Regression Models Course Project

Sangwon Han

2018-12-27

Executive Summary

In this project, I explore the relationship between a set of variables and miles per gallon (MPG) (outcome). I am particularly interested in the following two questions:

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

Exploratory analysis

Load data, and look at it. Supplemental figure (A1 on Appendix) shows scatter plots of data

```r
data(mtcars)
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 ...
```

Transform the numeric variables into factor variables

```r
mtcars$cyl  <- factor(mtcars$cyl)
mtcars$vs   <- factor(mtcars$vs)
mtcars$am   <- factor(mtcars$am,labels=c("automatic","manual"))
mtcars$gear <- factor(mtcars$gear)
mtcars$carb <- factor(mtcars$carb)
```

Initially, we compare tha mpg values between automatic and manual groups by means of linear regression with dummy variable.

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

## 
## 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 ***
## ammanual       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
```

The results showed that average MPG for automatic group is 17.147, while manual group has a mean value of 7.245 higher than automatic group (Box and whisker plot of mpg vs transmission is available in A2 of Appendix section). The difference is statistically significant (p = 0.000285)

However, this results did not consider the confounding effect. To deal with confounders, multivariate regression analysis is required.

Regression analysis

At first, I built linear regression models including all different variables. The model selection is based on stepwise selection using both forward selection and backward elimination method.

```
multi_model <- lm(mpg ~ ., data = mtcars)
best_multi <- step(multi_model, direction = "both")

## Start:  AIC=76.4
## mpg ~ cyl + disp + hp + drat + wt + qsec + vs + am + gear + carb
## 
##        Df Sum of Sq    RSS    AIC
## - carb  5   13.5989 134.00 69.828
## - gear  2    3.9729 124.38 73.442
## - am    1    1.1420 121.55 74.705
## - qsec  1    1.2413 121.64 74.732
## - drat  1    1.8208 122.22 74.884
## - cyl   2   10.9314 131.33 75.184
## - vs    1    3.6299 124.03 75.354
## <none>              120.40 76.403
## - disp  1    9.9672 130.37 76.948
## - wt    1   25.5541 145.96 80.562
## - hp    1   25.6715 146.07 80.588
##
## Step:  AIC=69.83
## mpg ~ cyl + disp + hp + drat + wt + qsec + vs + am + gear
##
##        Df Sum of Sq    RSS    AIC
## - gear  2    5.0215 139.02 67.005
## - disp  1    0.9934 135.00 68.064
## - drat  1    1.1854 135.19 68.110
## - vs    1    3.6763 137.68 68.694
## - cyl   2   12.5642 146.57 68.696
## - qsec  1    5.2634 139.26 69.061
## <none>              134.00 69.828
## - am    1   11.9255 145.93 70.556
## - wt    1   19.7963 153.80 72.237
## - hp    1   22.7935 156.79 72.855
## + carb  5   13.5989 120.40 76.403
##
## Step:  AIC=67
## mpg ~ cyl + disp + hp + drat + wt + qsec + vs + am
##
##        Df Sum of Sq    RSS    AIC
## - drat  1    0.9672 139.99 65.227
## - cyl   2   10.4247 149.45 65.319
## - disp  1    1.5483 140.57 65.359
## - vs    1    2.1829 141.21 65.503
## - qsec  1    3.6324 142.66 65.830
## <none>              139.02 67.005
## - am    1   16.5665 155.59 68.608
## - hp    1   18.1768 157.20 68.937
## + gear  2    5.0215 134.00 69.828
## - wt    1   31.1896 170.21 71.482
## + carb  5   14.6475 124.38 73.442
##
## Step:  AIC=65.23
## mpg ~ cyl + disp + hp + wt + qsec + vs + am
##
##        Df Sum of Sq    RSS    AIC
## - disp  1    1.2474 141.24 63.511
## - vs    1    2.3403 142.33 63.757
## - cyl   2   12.3267 152.32 63.927
## - qsec  1    3.1000 143.09 63.928
## <none>              139.99 65.227
## + drat  1    0.9672 139.02 67.005
## - hp    1   17.7382 157.73 67.044
## - am    1   19.4660 159.46 67.393
## + gear  2    4.8033 135.19 68.110
## - wt    1   30.7151 170.71 69.574
## + carb  5   13.0509 126.94 72.095
##
## Step:  AIC=63.51
## mpg ~ cyl + hp + wt + qsec + vs + am
##
##        Df Sum of Sq    RSS    AIC
## - qsec  1     2.442 143.68 62.059
## - vs    1     2.744 143.98 62.126
## - cyl   2    18.580 159.82 63.466
## <none>              141.24 63.511
## + disp  1     1.247 139.99 65.227
## + drat  1     0.666 140.57 65.359
## - hp    1    18.184 159.42 65.386
## - am    1    18.885 160.12 65.527
## + gear  2     4.684 136.55 66.431
## - wt    1    39.645 180.88 69.428
## + carb  5     2.331 138.91 72.978
##
## Step:  AIC=62.06
## mpg ~ cyl + hp + wt + vs + am
##
##        Df Sum of Sq    RSS    AIC
## - vs    1     7.346 151.03 61.655
## <none>              143.68 62.059
## - cyl   2    25.284 168.96 63.246
## + qsec  1     2.442 141.24 63.511
## - am    1    16.443 160.12 63.527
## + disp  1     0.589 143.09 63.928
## + drat  1     0.330 143.35 63.986
## + gear  2     3.437 140.24 65.284
## - hp    1    36.344 180.02 67.275
## - wt    1    41.088 184.77 68.108
## + carb  5     3.480 140.20 71.275
##
## Step:  AIC=61.65
## mpg ~ cyl + hp + wt + am
##
##        Df Sum of Sq    RSS    AIC
## <none>              151.03 61.655
## - am    1     9.752 160.78 61.657
## + vs    1     7.346 143.68 62.059
## + qsec  1     7.044 143.98 62.126
## - cyl   2    29.265 180.29 63.323
## + disp  1     0.617 150.41 63.524
## + drat  1     0.220 150.81 63.608
## + gear  2     1.361 149.66 65.365
## - hp    1    31.943 182.97 65.794
## - wt    1    46.173 197.20 68.191
## + carb  5     5.633 145.39 70.438
```

The best multivariate model (having lowest AIC) consists cyl, hp, wt, and am as predictors.

