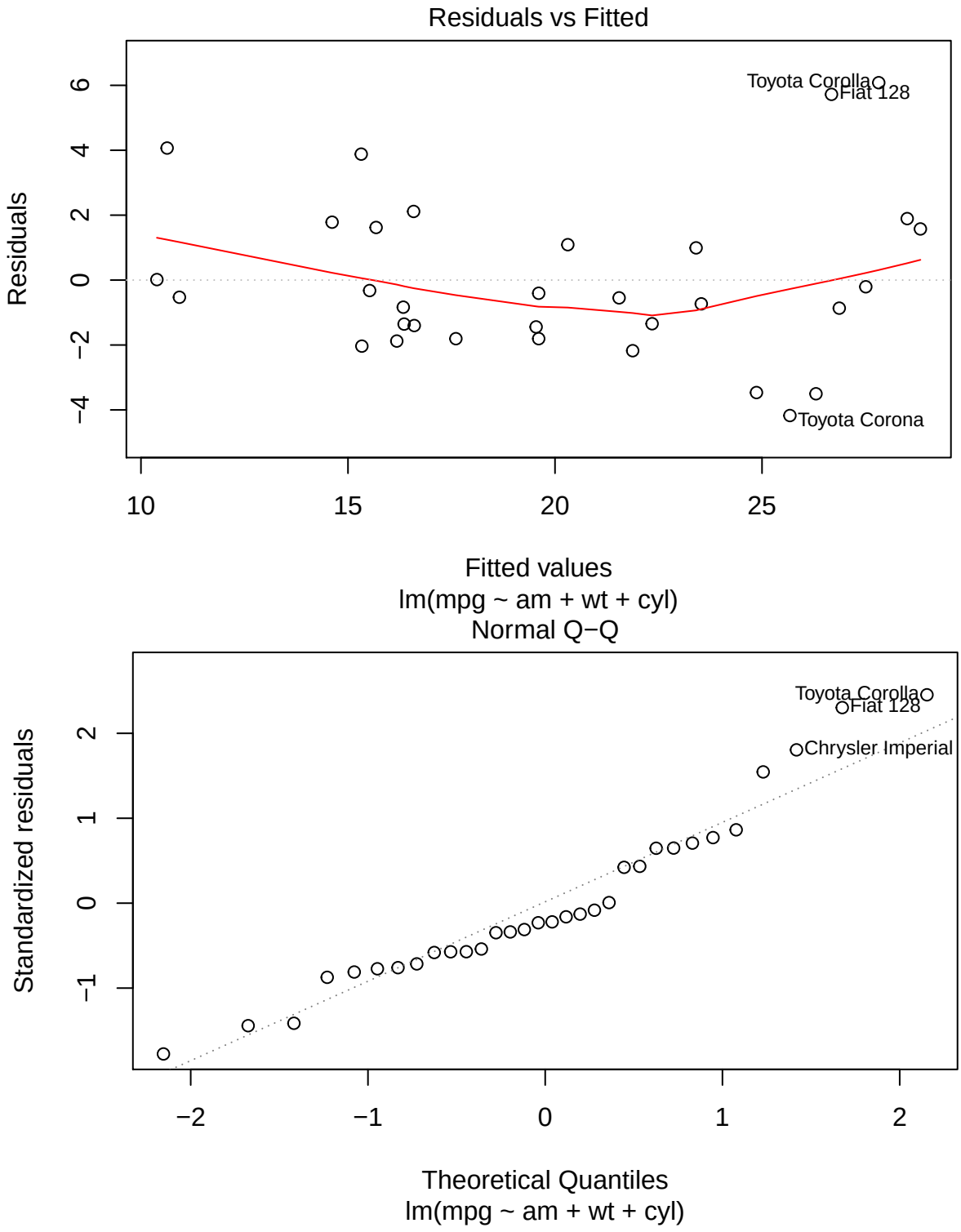
From prior articles[3] I believe that multiple approaches to building a regression model should be employed. I chose two approaches: nested model selection using ANOVA and using stepAIC.

