Linear Regression

Jurgen Tas

15 Jul 2015

Executive Summary

In this project, the Motor Trend Car Road Tests (mtcars) dataset is analyzed. The data, extracted from the 1974 Motor Trend US magazine, comprises fuel consumption and 10 aspects of automobile design and performance for 32 automobiles. The goal is to explore the relationship between a set of variables and miles per gallon (MPG) (outcome). Two questions are adressed:

1) Is an automatic or manual transmission better for MPG?
2) Can we quantify the MPG difference between automatic and manual transmissions?

Data Exploration

First, the mtcars dataset and the library we need are loaded.

```
data(mtcars)
library(car)
```

The structure of the mtcars dataset is listed by using:

```
str(mtcars)
```

```
## '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  : num  0 0 1 1 0 1 0 1 1 1 ...
##  $ am  : num  1 1 1 0 0 0 0 0 0 0 ...
##  $ gear: num  4 4 4 3 3 3 3 4 4 4 ...
##  $ carb: num  4 4 1 1 2 1 4 2 2 4 ...
```

To investigate the impact of am on mpg, a box plot is made (see Appendix). This type of plot shows the distribution of the mpg variable for both transmission types. We observe that median mpg for automatic and manual transmissions differ. The median mileage is higher for manual transmissions.

Next, consider the correlations among the variables in the dataset. Mileage is strongly correlated ($|corr| > 80\%$) with wt, cyl and disp. These variables are also strongly positevely correlated to each other. Strong correlations among predictor variables leads to confounders in the regression.

```
round(cor(mtcars), 2)
```

```
##        mpg   cyl  disp    hp  drat    wt  qsec    vs    am  gear  carb
## mpg   1.00 -0.85 -0.85 -0.78  0.68 -0.87  0.42  0.66  0.60  0.48 -0.55
## cyl  -0.85  1.00  0.90  0.83 -0.70  0.78 -0.59 -0.81 -0.52 -0.49  0.53
## disp -0.85  0.90  1.00  0.79 -0.71  0.89 -0.43 -0.71 -0.59 -0.56  0.39
## hp   -0.78  0.83  0.79  1.00 -0.45  0.66 -0.71 -0.72 -0.24 -0.13  0.75
## drat  0.68 -0.70 -0.71 -0.45  1.00 -0.71  0.09  0.44  0.71  0.70 -0.09
## wt   -0.87  0.78  0.89  0.66 -0.71  1.00 -0.17 -0.55 -0.69 -0.58  0.43
## qsec  0.42 -0.59 -0.43 -0.71  0.09 -0.17  1.00  0.74 -0.23 -0.21 -0.66
## vs    0.66 -0.81 -0.71 -0.72  0.44 -0.55  0.74  1.00  0.17  0.21 -0.57
## am    0.60 -0.52 -0.59 -0.24  0.71 -0.69 -0.23  0.17  1.00  0.79  0.06
## gear  0.48 -0.49 -0.56 -0.13  0.70 -0.58 -0.21  0.21  0.79  1.00  0.27
## carb -0.55  0.53  0.39  0.75 -0.09  0.43 -0.66 -0.57  0.06  0.27  1.00
```

Model Selection

Three nested linear regressions models using wt and am as predictor variables and mpg as the outcome variable are build. The (nested) models are compared using the anova() function in R. The first model only takes am as predictor variable. We compare this model with model 2, taking am and wt as predictor variables. The last model takes the interaction between am and wt into consideration.

```
mtcars$am = factor(mtcars$am,labels=c('Auto','Man'))
mtcars$cyl = factor(mtcars$cyl)
model1 = lm(mpg ~ am, data = mtcars)
model2 = lm(mpg ~ am + wt, data = mtcars)
model3 = lm(mpg ~ am + wt + am:wt, data = mtcars)
anova(model1, model2, model3)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ am + wt
## Model 3: mpg ~ am + wt + am:wt
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     30 720.90
## 2     29 278.32  1    442.58 65.913 7.717e-09 ***
## 3     28 188.01  1     90.31 13.450  0.001017 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Using the third model leads to significant improvement compared to using the first two models. We decide to base our conclusions on model 1 and model 3. Residual plots are shown in the appendix of this document.

Conclusions

