Regression Models Project - Cars

Bowen Zhang

Summary

In this project, we will analyze the mtcars dataset to answer the following questions:

- Is an automatic or manual transmission better for miles per gallon (MPG)?
- How different is the MPG between automatic and manual transmissions?

With inear regression we find that manual cars have a significant higher average MPG. However as we include more significant variables into our model, we see the difference diminishes almost entirely as as the other variables can account for most of the effect in mpg.

Loading our Data

We will be using the mtcars dataset to perform this analysis. Here is a preview:

```
data(mtcars)
head(mtcars, 3)
```

```
##                mpg cyl disp  hp drat    wt  qsec vs am gear carb
## Mazda RX4     21.0   6  160 110 3.90 2.620 16.46  0  1    4    4
## Mazda RX4 Wag 21.0   6  160 110 3.90 2.875 17.02  0  1    4    4
## Datsun 710    22.8   4  108  93 3.85 2.320 18.61  1  1    4    1
```

Exploratory Data Analysis

Let's take a quick look at the average MPG of automatic vs manual transmission. We have also visualized the difference on a boxplot (Appendix A).

```
# 1 is manual, 0 is automatic
round(aggregate(mpg~am, data = mtcars, mean),1)
```

```
##   am  mpg
## 1  0 17.1
## 2  1 24.4
```

T-Test (Appendix B): From this we can test for a significant difference in means between MPG for manual and automatic cars. From a quick **t-test** (Appendix B), we find that the **p-value = 0.001374 and the confidence interval at 95% is -11.280194 to -3.209684.**

Findings: From the t-test we see that the confidence interval does not include zero and the p-value of 0.001374 is less than 0.05. This allows us to accept the alternate hypothesis that there is a significant difference in means between manual and automatic MPG. Our next step is to further quantify this through regression analysis.

Regression Analysis

Linear Regression

Regression with two variables: mpg and am

```r
fit1 <- lm(mpg ~ factor(am), data=mtcars)
summary(fit1)$coef
```

```
##              Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 17.147368   1.124603 15.247492 1.133983e-15
## factor(am)1  7.244939   1.764422  4.106127 2.850207e-04
```

```r
summary(fit1)$r.squared
```

```
## [1] 0.3597989
```

We do find that the p-values are low, however the R-squared is only 0.3598. This suggests we should try this with more variables.

Analysis of Variance (Appendix C)

From Appendix C summary table, we see that the **significant variables with p-values less than 0.05 are cyl**, disp, and wt**. We can now include these variables in a multivariate regression. We also checked these for normality and Homoscedasticity - so we can trust these p-values (Appendix D).

Multivariate Regression

See plot (Appendix E) for reference.

```r
fit2 <- lm(mpg~ am + wt + cyl + disp, data = mtcars)
summary(fit2)$coef
```

```
##                 Estimate Std. Error    t value     Pr(>|t|)
## (Intercept) 40.898313414 3.60154037 11.3557837 8.677574e-12
## am           0.129065571 1.32151163  0.0976651 9.229196e-01
## wt          -3.583425472 1.18650433 -3.0201537 5.468412e-03
## cyl         -1.784173258 0.61819218 -2.8861142 7.581533e-03
## disp         0.007403833 0.01208067  0.6128661 5.450930e-01
```

```r
summary(fit2)$r.squared
```

```
## [1] 0.8326661
```

Findings: Our R-Squared is now 0.8327, meaning our model can explain for 83% of the variance. From the multivariate model, we accounted for the confounding variables cyl and wt and saw that the difference between manual and automatic transmissions is about **0.13 MPG**.