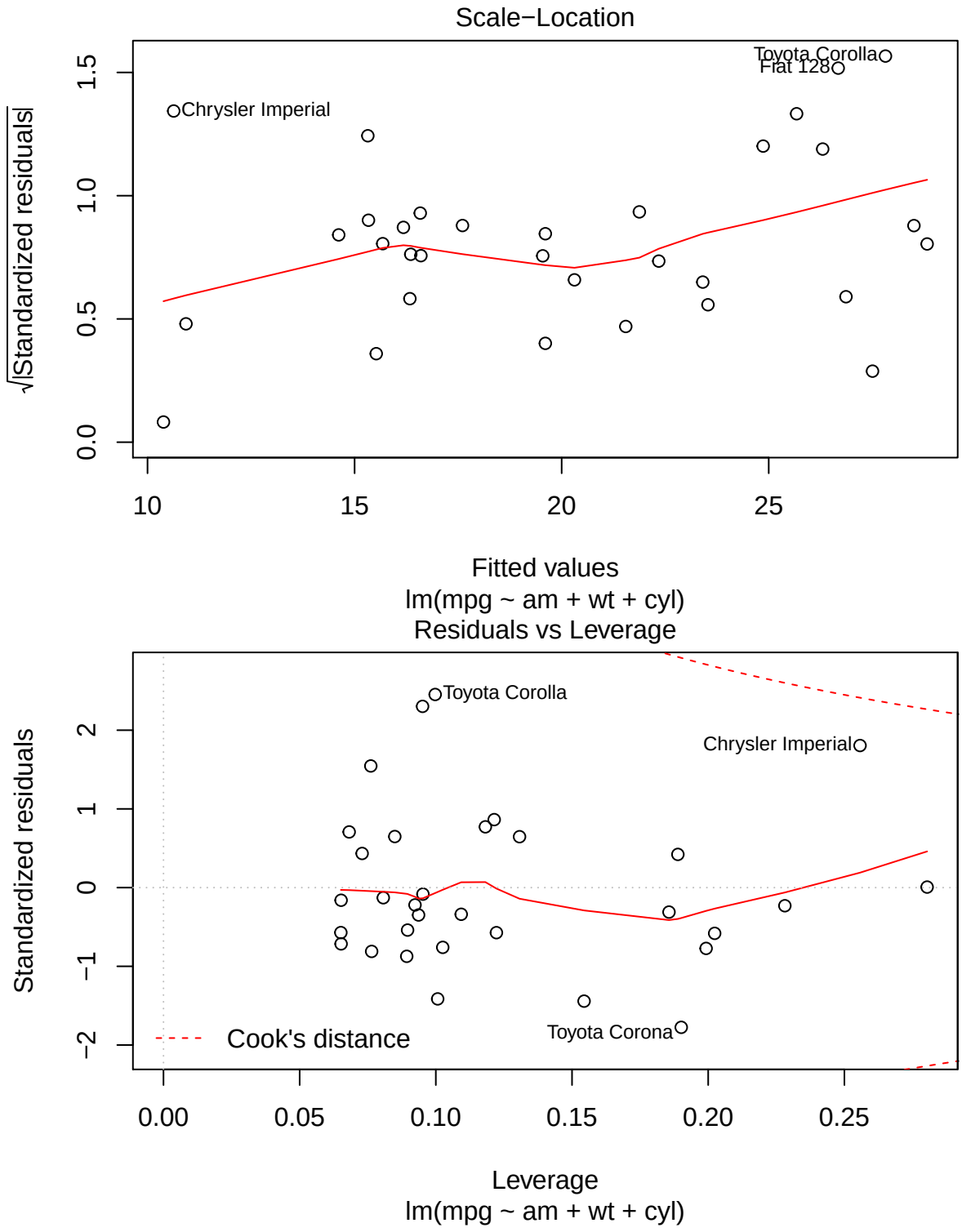
For the nested model section approach using ANOVA the best regression model was (mpg ~ am + wt + cyl) which did not find transmission type to be statistically significant and weight and cylinders were statistically significant confouders (at an alpha of 0.05). Using stepAIC the best regrsesion model was (mpg ~ wt + qsec + am) which did find transmission type to be statistically significant with qsec and weight being statistically significant confounders (at an alpha of 0.05). I concluded that transmission type was not statistically significant and other variables were confounders.

