Mileage and Transmission

Aseem Rohatgi

Overview

We at *Motor Trend* looked at a collection of cars and explored the relationship between a set of variables describing the car and miles per gallon that the car achieved. Particularly, we were interested in answering the following two questions:

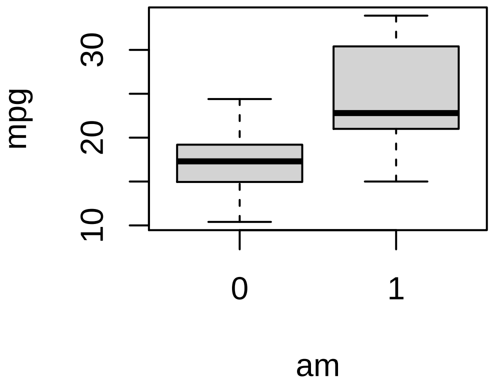
- "Is an automatic or manual transmission better for MPG"
- "Quantify the MPG difference between automatic and manual transmissions"

We found that:

- Manual transmission provided better mileage than automatic mileage.
- The difference in mileage between the two transmissions were around 7.2 miles per gallon. However, this huge difference was reduced by a lot when the number of cylinders in the car were also considered.
  - 4 cylinders, difference of 4.1 miles/gal
  - 6 cylinders, difference of 0.2 miles/gal
  - 8 cylinders, difference of 0.6 miles/gal

Data

We used the data we collected on 32 automobiles('73 & '74 models). We have stored the data in a data frame called `mtcars`. We are interested in the how MPG, stored in the variable `mpg`, change with the transmission type, stored in the variable `am` as 0 for automatic, and 1 for manual. Let's look at a boxplot for the data.



This initial look at the data suggests that manual transmission seems to be better. We performed regression analysis to solidify this exploratory observation with statistical significance.