```
summary(best_multi)

## 
## Call:
## lm(formula = mpg ~ cyl + hp + wt + am, data = mtcars)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -3.9387 -1.2560 -0.4013  1.1253  5.0513 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 33.70832    2.60489  12.940 7.73e-13 ***
## cyl6        -3.03134    1.40728  -2.154  0.04068 *  
## cyl8        -2.16368    2.28425  -0.947  0.35225    
## hp          -0.03211    0.01369  -2.345  0.02693 *  
## wt          -2.49683    0.88559  -2.819  0.00908 ** 
## ammanual     1.80921    1.39630   1.296  0.20646    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 2.41 on 26 degrees of freedom
## Multiple R-squared:  0.8659, Adjusted R-squared:  0.8401 
## F-statistic: 33.57 on 5 and 26 DF,  p-value: 1.506e-10
```

The results shows that after adjusting confounders(cyl, hp, and wt), manual group has a mean mpg value of being 1.80921 higher than automatic transmission group, however, the difference is not statistically significant (p = 0.20646)

The adjusted R-squared value of the best multivariate model is 0.84, wherease that of initial single regression model is 0.34. The anova analysis shows that adding cyl, hp, and wt contribute to enhance the model fit significantly.

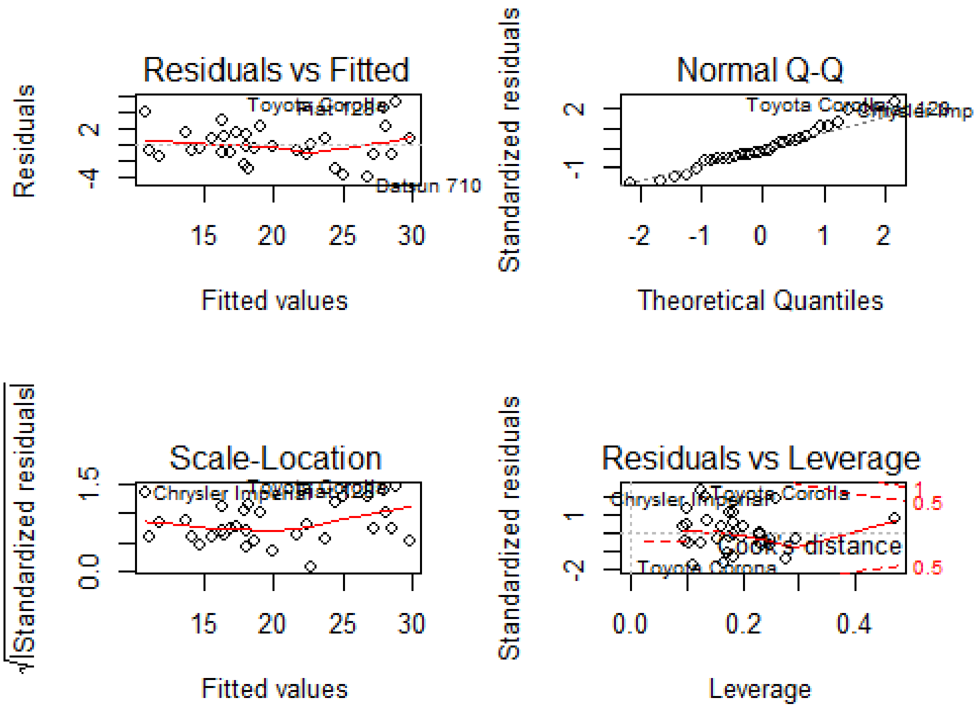
```
anova(lm_am, best_multi)