Normalacy and residuals

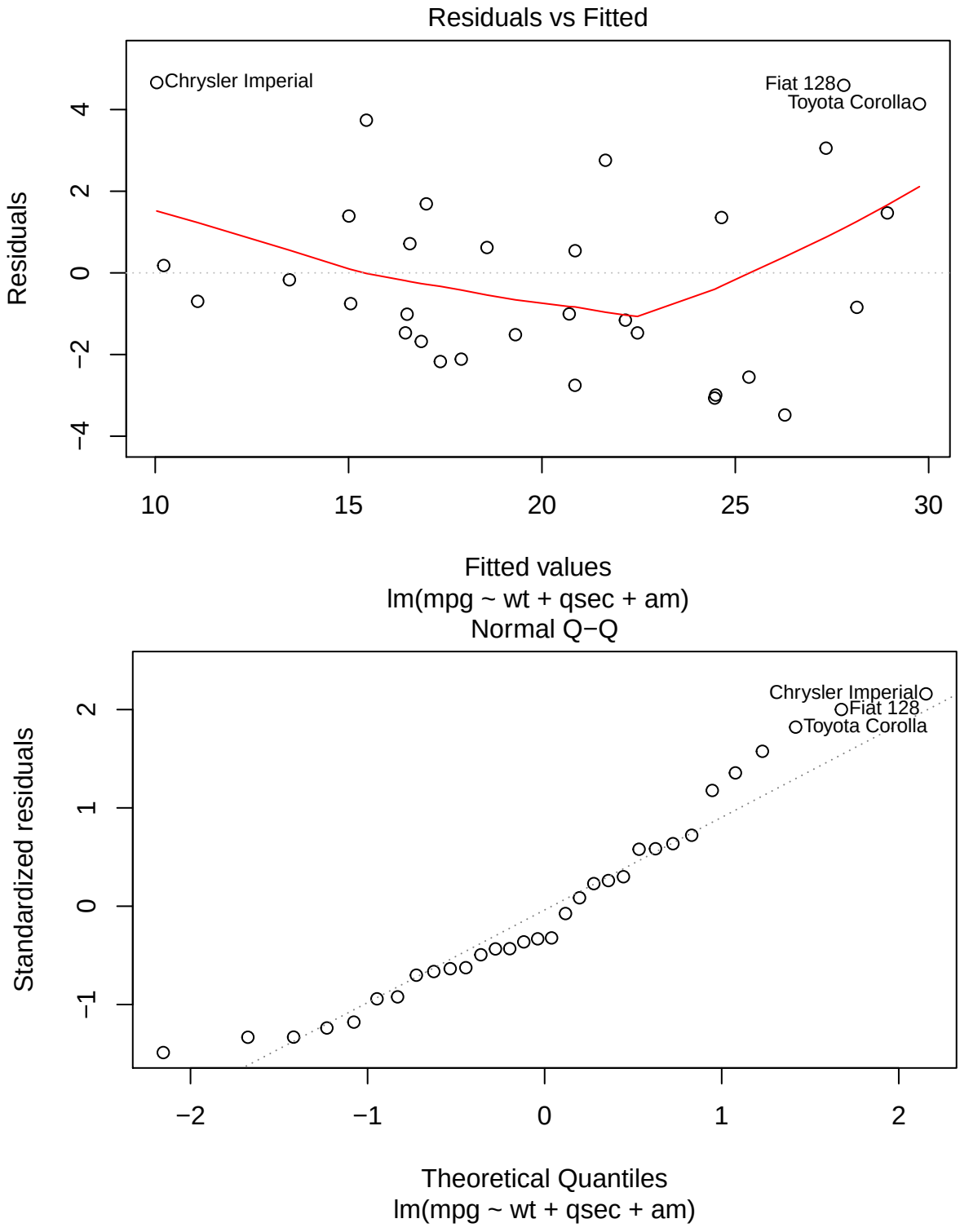
The residuals support the claim of uncorrelated and homskedastic errors, which are needed for standard hypothesis tests. Since the sample size is small, lets take a look at the qq plot to check for normality of errors. qqnorm(residuals(reg3), main="QQ Plot") qqline(residuals(reg3))

