Regression Models Course Project

Ananya Mantravadi

04/12/2021

Executive Summary

Motor Trend is a magazine about the automobile industry. Looking at a data set of a collection of cars (`mtcars`), we are interested in exploring the relationship between a set of variables and miles per gallon (MPG) (outcome). We are 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

```
data("mtcars")
head(mtcars,1)
```

```
##           mpg cyl disp  hp drat   wt  qsec vs am gear carb
## Mazda RX4  21   6  160 110  3.9 2.62 16.46  0  1    4    4
```

Exploratory Data Analysis - Average and Correlation

```
aggregate(mpg ~ factor(am,labels=c("Automatic","Manual")), mtcars, mean)
```

```
##   factor(am, labels = c("Automatic", "Manual"))      mpg
## 1                                      Automatic 17.14737
## 2                                         Manual 24.39231
```

We can see that manual car has higher MPG.

```
round(cor(mtcars), 2)[1, ]
```

```
##   mpg   cyl  disp    hp  drat    wt  qsec    vs    am  gear  carb
##  1.00 -0.85 -0.85 -0.78  0.68 -0.87  0.42  0.66  0.60  0.48 -0.55
```

`cyl`, `disp`, `hp` and `wt` show high correlation.

Regression Analysis

Let us build a few linear models, in increasing complexity as follows:

```
model1 <- lm(mpg ~ factor(am) - 1, data = mtcars)
model2 <- update(model1, . ~ . + wt)
model3 <- update(model2, . ~ . + factor(cyl))
model4 <- update(model3, . ~ . + hp)
model5 <- update(model4, . ~ . + disp)
```

We can use ANOVA to understand how these models compare:

```
anova(model1, model2, model3,model4,model5)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ factor(am) - 1
## Model 2: mpg ~ factor(am) + wt - 1
## Model 3: mpg ~ factor(am) + wt + factor(cyl) - 1
## Model 4: mpg ~ factor(am) + wt + factor(cyl) + hp - 1
## Model 5: mpg ~ factor(am) + wt + factor(cyl) + hp + disp - 1
##   Res.Df    RSS Df Sum of Sq       F    Pr(>F)
## 1     30 720.90
## 2     29 278.32  1    442.58 73.5623 6.452e-09 ***
## 3     27 182.97  2     95.35  7.9244   0.00216 **
## 4     26 151.03  1     31.94  5.3093   0.02980 *
## 5     25 150.41  1      0.62  0.1025   0.75149
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

We can clearly observe that Model 4 performs the best. It includes the weight, cylinder type, and horsepower in addition to the transmission type. Adding displacement in Model 5 doesn’t give us significant improvement as it is highly correlated to the other regressors. The summary of this model is:

```
summary(model4)
```

```
##
## Call:
## lm(formula = mpg ~ factor(am) + wt + factor(cyl) + hp - 1, 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|)
## factor(am)0  33.70832    2.60489  12.940 7.73e-13 ***
## factor(am)1  35.51754    2.03171  17.482 6.81e-16 ***
## wt           -2.49683    0.88559  -2.819  0.00908 **
## factor(cyl)6 -3.03134    1.40728  -2.154  0.04068 *
## factor(cyl)8 -2.16368    2.28425  -0.947  0.35225
## hp           -0.03211    0.01369  -2.345  0.02693 *
## ---
## 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:  0.9892, Adjusted R-squared:  0.9868
## F-statistic: 398.6 on 6 and 26 DF,  p-value: < 2.2e-16
```

In this model, we can see that the cylinder type, horsepower, and the weight are all negatively correlated to the mileage. The R-squared value of this model tells us that it explains a very good 99% of the variance in `mpg`. Now let's see the confidence intervals for the predicted mileage for a hypothetical car:

```
# Predict
model4.prediction <- predict(model4, data.frame(am = c(0,1), wt = mean(mtcars$wt),
                     cyl = c(6), hp = mean(mtcars$hp)), interval="confidence")
print(model4.prediction)
```