## Analysis of Variance Table
## 
## Model 1: mpg ~ am
## Model 2: mpg ~ cyl + hp + wt + am
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)    
## 1     30 720.90                                  
## 2     26 151.03  4    569.87 24.527 1.688e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Residual and diagnositics

Make residual plots for the best multivariate model to examine any heteroskedacity or non-normality

```
par(mfrow = c(2,2))
plot(best_multi)
```



No significant heteroskedacity or non-normality was found.

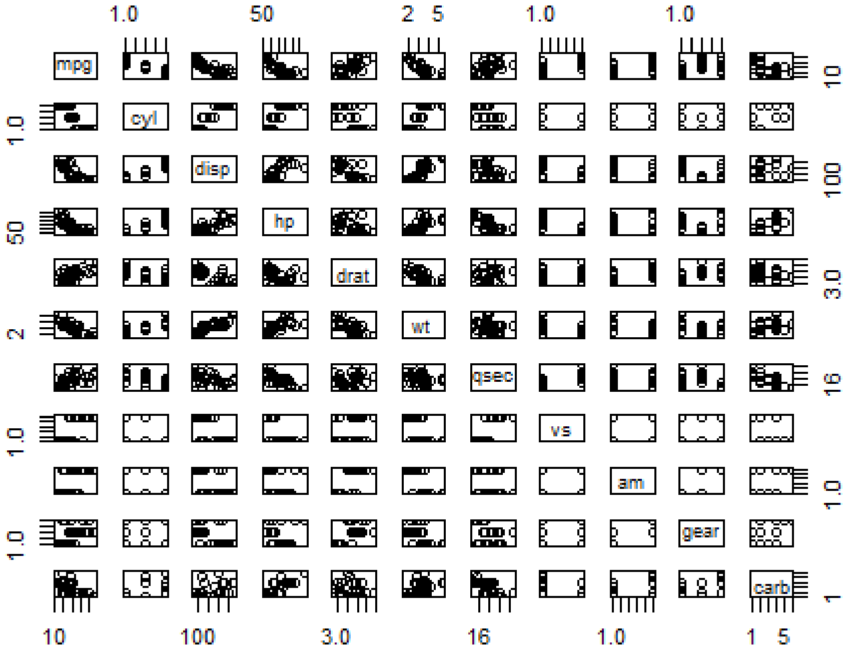
Conclusions

- Cars having manual transmission has a mean miles per gallon of being 7.245 higher than those having automatic transmission.
- However, after considering the influences of horsepower, number of cylinders and weight. group, the difference was decreased into 1.809 and no longer statistically significant

Appendix

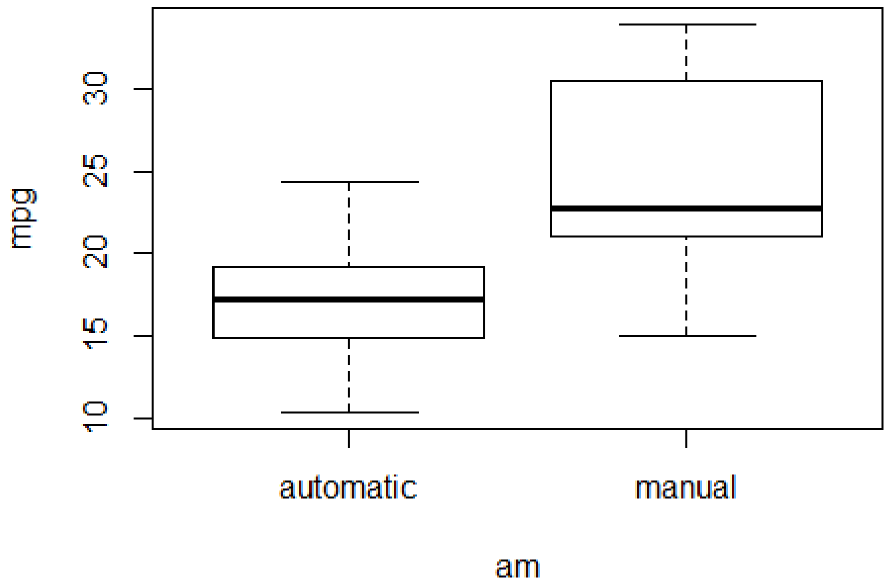
A1. Scatter plot matrix for mtcars dataset

```
pairs(mpg~., data = mtcars)
```



A2. Box and whisker plot of mpg vs transmission

```
plot(mpg ~ am, data = mtcars)
```


mpg
10
15
20
25
30
automatic
manual
am