```
summary(model1)
```

```
##
## Call:
## lm(formula = mpg ~ 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 ***
## amMan          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
```

```
confint(model1)
```

```
##                2.5 %   97.5 %
## (Intercept) 14.85062 19.44411
## amMan        3.64151 10.84837
```

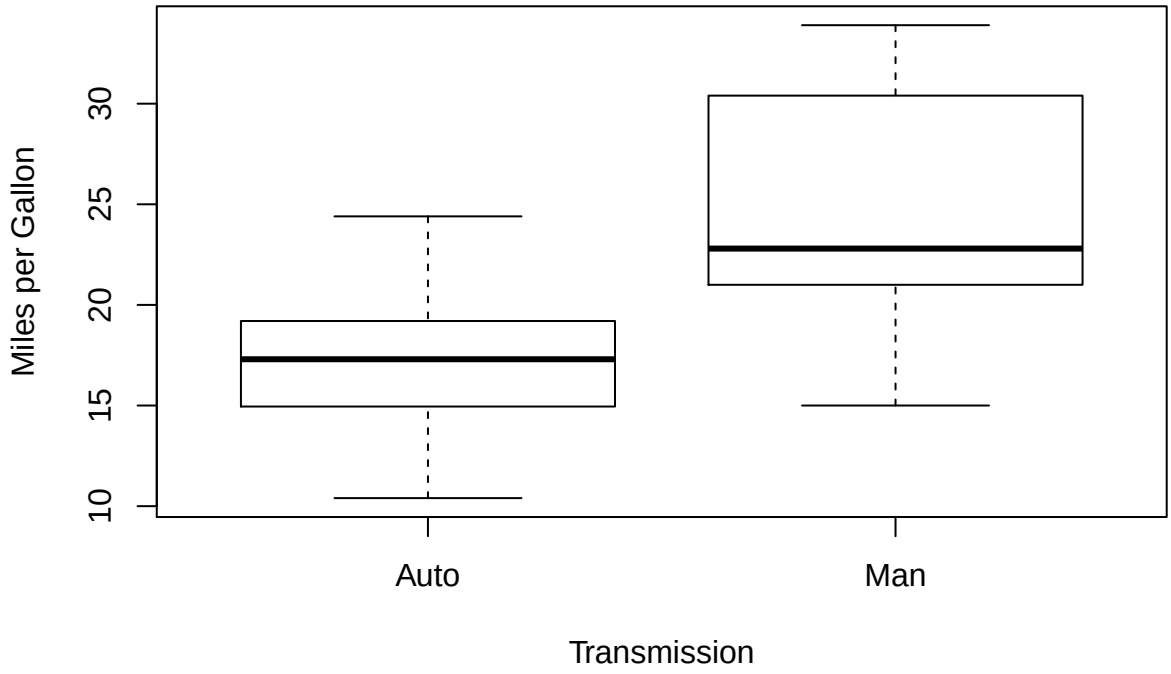
```
summary(model3)
```