```
##        fit      lwr      upr
## 1 17.93400 15.47757 20.39043
## 2 19.74321 17.32716 22.15927
```

Residual Plots and Diagnostics

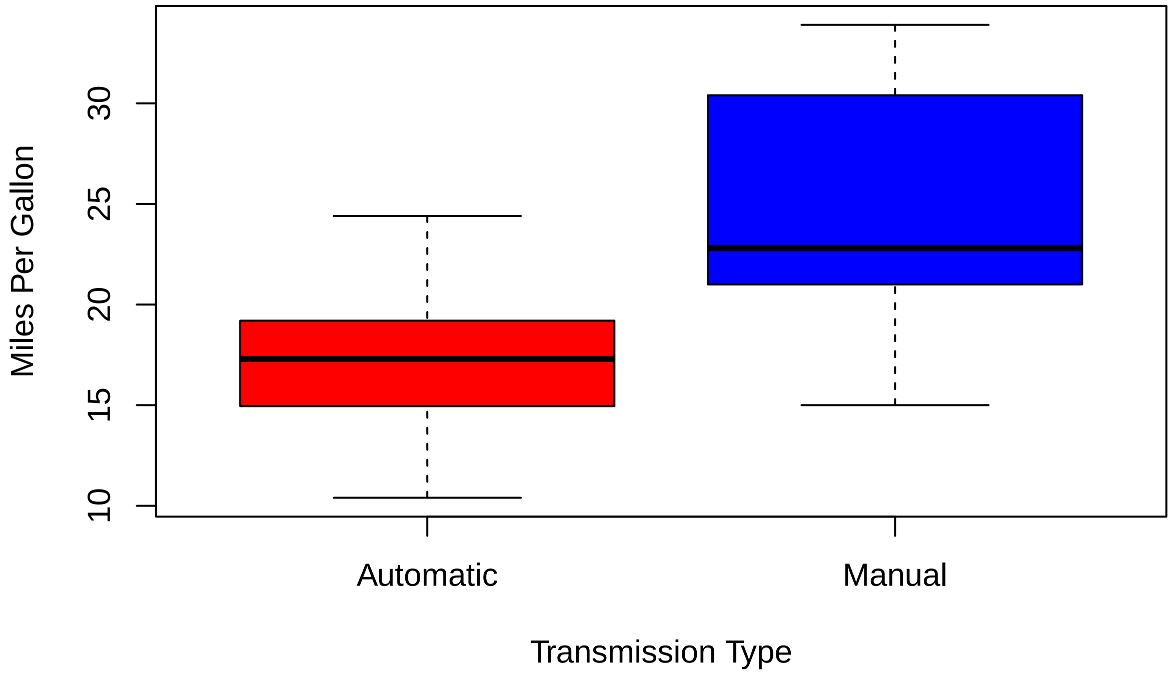
We now examine residual plots and compute regression diagnostics of our model (in the appendix) to identify any outliers in the data set. We can see that the points in the Residuals vs Fitted plot seem to be randomly scattered on the plot and verify the independence condition. The Normal Q-Q plot consists of the points which mostly fall on the line indicating that the residuals are normally distributed. The Scale-Location plot consists of points scattered in a constant band pattern, indicating constant variance.

Conclusions

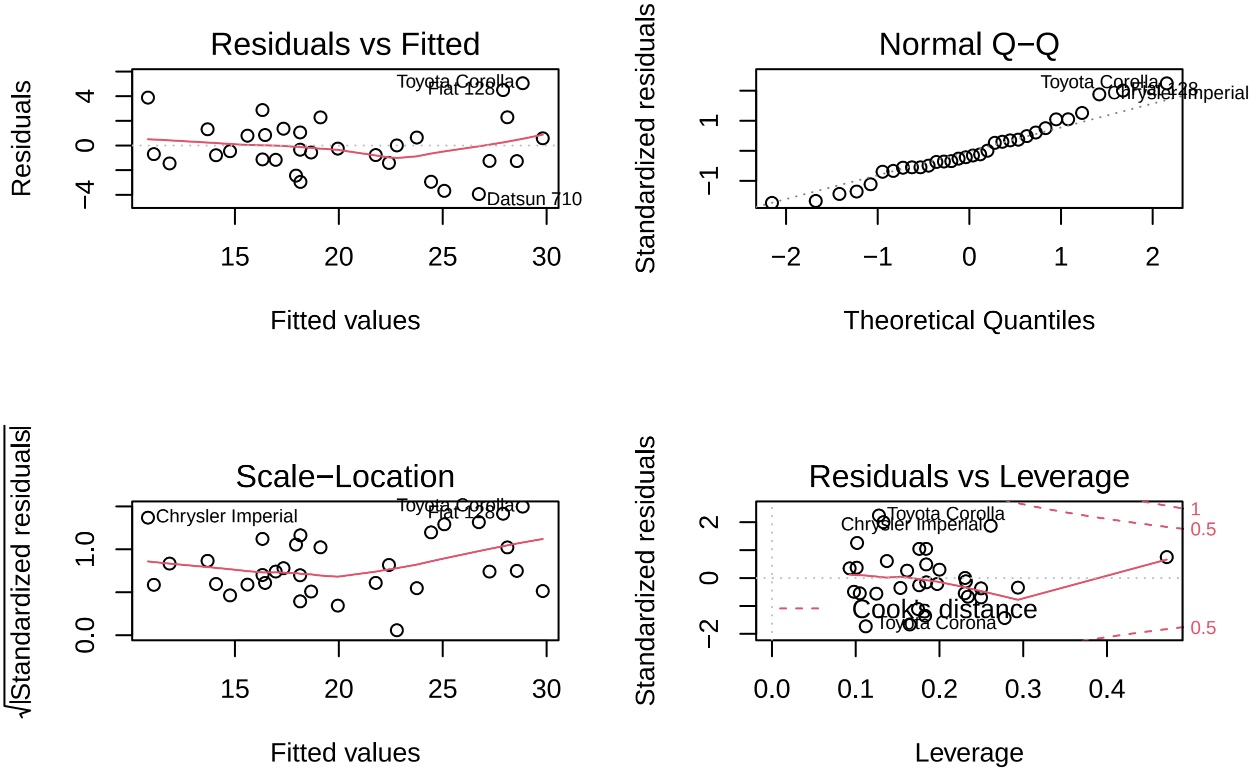
We observe that our hypothetical car having automatic transmission has an expected `mpg` of about 18 miles/gallon, and 20 miles/gallon if it has manual transmission. We find that manual transmission cars in this data set have better `mpg` by about 1.8 miles/gallon compared to automatic transmission cars. Although the manual transmission looks to have better mileage, we should note that the 95% confidence level intervals are not exclusive and is not statistically significant.

Appendix

```
boxplot(mpg ~ factor(am,labels=c("Automatic","Manual")), data = mtcars,
        col = (c("red","blue")), ylab = "Miles Per Gallon", xlab = "Transmission Type")
```



```
par(mfrow = c(2,2))
plot(model4)
```


Residuals vs Fitted
Residuals
Fitted values
Toyota Corolla
Fiat 128
Datsun 710
−4
0
4
15
20
25
30
Normal Q−Q
Standardized residuals
Theoretical Quantiles
Toyota Corolla
Fiat 128
Chrysler Imperial
−1
1
−2
−1
0
1
2
Scale−Location
√|Standardized residuals|
Fitted values
Chrysler Imperial
Toyota Corolla
Fiat 128
0.0
1.0
15
20
25
30
Residuals vs Leverage
Standardized residuals
Leverage
Toyota Corolla
Chrysler Imperial
Toyota Corona
Cook's distance
1
0.5
0.5
−2
0
2
0.0
0.1
0.2
0.3
0.4