Regression Assignment: Car Miles per Gallon Analysis

JB

October 26, 2017

Executive Summary

This assignment will evaluate the impact of different design and property variables of cars in the **mtcars** dataset on miles per gallon (mpg). The evaluation will assess the impact of transmission type (am -automatic vs. manual) and other variables on mpg with an exploratory analysis, regression and hypothesis testing.

Exploratory Analysis, Hypothesis Test and Regression

The plot of Figure 1 in the Appendix shows an inital assessment of mpg vs. transmission type. Manual mpg seems somewhat larger qualitatitively than automatic mpg. A two-tailed, two-sample student's t-test will be used to investigate this question further, comparing mpg for automatic vs. manual transmission.

```
car_data <- mtcars
t_test <-t.test(data=car_data, mpg~am)
```

```
t_test$p.value
```

```
## [1] 0.001373638
```

From this result, p = 0.001 implies that there could be a difference in mean mpg between the transmission types, assuming a significance level of α=0.05.

A linear regression of the mpg vs. transmission was also done to investigate the strength of the relationship between these two variables.

```
model_transmission <- lm(mpg~am, data = car_data)
summary_lm <- summary(model_transmission)
print(summary_lm$r.squared)
```

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

The results show that the correlation between mpg and transmission types seems weak ($R^2 = 0.36$) and that only about 36% of the variation in mpg can be explained by this model (see Figure 2 in Appendix).

To improve the regression, a multivariate regression model will be done to include additional regressor variables beyond transmission. The following code shows the correlation between different regressor variables. The code determines the correlation between mpg (or weight) and every other regressor.

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

```
##        mpg        cyl       disp         hp       drat         wt
##  1.0000000 -0.8521620 -0.8475514 -0.7761684  0.6811719 -0.8676594
##       qsec         vs         am       gear       carb
##  0.4186840  0.6640389  0.5998324  0.4802848 -0.5509251
```

```
cor(mtcars)[6,]
```

```
##        mpg        cyl       disp         hp       drat         wt
## -0.8676594  0.7824958  0.8879799  0.6587479 -0.7124406  1.0000000
##       qsec         vs         am       gear       carb
## -0.1747159 -0.5549157 -0.6924953 -0.5832870  0.4276059
```

From these results, mpg appears highly negatively correlated with weight, disp (engine volume),cyl (number of cylinders) and hp (horsepower). However, weight is highly correlated with cyl, disp and hp so adding these variables in addition to weight might not add all that much value to the model. qsec (1/4 mile time) is very weakly correlated with weight and could help the model.

Here are some examples of expanding the model (without intercepts to avoid am factor split) to include more variables based on these arguments. To save space the summary of only the 3rd model will be printed.

```
mr_model <- update(model_transmission,mpg~am+wt-1, data = car_data)
mr_model2 <- update(model_transmission, mpg~am+wt+qsec-1, data=car_data)
mr_model3 <- update(model_transmission, mpg~am+wt+qsec+disp-1, data=car_data)

summary_mr <- summary(mr_model)
summary_mr2 <- summary(mr_model2)
summary_mr3 <- summary(mr_model3)
print(summary_mr3$coefficients)
```

```
##         Estimate  Std. Error   t value     Pr(>|t|)
## am    4.18085430 1.013616073  4.124692 3.005272e-04
## wt   -4.61279456 1.158173236 -3.982819 4.400086e-04
## qsec  1.70550996 0.127485705 13.378049 1.099649e-13
## disp  0.01202006 0.008891454  1.351866 1.872383e-01
```

```
print(summary_mr3$r.squared)
```

```
## [1] 0.9879113
```

From these results, the regression model is improved by the addition of the wt and qsec variables ($R^2 = 0.987$). The p-values for each coefficient show significance and they are unlikely to be zero. Adding the disp variable and coefficient did not really improve the model, and this coefficient did not show significance to be different than zero. This was the case when trying the other remaining variables as well beyond wt and qsec.