```
##
## Call:
## lm(formula = mpg ~ am + wt + am:wt, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.6004 -1.5446 -0.5325  0.9012  6.0909
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  31.4161     3.0201  10.402 4.00e-11 ***
## amMan        14.8784     4.2640   3.489  0.00162 **
## wt           -3.7859     0.7856  -4.819 4.55e-05 ***
## amMan:wt     -5.2984     1.4447  -3.667  0.00102 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.591 on 28 degrees of freedom
## Multiple R-squared:  0.833,  Adjusted R-squared:  0.8151
## F-statistic: 46.57 on 3 and 28 DF,  p-value: 5.209e-11
```

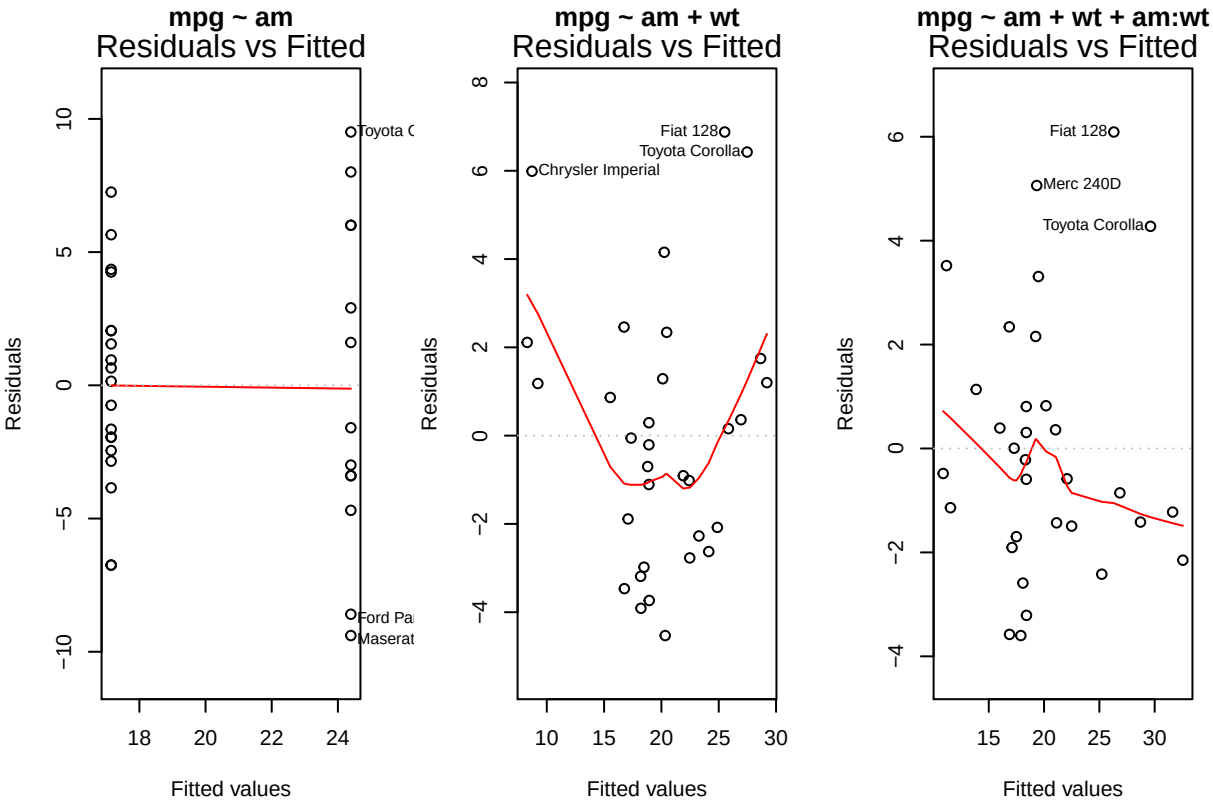
Model 1 explains 36% of the variance, model 4 explains 83% of the variance. The beta values, for both models, are statistically significant. Based on model 1, we estimate with 95% confidence that the mpg difference between automatic and manual transmission cars is between 3.6 and 10.8.

But, if we look deeper into this matter, we observe that if wt increases, the drop in mpg is steeper for manual transmission cars; i.e. see Appendix. This effect is captured by model 3. Intuitively this makes sense: a heavier car requires more fuel. For manual transmission cars, the mpg drops 9.1 if weight increases with 1000lbs. For automatic transmission cars, the mpg drops 3.8 if weight increases with 1000lbs.

Appendix: Boxplots



Appendix: Regression diagnostics



Appendix: Interaction plot