Conclusion

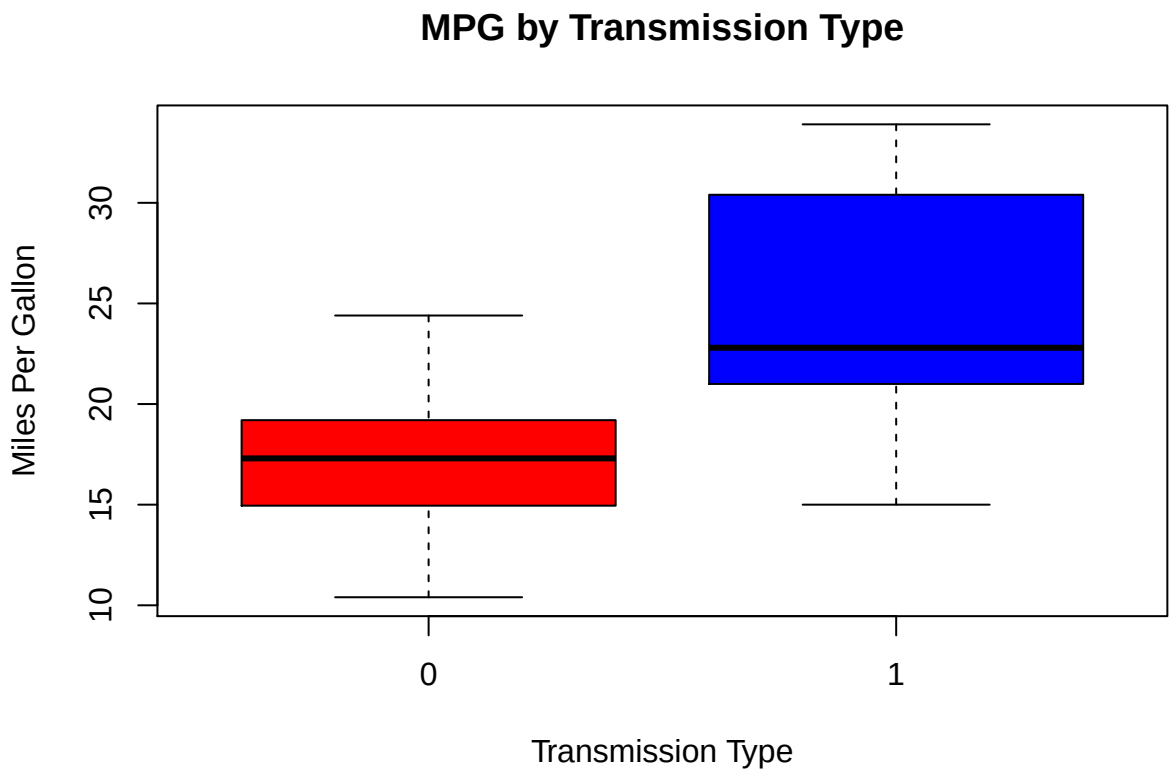
Based on just looking at the two variables transmission and mpg, we do see a significant difference, where manual cars have much higher miles per gallon. However when modeled with confounding variables like weight, cylinders, and displacement - the difference shrinks significantly to almost non-existent. This means that a big part of the difference is actually explained by the other confounding variables.

From our linear regression analysis, manual cars have a higher average by **7.245 mpg** over automatic cars. However, when accounting for other variables such as weight, cylinder, and displacement, the difference drops to **0.129 mpg**. This shows that other variables seem to be contributing more to the effect of mpg.

Appendix

A. Box plot of MPG by transmission type

```r
boxplot(mpg ~ am, data = mtcars, col = c("red", "blue"), ylab = "Miles Per Gallon",
        xlab = "Transmission Type", main = "MPG by Transmission Type")
```



B. T-Test

```
t.test(mtcars$mpg[mtcars$am==0], mtcars$mpg[mtcars$am==1])
```

```
##
##  Welch Two Sample t-test
##
## data:  mtcars$mpg[mtcars$am == 0] and mtcars$mpg[mtcars$am == 1]
## 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 of x mean of y
##  17.14737  24.39231
```

