Explore the Relationship Between MPG and Transmission Type

Course Project for Regression Models

Feng Li

26 Jan 2016

Executive Summary

In this report we will analyze `mtcars` data set, which is a data set containing fuel consumption (mpg) and 10 aspects of automobile design and performance for 32 automobiles extracted from the 1974 *Motor Trend* Magazine.
We'll explore the relationship between mpg and other variables, with focus on two questions:
1. Is an automatic or manual transmission better for mpg?
2. What's the mpg difference between automatic and manual transmissions?

Data Processing

We convert categorical variables `cyl` , `vs` , `am` , `gear` , `crab` into factors. Note that variable `am` has two values, among which 0 is automatic transmission, and 1 is manual transmission. We split the data into two subsets according to `am` .

Exploratory Analysis

To get a general idea about the difference of mpg between automatic and manual transmissions, we make a boxplot and a density plot(refer to appendix). Both plots show that manual cars have higher mpg, but with more variation as well.

Statistical Inference

The mean mpg of manual transmission cars is 7.245 higher than that of automatic transmission cars. Is this a significant difference? We'll perform a one-sided t-test to find it out.

As the t-test result shows, the p-value is 0.000687, and the 95% confidence interval (-Inf,-3.913256) is below 0, so we reject the null hypothesis, and are in favor of the alternative hypothesis that true mean value of mpg for manual transmission cars are higher than that of automatic ones.

Linear Regression Analysis

In this part, we first fit a linear model between `am` and the outcome `mpg` . The result shows that both slope and intercept coefficients are significant at 0.05 significant level, but the adjusted R-squared is 0.3385, which means this model can only explain 33.85% of the variance of the `mpg` variable. The residual standard error is 4.902 on 30 degrees of freedom.

```
model.lm<- lm(mpg ~ am, data = mtcars)
summary(model.lm)
```

Model Selection

So we expand our predictor scope to other 9 variables in the dataset. Specifically, we use `step()` function to choose optimal model from a collection of models using the subsets of the variables.

```
model.best<- step(lm(mpg ~ ., data = mtcars), direction = "backward")
summary(model.best)
```

The best model it returns uses `cyl` , `hp` , `wt` , `am` as predictors. Its adjusted R-squared is 0.8401, much larger than only taking `am` as predictor.

The coefficient for `am1` is 1.809 with a standard error 1.396, which can be interepreted as, other variables held constant, a change from automatic to manual transmission will increase the mpg by 1.809 miles per gallon.

However, the p-value associated with the transmission type variable is 0.207, well above the 0.05 significant level.

So we remove the `cyl` variable to fit a new model. And we compare these nested models using `anova()` function.

```
## Analysis of Variance Table
## 
## Model 1: mpg ~ am
## Model 2: mpg ~ hp + wt + am
## Model 3: mpg ~ cyl + hp + wt + am
##   Res.Df    RSS Df Sum of Sq       F    Pr(>F)    
## 1     30 720.90                                   
## 2     28 180.29  2    540.61 46.5343 2.566e-09 ***
## 3     26 151.03  2     29.27  2.5191       0.1 .  
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The anova test result shows that, adding `hp` and `wt` to the Model 1 reduces the residual sum square dramatically from 720.90 to 180.29, and the p-value for F-test is significant, but very weak evidence for `mpg` effect when adding `cyl` to Model 2.

So we choose Model 2 as the best fit: `lm(mpg ~ hp + wt + am, data = mtcars)` . The estimated coefficient of `am1` (with `am0` as base) is 2.084, which can be interpreted as an increase of 2.084 `mpg` when we change from automatical to a manual transmission. The 95% confidence interval for `am1` coefficient is (-0.736, 4.903), that is to say, we’re 95% confident that the true difference of `mpg` lies in this interval.