The analysis of variance (ANOVA) was also used to compare models. The ANOVA p-values show that adding the weight and qsec variables help to improve the model but disp is not adding value beyond this. Other variables to replace disp also did not provide improvement. Figure 3 shows the residuals of mr_model2. There is no obvious pattern for variance of the residuals with predicted values and they appear to follow a normal distribution.

```
anova_results <- anova(mr_model,mr_model2, mr_model3)
cat("Model 2(am+wt+qsec): ",anova_results$`Pr(>F)`[2])
```

```
## Model 2(am+wt+qsec):  1.538901e-15
```

```
cat("Model 3(am+wt+qsec+disp): ",anova_results$`Pr(>F)`[3])
```

```
## Model 3(am+wt+qsec+disp):  0.1872383
```

Conclusion

There does appear to be a difference in mean mpg with transmission type, with manual transmission delivering about 7.3 more mean mpg. However, mpg is not well-predicted by transmission type alone; adding weight and 1/4 mile time help to improve the predicted mpg. The weight coefficient is negative, implying that more weight reduces the mpg, while the 1/4 mile time coefficient is positive, showing that the longer time to reach a 1/4 mile could be related to improved fuel efficiency.

Appendix

Figure 1: Exploratory box plot of miles per gallon for automatic and manual transmission types.

```
boxplot(data= car_data,mpg~am, main="Car Mileage vs. Transmission Type",
xlab="Transmission Type", ylab="Miles Per Gallon",
col = c("gold"), names=c("Automatic", "Manual"))
```