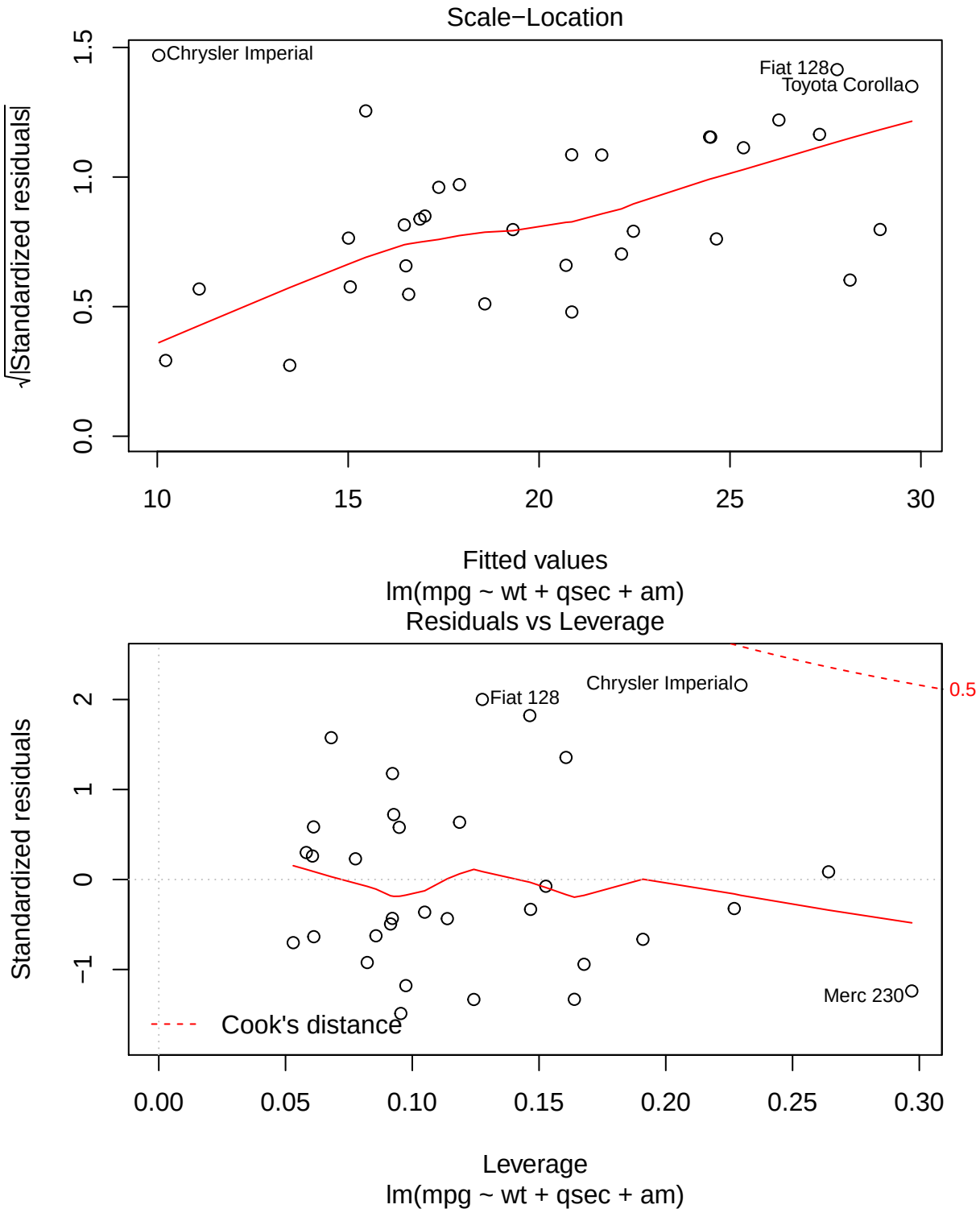
```
#residuals
plot(lmcyl)
```


Residuals vs Fitted
Residuals
Fitted values
lm(mpg ~ am + wt + cyl)
Toyota Corolla
Fiat 128
Toyota Corona
Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(mpg ~ am + wt + cyl)
Toyota Corolla
Fiat 128
Chrysler Imperial




```
plot(bestmodel)
```


Residuals vs Fitted
Chrysler Imperial
Fiat 128
Toyota Corolla
Residuals
4
2
0
−2
−4
10
15
20
25
30
Fitted values
lm(mpg ~ wt + qsec + am)
Normal Q−Q
Chrysler Imperial
Fiat 128
Toyota Corolla
Standardized residuals
2
1
0
−1
−2
−1
0
1
2
Theoretical Quantiles
lm(mpg ~ wt + qsec + am)



References

1. mtcars dataset
2. Gear article

3. fivethirtyeight
4. QQNorm
5. Shapiro-Wilk normality test
6. r-tutur.com
7. Variances explained
8. Histogram Blog

Appendix

Figure set 1: basic data exploration

```
##       mpg             cyl             disp             hp
##  Min.   :10.40   Min.   :4.000   Min.   : 71.1   Min.   : 52.0
##  1st Qu.:15.43   1st Qu.:4.000   1st Qu.:120.8   1st Qu.: 96.5
##  Median :19.20   Median :6.000   Median :196.3   Median :123.0
##  Mean   :20.09   Mean   :6.188   Mean   :230.7   Mean   :146.7
##  3rd Qu.:22.80   3rd Qu.:8.000   3rd Qu.:326.0   3rd Qu.:180.0
##  Max.   :33.90   Max.   :8.000   Max.   :472.0   Max.   :335.0
##       drat             wt             qsec       vs             am
##  Min.   :2.760   Min.   :1.513   Min.   :14.50   0:18   Automatic:19
##  1st Qu.:3.080   1st Qu.:2.581   1st Qu.:16.89   1:14   Manual   :13
##  Median :3.695   Median :3.325   Median :17.71
##  Mean   :3.597   Mean   :3.217   Mean   :17.85
##  3rd Qu.:3.920   3rd Qu.:3.610   3rd Qu.:18.90
##  Max.   :4.930   Max.   :5.424   Max.   :22.90
##       gear            carb
##  Min.   :3.000   Min.   :1.000
##  1st Qu.:3.000   1st Qu.:2.000
##  Median :4.000   Median :2.000
##  Mean   :3.688   Mean   :2.812
##  3rd Qu.:4.000   3rd Qu.:4.000
##  Max.   :5.000   Max.   :8.000
```

```
## '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  : Factor w/ 2 levels "0","1": 1 1 2 2 1 2 1 2 2 2 ...
##  $ am  : Factor w/ 2 levels "Automatic","Manual": 2 2 2 1 1 1 1 1 1 1 ...
##  $ gear: num  4 4 4 3 3 3 3 4 4 4 ...
##  $ carb: num  4 4 1 1 2 1 4 2 2 4 ...
```