```
##                Estimate  Std. Error   t value     Pr(>|t|)
## (Intercept) 34.00287512 2.642659337 12.866916 2.824030e-13
## hp          -0.03747873 0.009605422 -3.901830 5.464023e-04
## wt          -2.87857541 0.904970538 -3.180850 3.574031e-03
## am1          2.08371013 1.376420152  1.513862 1.412682e-01
```

Diagnose

To evaluate the performance of the selected model, we make the residual plots(refer to appendix). We can see a little curve in Residuals vs. Fitted plot.

The Normal Q-Q plot indicates a little departure from normality on two ends. The residual for the `"Toyota Corolla"` , `"Fiat 128"` " and `"Chrysler Imperial"` are called out, because they exert some influence on the shape of the line.

In the Residuals vs. Leverage plot, there are no outliers beyond the 0.5 bands, but the above three points again lie much nearer to the 0.5 band than others.

To further examine whether these three points are outliers, we calculate the `hatvalues` and `dfbetas` . The hatvalues of `"Toyota Corolla"` , `"Fiat 128"` " and `"Chrysler Imperial"` is 0.107, 0.111, 0.230 respectively, much below the maxium value 0.412.

However, the absolute `dfbetas` value of `"Chrysler Imperial"` for the intercept and `wt` are both above 0.9, which means this point has particularly large influence.