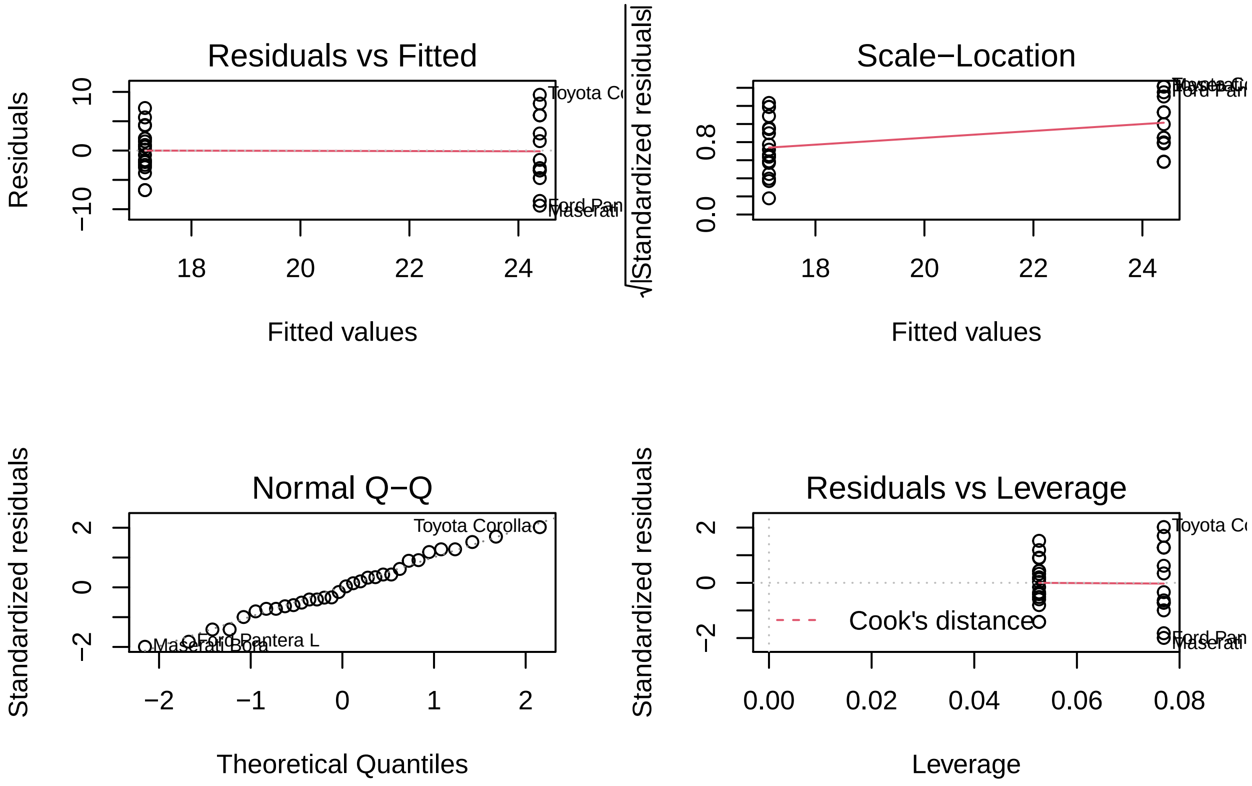
Analysis

Assessing if there is a difference

First, we performed a t-test to provide a solid statistical background to our observation of difference in mileage obtained from the two transmission types. The test provided a p-value of 0.0013736, suggesting a strong difference between the means of `mpg` under the two forms of transmission. Our estimates were 17.1473684 for automatic transmission, and 24.3923077 for manual(See Appendix for detailed code). Thus, we concluded that automobiles with a manual transmission provide a better mileage than those with an automatic transmission.

Quantifying the difference

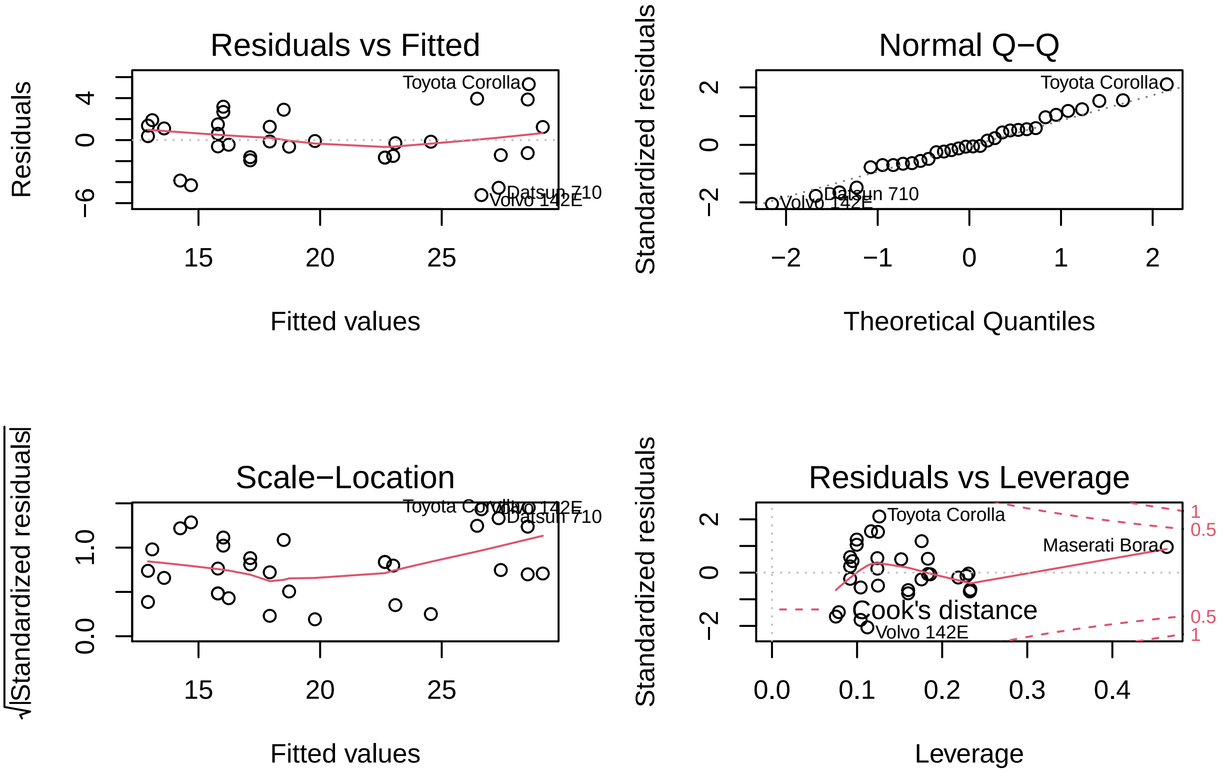
To answer our second question of quantifying our result, we performed a regression analysis on `mpg` with `am` as a regressor. Our model was fairly robust as can be seen from the following diagnostic plots.