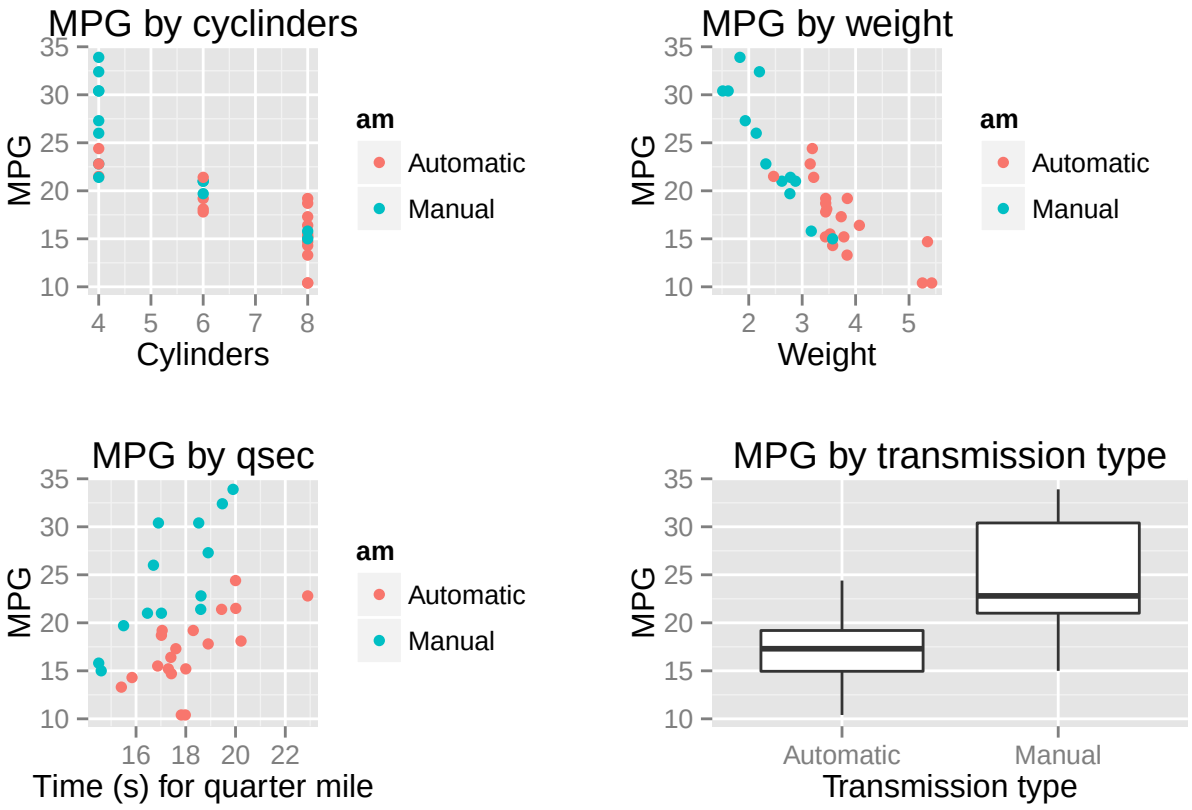


Figure set 2: MPG distribution

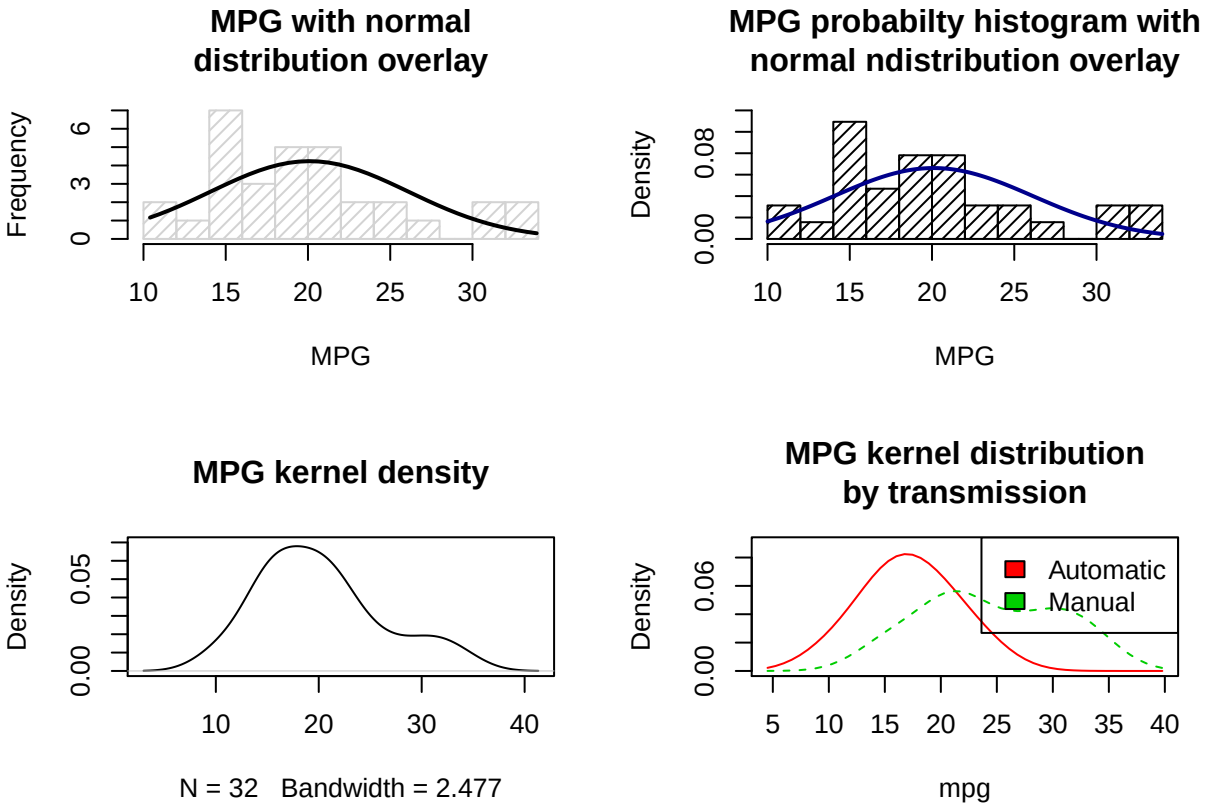


Figure set 3

ANOVA

```
## Analysis of Variance Table
##
## Model  1: mpg ~ am
## Model  2: mpg ~ am + wt
## Model  3: mpg ~ am + wt + cyl
## Model  4: mpg ~ am + wt + cyl + disp
## Model  5: mpg ~ am + wt + cyl + disp + hp
## Model  6: mpg ~ am + wt + cyl + disp + hp + drat
## Model  7: mpg ~ am + wt + cyl + disp + hp + drat + qsec
## Model  8: mpg ~ am + wt + cyl + disp + hp + drat + qsec + vs
## Model  9: mpg ~ am + wt + cyl + disp + hp + drat + qsec + vs + gear
## Model 10: mpg ~ am + wt + cyl + disp + hp + drat + qsec + vs + gear + carb
##    Res.Df    RSS Df Sum of Sq       F    Pr(>F)
## 1      30 720.90
## 2      29 278.32  1    442.58 63.0133 9.325e-08 ***
## 3      28 191.05  1     87.27 12.4257   0.00201 **
## 4      27 188.43  1      2.62  0.3732   0.54782
## 5      26 163.12  1     25.31  3.6030   0.07151 .
## 6      25 162.43  1      0.69  0.0977   0.75768
## 7      24 149.09  1     13.34  1.8999   0.18260
## 8      23 148.87  1      0.22  0.0309   0.86214
## 9      22 147.90  1      0.97  0.1384   0.71365
## 10     21 147.49  1      0.41  0.0579   0.81218
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

stepAIC

```
##
## Call:
## lm(formula = mpg ~ wt + qsec + am, data = M1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4811 -1.5555 -0.7257  1.4110  4.6610
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   9.6178     6.9596   1.382 0.177915
## wt           -3.9165     0.7112  -5.507 6.95e-06 ***
## qsec          1.2259     0.2887   4.247 0.000216 ***
## amManual      2.9358     1.4109   2.081 0.046716 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.459 on 28 degrees of freedom
## Multiple R-squared:  0.8497, Adjusted R-squared:  0.8336
## F-statistic: 52.75 on 3 and 28 DF,  p-value: 1.21e-11
```