Conclusion

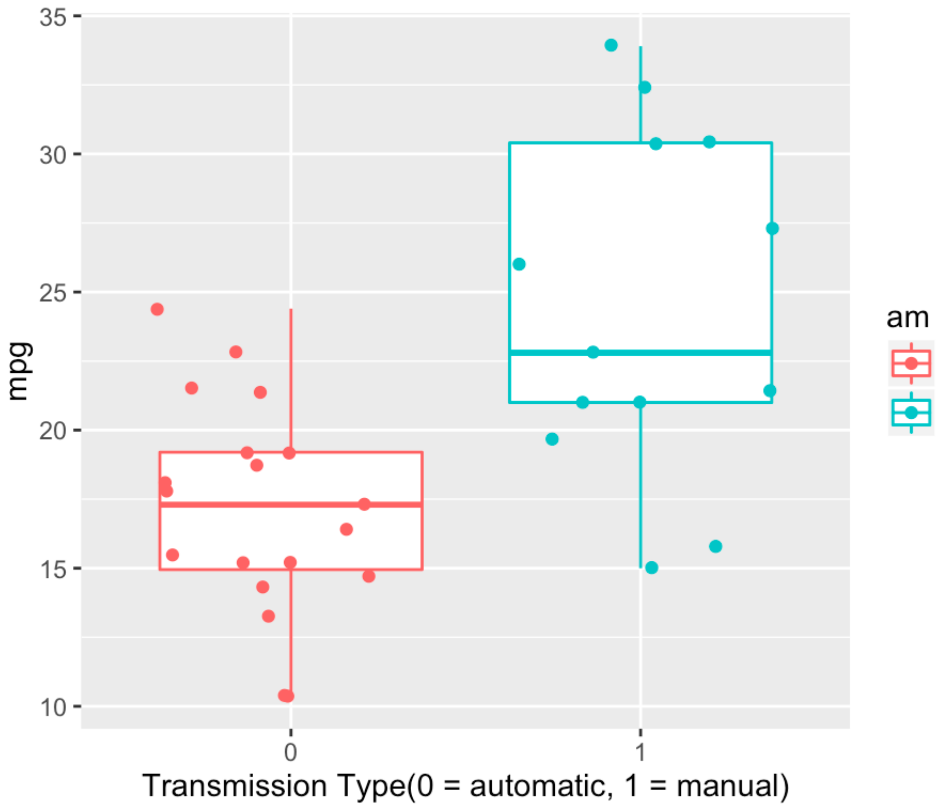
With the analysis above, we’re ready to answer the questions:

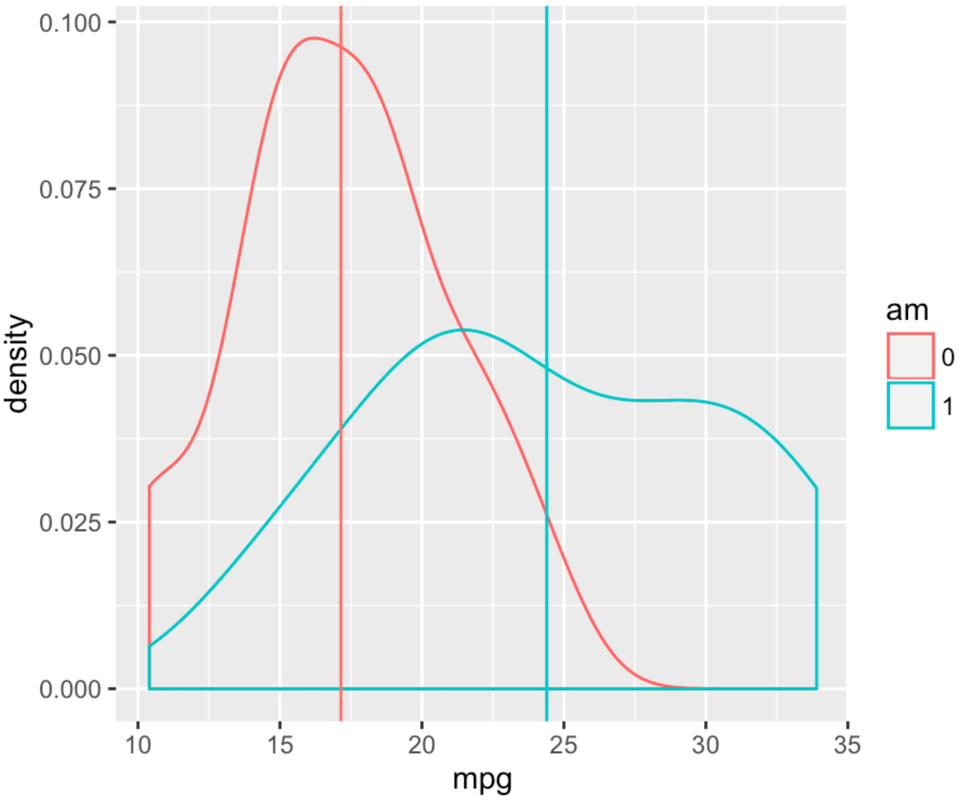
1. Manual transmission is better for mpg than automatic transition, that is, manual cars have higher mpg value.
2. There is an estimated decline of 2.084 miles per gallon when shifting from manual to automatic transition. We’re 95% confident that the true dicline lies in the interval (-0.736, 4.903).

Appendix

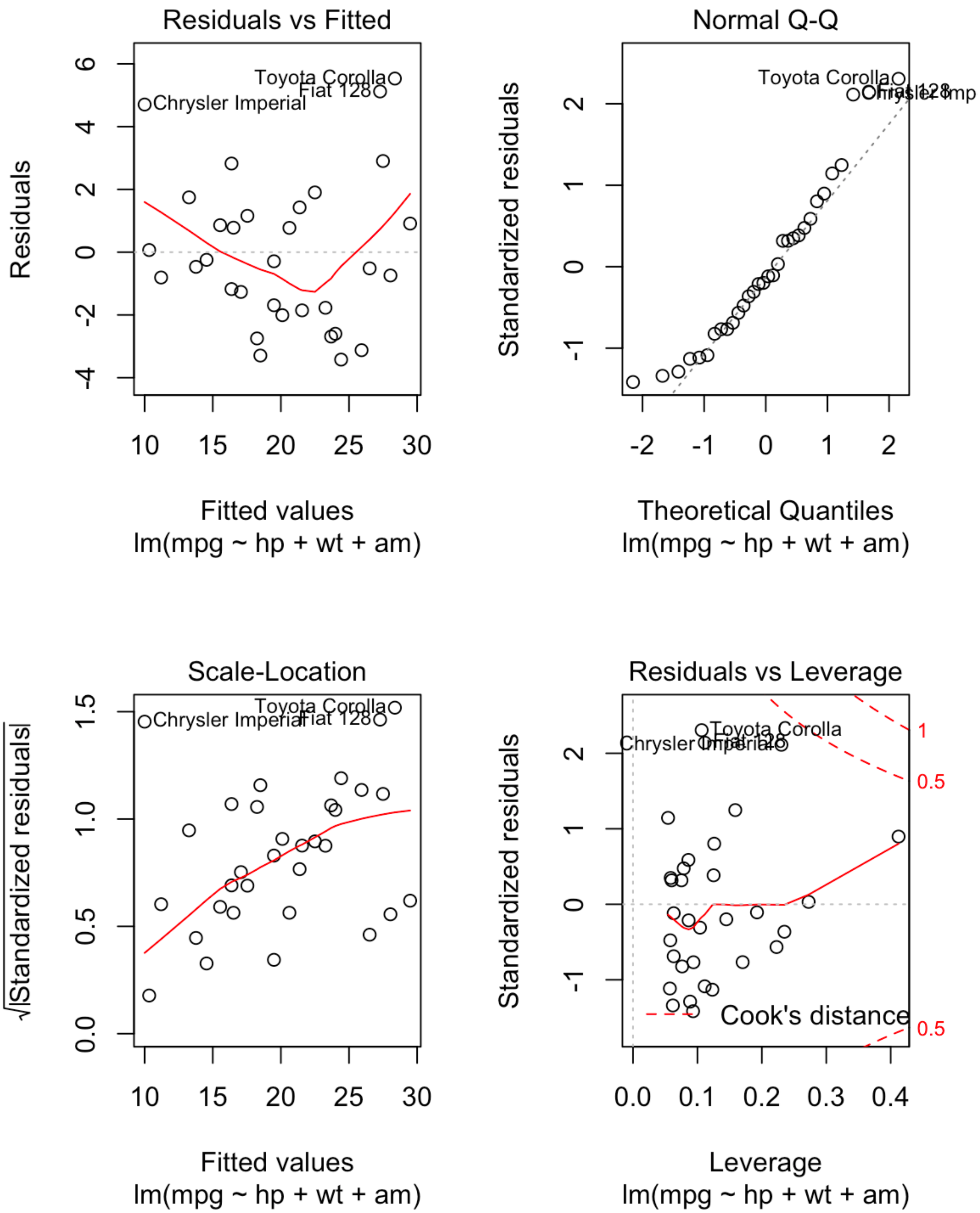
1. Explorary Plots

Two vertical lines in the density plot show the mean of mpg for each transmission.





2. Residual Plot


Residuals vs Fitted
Residuals
-4
-2
0
2
4
6
Toyota Corolla
Fiat 128
Chrysler Imperial
10
15
20
25
30
Fitted values
lm(mpg ~ hp + wt + am)
Normal Q-Q
Standardized residuals
-1
0
1
2
Toyota Corolla
Fiat 128
Chrysler Imp
-2
-1
0
1
2
Theoretical Quantiles
lm(mpg ~ hp + wt + am)
Scale-Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
Toyota Corolla
Chrysler Imperial
Fiat 128
10
15
20
25
30
Fitted values
lm(mpg ~ hp + wt + am)
Residuals vs Leverage
Standardized residuals
-1
0
1
2
Toyota Corolla
Chrysler Imperial
Fiat 128
1
0.5
Cook's distance
0.5
0.0
0.1
0.2
0.3
0.4
Leverage
lm(mpg ~ hp + wt + am)