Regression model course project

Fan Ouyang

Executive Summary

In this report, I will examine the mtcars data set and explore how miles per gallon (MPG) is affected by different variables. In particularly, I will answer the following two questions: (1) Is an automatic or manual transmission better for MPG, and (2) Quantify the MPG difference between automatic and manual transmissions.

Parameters of this dataset–mpg : Miles/(US) gallon, cyl : Number of cylinders, disp : Displacement (cu.in.), hp : Gross horsepower, drat : Rear axle ratio, wt : Weight (lb/1000), qsec : 1/4 mile time, vs : V/S, am : Transmission (0 = automatic, 1 = manual), gear : Number of forward gears, carb : Number of carburetors

Results: Manual transmission has an MPG 1.8 greater than an automatic transmission.

Exploratory Analysis

```
library(ggplot2)
library(datasets)
data(mtcars)
```

Correlation matrix Analysis

I explore various relationships between variables of interest in the dataset.

```
res <- cor(mtcars)
round(res, 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
```

```
symnum(res, abbr.colnames = FALSE)
```

```
##      mpg cyl disp hp drat wt qsec vs am gear carb
## mpg  1
## cyl  +   1
## disp +   *   1
## hp   ,   +   ,    1
## drat ,   ,   ,    .  1
## wt   +   ,   +    ,  ,    1
## qsec .   .   .    ,          1
```

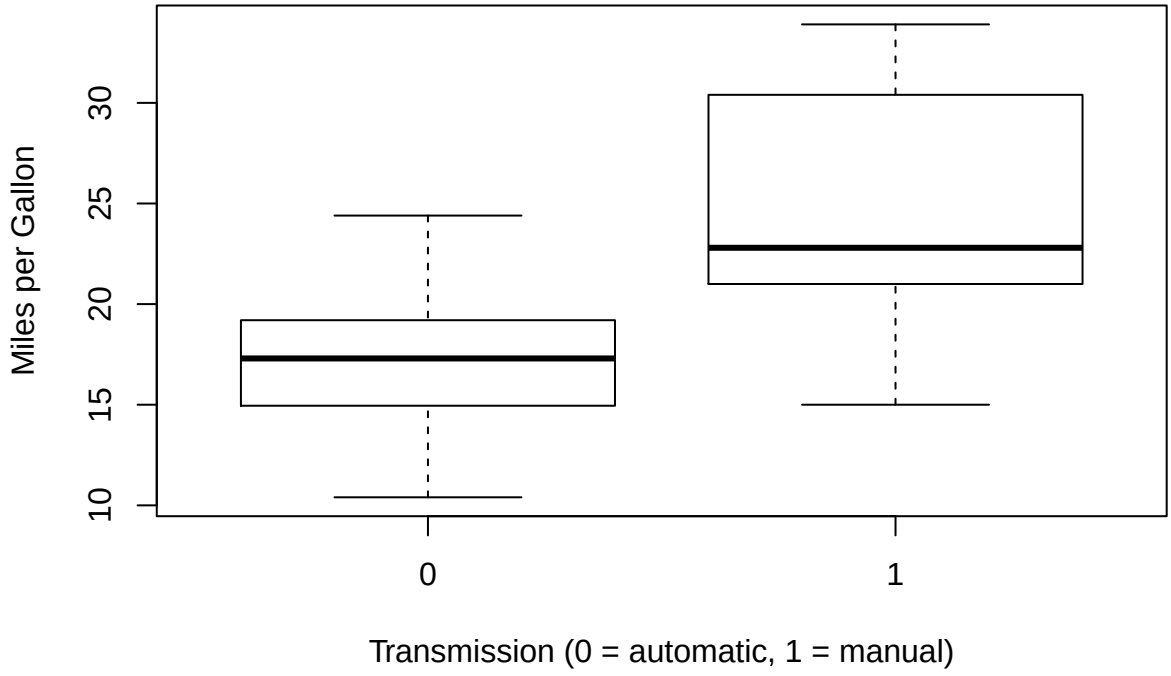
```
## vs   ,   +   ,    ,  .    .  ,    1
## am   .   .   .       ,    ,          1
## gear .   .   .       ,    .          ,  1
## carb .   .   .    ,       .  ,    .          1
## attr(,"legend")
## [1] 0 ' ' 0.3 '.' 0.6 ',' 0.8 '+' 0.9 '*' 0.95 'B' 1
```

The correlation table shows that cyl, disp, hp, drat, wt, vs and am have a strong correlation with mpg. Further summary shows that cyl, disp, wt have the most strongest correlation with mpg.

Question 1: Is an automatic or manual transmission better for MPG?

Let's take a look at the consumption distribution (or MPG distribution) by transmission system :

```
mtcars$vs <- as.factor(mtcars$vs)
mtcars$am <- as.factor(mtcars$am)
boxplot(mpg ~ am, data = mtcars, xlab = "Transmission (0 = automatic, 1 = manual)", ylab="Miles per Gal
```



There is a difference between automatic and manual distribution systems. Manual distribution systems are mostly less fuel consuming.

Question 2 : Quantify the MPG difference between automatic and manual transmissions.