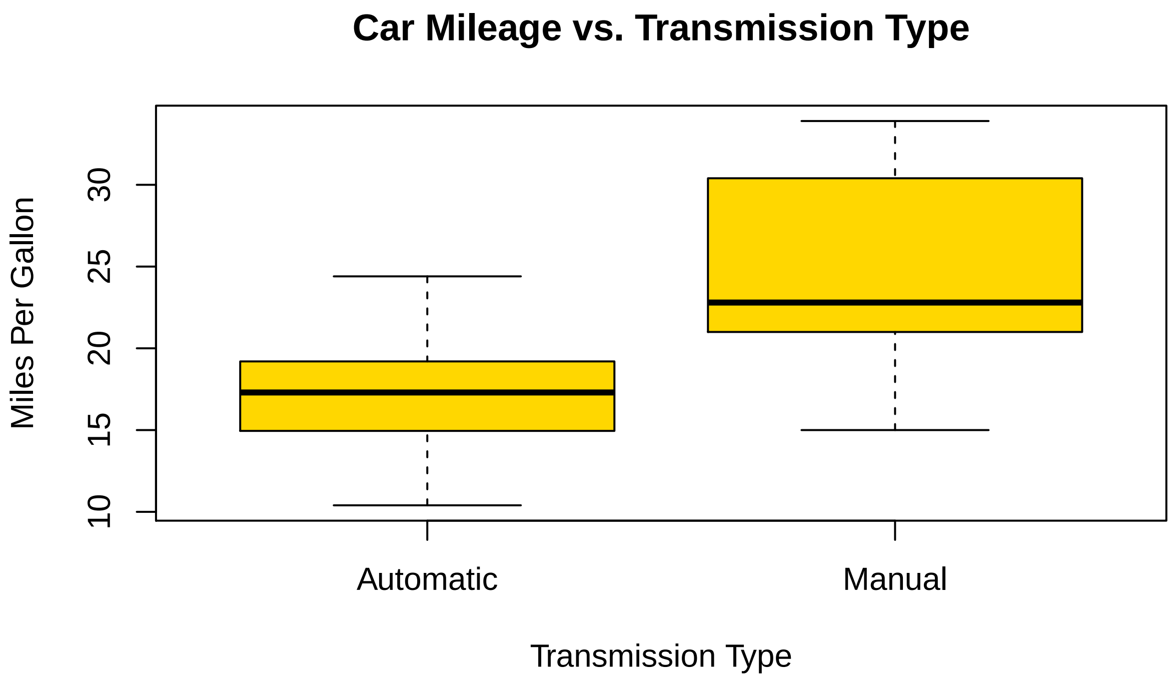


Figure 2: Linear regression plot of MPG vs transmission type and the associated residuals plot.

```
par(mfrow=c(1,2))
plot_lm <- plot(car_data$am, car_data$mpg, xlab = "Transmission Type",
ylab = "Miles per gallon", pch = 21, bg = "lightgreen", col = "black",xaxt = 'n',cex=1.5)
axis(side = 1, at = c(0,1),labels = c("Automatic", "Manual"))
abline(model_transmission, lwd = 2)

resid_plot <- plot(car_data$am, resid(model_transmission), xlab = "Transmission Type",
ylab = "Residuals", pch = 21,
bg = "lightgreen", col = "black", xaxt = 'n', cex=1.5)
axis(side = 1, at = c(0,1),labels = c("Automatic", "Manual"))
abline(h=0, lwd = 2)
```