C. Analysis of variance summary

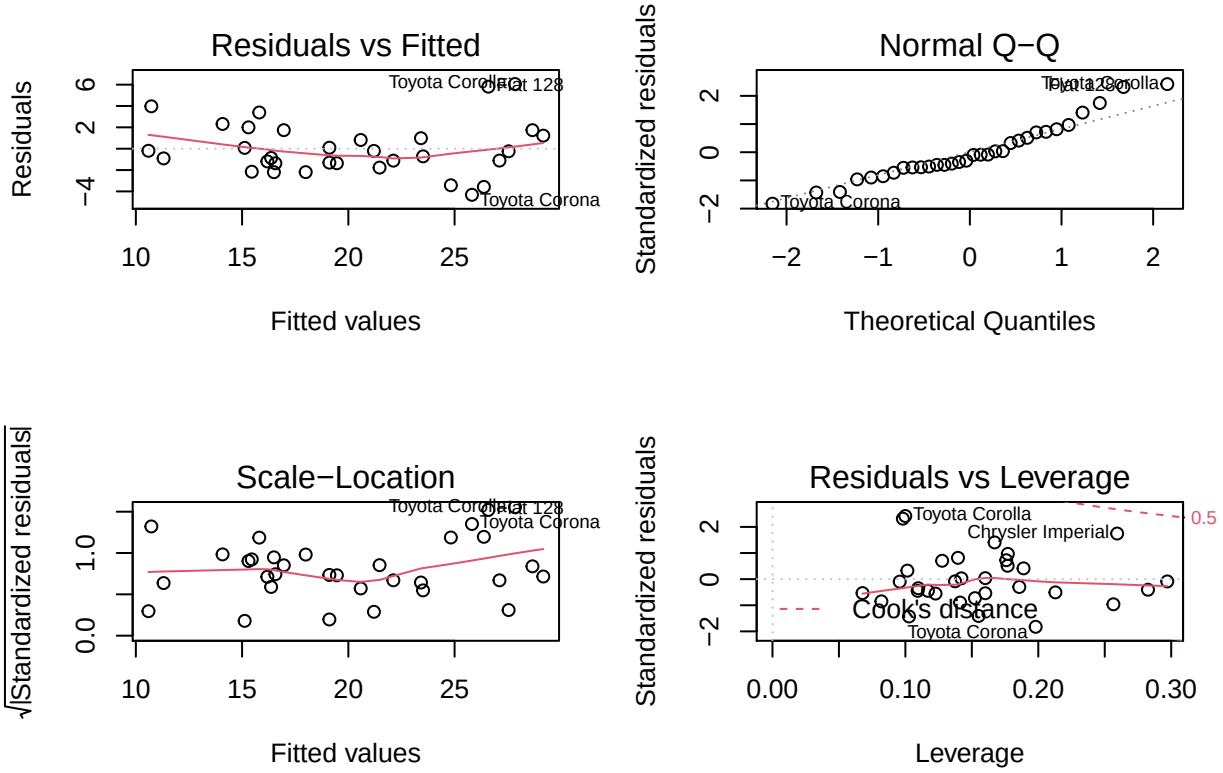
```
summary(aov(mpg~., data=mtcars))
```

```
##             Df Sum Sq Mean Sq F value   Pr(>F)
## cyl          1  817.7   817.7 116.425 5.03e-10 ***
## disp         1   37.6    37.6   5.353  0.03091 *
## hp           1    9.4     9.4   1.334  0.26103
## drat         1   16.5    16.5   2.345  0.14064
## wt           1   77.5    77.5  11.031  0.00324 **
## qsec         1    3.9     3.9   0.562  0.46166
## vs           1    0.1     0.1   0.018  0.89317
```

```
## am           1   14.5    14.5   2.061  0.16586
## gear         1    1.0     1.0   0.138  0.71365
## carb         1    0.4     0.4   0.058  0.81218
## Residuals   21  147.5     7.0
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

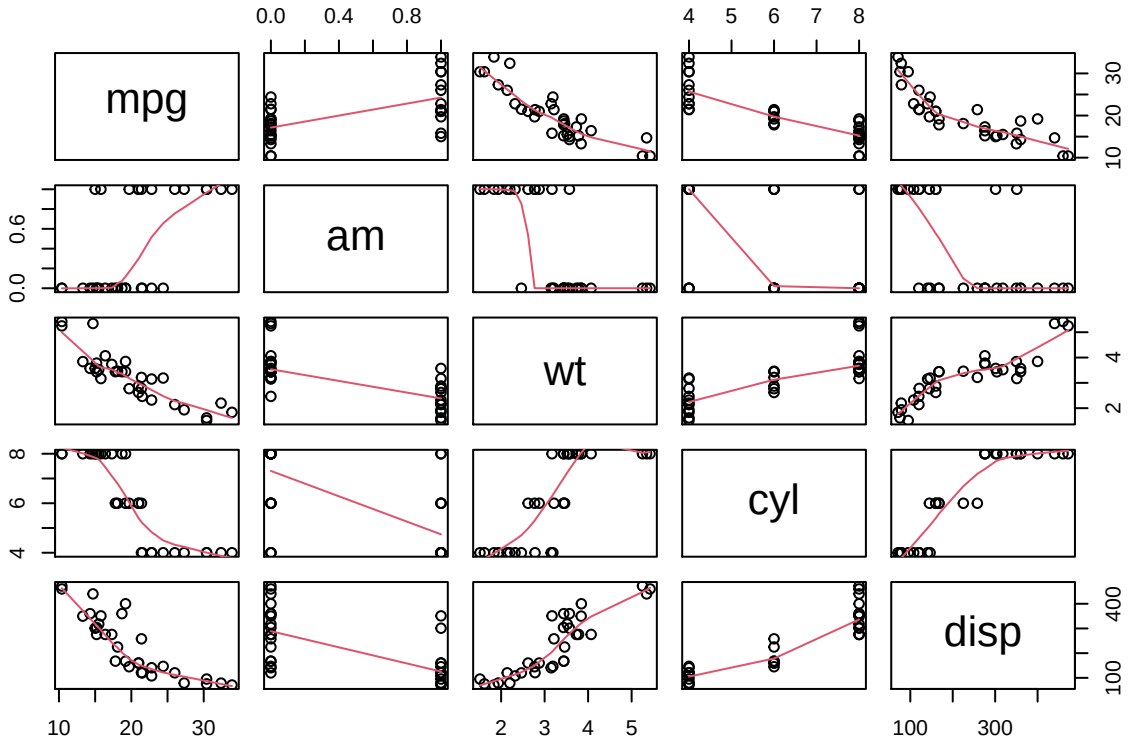
D. Residual Plot

```
par(mfrow = c(2,2))
plot(fit2)
```



E. Multivariate regression pairs

```
pairs(mpg ~ am + wt + cyl + disp, data = mtcars, panel = panel.smooth)
```


mpg
am
wt
cyl
disp
0.0
0.4
0.8
4
5
6
7
8
10
20
30
0.0
0.6
2
4
4
6
8
100
400
10
20
30
2
3
4
5
100
300