Regression analysis

I build linear regression models using different variables in order to find the best fit and compare it with the base model.

```
initialmodel <- lm(mpg ~ ., data = mtcars)
basemodel <- lm(mpg ~ am, data = mtcars)
bestmodel<-lm(mpg ~ cyl+ disp+ wt + am, data = mtcars)
summary(bestmodel)
```

```
##
## Call:
## lm(formula = mpg ~ cyl + disp + wt + am, data = mtcars)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -4.318 -1.362 -0.479  1.354  6.059
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 40.898313   3.601540  11.356 8.68e-12 ***
## cyl         -1.784173   0.618192  -2.886  0.00758 **
## disp         0.007404   0.012081   0.613  0.54509
## wt          -3.583425   1.186504  -3.020  0.00547 **
## am1          0.129066   1.321512   0.098  0.92292
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.642 on 27 degrees of freedom
## Multiple R-squared:  0.8327, Adjusted R-squared:  0.8079
## F-statistic: 33.59 on 4 and 27 DF,  p-value: 4.038e-10
```

```
anova(basemodel, bestmodel)
```

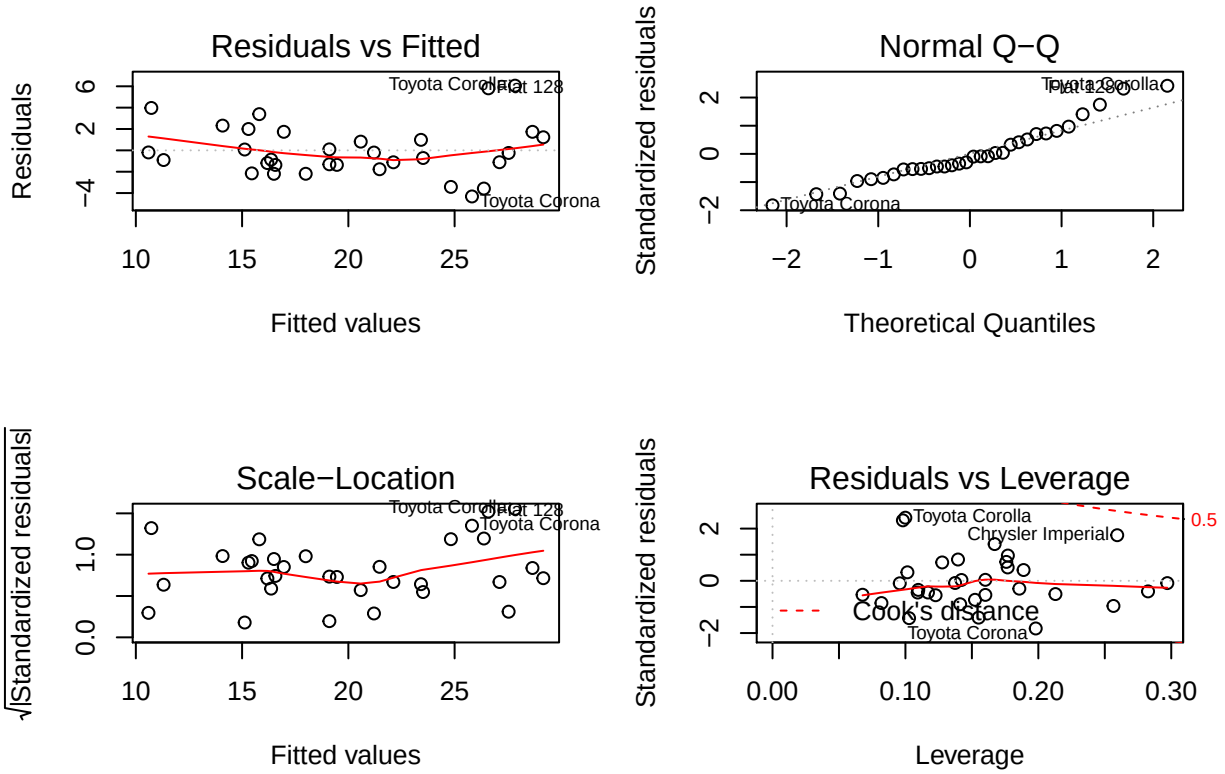
```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ cyl + disp + wt + am
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     30 720.90
## 2     27 188.43  3    532.47 25.433 5.034e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Results show the p-value obtained is highly significant and we reject the null hypothesis that the confounder variables cyl, disp and wt don't contribute to the accuracy of the model. After computation our best model get an R-squared adjusted of 0.83. wt, hp and cyl are the variables that best explain miles per gallon consumption if we look at the asterix marks.

Model Residuals and Diagnostics

In this section, we have the residual plots of our regression model along with computation of regression diagnostics for our liner model. This excercise helped us in examining the residuals and finding leverage points to find any potential problems with the model.

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



In the residual plot, I don't see any pattern that causes us to believe that the Fuel consumption could be explained more by any other predictors available in the dataset.

Statistical Inference

Two samples t-test on the MPG parameter distinguishing auto vs manual systems. Null hypothesis: There is no difference between MPG means for automatic and manual transmissions.

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

```
##
##  Welch Two Sample t-test
##
## data:  mpg by am
## t = -3.7671, df = 18.332, p-value = 0.001374
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -11.280194  -3.209684
## sample estimates:
## mean in group 0 mean in group 1
##        17.14737        24.39231
```

Based on the t-test results, I reject the null hypothesis that the mpg distributions for manual and automatic transmissions are the same. So there are statistically significance between MPG means for automatic and manual transmissions. Manual transmission has an MPG 1.8 greater than an automatic transmission.

Conclusion

Using simple linear regression analysis, I conclude that there is a signficant difference between the mean MPG for automatic and manual transmission cars. Manual transmissions achieve a higher value of MPG compared to automatic transmission. This increase is approximately 1.8 MPG when switching from an automatic transmission to a manual one, with all other parameters held constant.