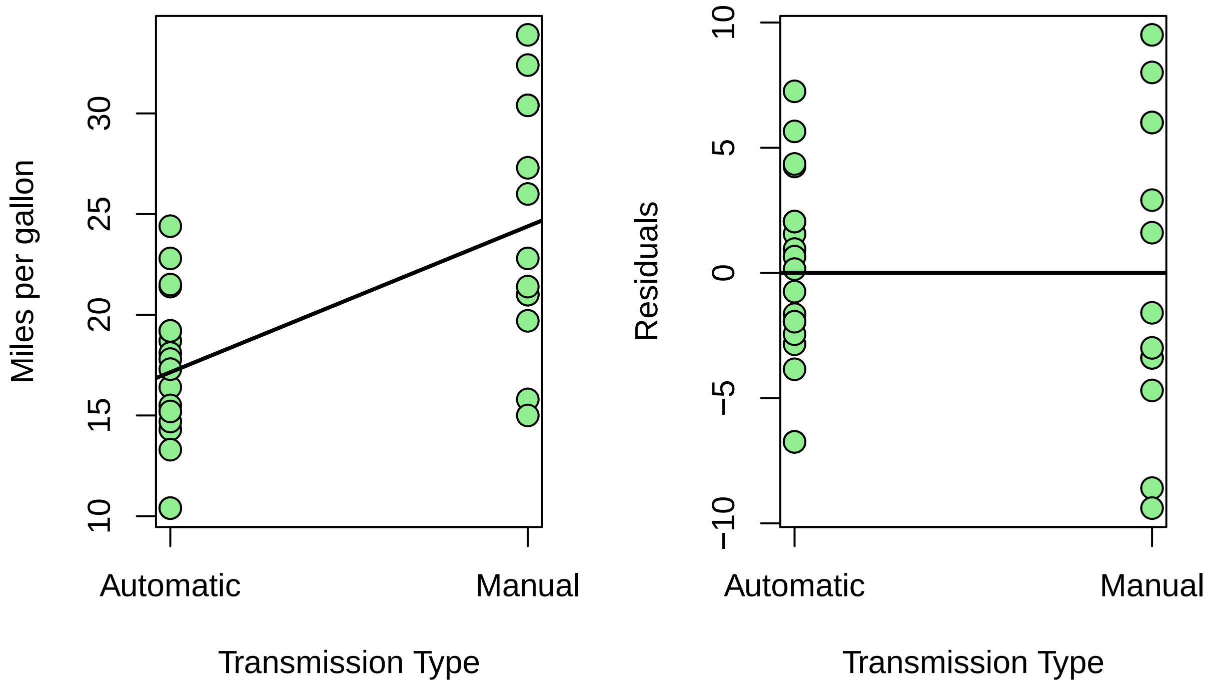
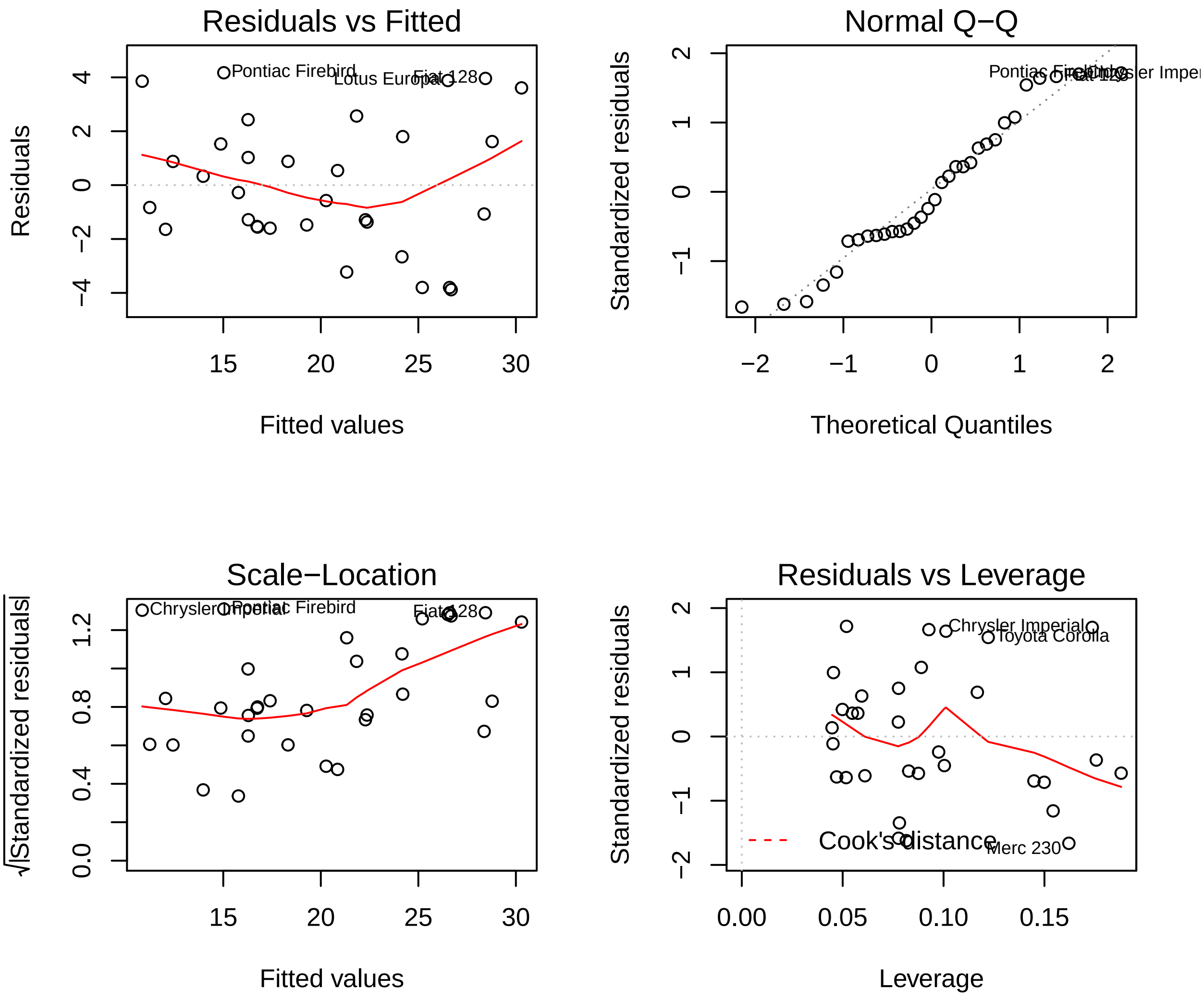


Figure 3: Residuals plot for the multivariate regression including am, weight and qsec (Model 2).

```
par(mfrow = c(2,2))
plot(mr_model2)
```


Residuals vs Fitted
Residuals
Fitted values
−4
−2
0
2
4
15
20
25
30
Pontiac Firebird
Lotus Europa
Fiat 128
Normal Q−Q
Standardized residuals
Theoretical Quantiles
−2
−1
0
1
2
Pontiac Firebird
Fiat 128
Chrysler Imperial
Scale−Location
√|Standardized residuals|
Fitted values
0.0
0.4
0.8
1.2
15
20
25
30
Chrysler Imperial
Pontiac Firebird
Fiat 128
Residuals vs Leverage
Standardized residuals
Leverage
−2
−1
0
1
2
0.00
0.05
0.10
0.15
Chrysler Imperial
Toyota Corolla
Merc 230
Cook's distance