However, our model was only able to explain about a third of the variance in `mpg`, as is evidenced by the R^2 value of 0.3597989. Hence, we began adding more variables in our model to explain more of the variance, while also making sure that the results we obtained were significant.

There were 4 variables which, when added to the model, raised the adjusted R^2 to above 70% - `cyl`,`disp`,`hp` and `wt`. These represented the number of cylinders, the displacement, the gross horsepower, and the weight of the automobile. After performing regression with different combinations of variables and also performing an Analysis of Deviance on those models, we made some interesting conclusions.

For instance, adding `wt` essentially rendered the transmission type useless, suggesting that mileage depends much more on the weight of the car than it does on its transmission type. `disp` did pretty much nothing when it was added at the end, leading us to drop it from our model. However, the addition of `cyl` and `hp` to the regressor `am` was effective in increasing the effectiveness of our model. This model, which contained 3 regressors, was able to explain about 80% of the variance in `mpg`(details in appendix). The diagnostics plots are given below.



Appendix

1. Boxplot

```
boxplot(mpg~am,mtcars)
```

2. t-test

```
t<-t.test(mpg~am,data=mtcars)
```

3. Diagnostic plots for first model

```
par(mfcol=c(2,2))
mdl1<-lm(mpg~factor(am),mtcars)
plot(mdl1)
```

4. First model summary

```
summary(mdl1)
```

```
##
## Call:
## lm(formula = mpg ~ factor(am), data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.3923 -3.0923 -0.2974  3.2439  9.5077
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   17.147      1.125  15.247 1.13e-15 ***
## factor(am)1    7.245      1.764   4.106 0.000285 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 4.902 on 30 degrees of freedom
## Multiple R-squared:  0.3598, Adjusted R-squared:  0.3385
## F-statistic: 16.86 on 1 and 30 DF,  p-value: 0.000285
```

5. Considering addition of different variables was first done by adding one variable. The models are stored in a list and can be summarized using `lapply(mdl2,summarise)`.

```
mdl2<-list()
for(i in c(2:8,10,11)){
  if(i %in% c(2,8,10,11))
    mdl2[[i]]<-lm(mpg~factor(am)+factor(mtcars[,i]),mtcars)
  else
    mdl2[[i]]<-lm(mpg~factor(am)+mtcars[,i],mtcars)
}
```

After this, comparison of models using different combinations of the chosen variables was performed using Analysis of Deviance. For example,

```
anova(lm(mpg~factor(am),mtcars),
      lm(mpg~factor(am)+factor(cyl),mtcars),
      lm(mpg~factor(am)+factor(cyl)+hp,mtcars))
```

6. Diagnostic plots for the final model

```
mdl_f<-lm(mpg~factor(am)+factor(cyl)+hp,mtcars)
par(mfrow=c(2,2))
plot(mdl_f)
```

7. Final model summary

```
summary(mdl_f)
```

```
##
## Call:
## lm(formula = mpg ~ factor(am) + factor(cyl) + hp, data = mtcars)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -5.231 -1.535 -0.141  1.408  5.322
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  27.29590    1.42394  19.169  < 2e-16 ***
## factor(am)1   4.15786    1.25655   3.309  0.00266 **
## factor(cyl)6 -3.92458    1.53751  -2.553  0.01666 *
## factor(cyl)8 -3.53341    2.50279  -1.412  0.16943
## hp           -0.04424    0.01458  -3.035  0.00527 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.703 on 27 degrees of freedom
## Multiple R-squared:  0.8249, Adjusted R-squared:  0.7989
## F-statistic: 31.79 on 4 and 27 DF,  p-value: 7.401e-10
```