Regression Models Course Project

Asfa L.

Executive Summary

This project takes a closer look at the automobile industry. Specifically, we will look at a data set of a collection of cars and explore the relationship between a set of variables and miles per gallon (MPG) (outcome). The data set being used is the mtcars data set and we will aim to answer the following two questions: “Is an automatic or manual transmission better for MPG?” and “Quantify the MPG difference between automatic and manual transmissions.” Our analysis ultimately concludes that manual cars are better for MPG, qsec and weight along wth transmission have a significant affect on mpg, and it will provide a method for quantifying how many more mpg a manual car will have compared to an automatic car.

Exploratory Data Analysis

The mtcars data set contains 32 observations (automobile brand/type) of 11 variables. Let’s load the data set, see what it looks like and do some plots to visually see the data.

```
data(mtcars)
head(mtcars)
summary(mtcars)
str(mtcars)
```

Now that we have a better idea of what the data looks like lets start by comparing mpg to am (transmission) where 0=automatic and 1=manual. (See Appendix for output)

```
boxplot(mpg ~ am, data= mtcars, xlab="Transmission (0 = Automatic, 1 = Manual)", ylab="MPG", main="MPG vs. Transmission")
```

A manual transmission seems to have a higher mean MPG and overall higher MPG than automatic transmission. Now, Let’s look at all the variables in a pairs plot. (See Appendix for output)

```
pairs(mtcars, panel=panel.smooth, main="mtcars pairs plot")
```

Variables that seem to have a high correlation to MPG are cyl, disp, hp, qsec, and wt.

Hypothesis Testing

Lets confirm our initial statement that that there is a difference in mpg based on transmission. The null hypothesis is that there is no difference in mpg based on transmission. (See Appendix for output)

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

The p value is less than .05 so we can reject the null and conclude that there is a difference in MPG depending on transmission. We can see that the mean mpg for automatic is 17.14 whereas for manual it is about 7mpg higher at 24.4.

Regression Analysis

We can now fit a linear model based on the results of our hypothesis test. (See Appendix for output)

```
fit1<-lm(mpg ~ am, data=mtcars)
summary(fit1)
```

As expected the p-value is less than .05 which makes the am coefficient a meaningful addition to the model because changes in its value are related to changes in the response variable - mpg. However, the r squared value is quite low at approximately .34. Lets consider other models which include some of the other confounding variables. 1. All variables. (See Appendix for output)

```
 fitall<- lm(mpg ~ ., data=mtcars)
summary(fitall)
```

Here we see the r squared value jump to .81, and the standard error is lower as expected with more variables, however, none of the p-values for the coefficients are significant. 2. Let’s fit a model with the varibales we saw in our exploratory analysis as having a high correlation to mpg. (See Appendix for output)

```
fitsixvars<-lm(mpg ~ am + cyl + disp + hp + qsec + wt, data=mtcars)
summary(fitsixvars)
```

Only the pvalue for wt is significant, but the r squared value is .83. The wt coefficent is also the largest of all the variables so for a one unit change in mpg the wt decreases by 4 units. So the less weight the greater the fuel efficiency. 3. qsec(1/4 mile time) has the next highest coefficient, so lets add that to the model and consider the interaction between transmission and weight while holding all the other variables constant.

```
fitfin<-lm(mpg ~ am + wt + qsec + am*wt, data=mtcars)
summary(fitfin)
```

```
##
## Call:
## lm(formula = mpg ~ am + wt + qsec + am * wt, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.5076 -1.3801 -0.5588  1.0630  4.3684
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    9.723      5.899   1.648 0.110893
## am            14.079      3.435   4.099 0.000341 ***
## wt            -2.937      0.666  -4.409 0.000149 ***
## qsec           1.017      0.252   4.035 0.000403 ***
## am:wt         -4.141      1.197  -3.460 0.001809 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.084 on 27 degrees of freedom
## Multiple R-squared:  0.8959, Adjusted R-squared:  0.8804
## F-statistic: 58.06 on 4 and 27 DF,  p-value: 7.168e-13
```

We see here all the pvalues are significant and the rsquared value .88 which is the highest we have seen and means that the model explains 88% of the variation. The interaction term seems to be correct and make sense that weight and transmission would be correlated and have a significant effect on mpg. The standard error is also 2 which is the lowest we have seen. So, we choose this model.

Residual Plots and Diagnostics

Now we look at the residuals to validate our model. (See Appendix for output)

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

1.For the residual plot, the points are randomly dispersed around the horizontal axis, which shows a random pattern, indicating a good fit for a linear model. 2. The QQ Plot indicates the residuals are normally distributed as they mostly fall on the line. 3.The Scale-Location Plot uses the square root of the standardized residuals. Like the first plot, there should be no discernable pattern to the plot which is what it shows. 4.Residuals vs. Leverage plot - Leverage is a measure of how much each data point influences the regression. Here we do not see any outliers or points >.5. So there are no points that show a large residual which can distort the regression. In performing diagnostics, lets check the hatvalues which measures leverage and see if any points are greater than .5.

```
leverage <- round(hatvalues(fitfin), 3)
leverage[which(leverage > 0.5)]
```

```
## named numeric(0)
```

As we expected there are none. Lets also check the dfbetas which measures how much the coefficients change when the i-th case is deleted.

```
inf <- dfbetas(fitfin)
inf[which(inf > 1)]
```

```
## numeric(0)
```

No dfbetas greater than 1, so changes are considered not significant.

Conclusions

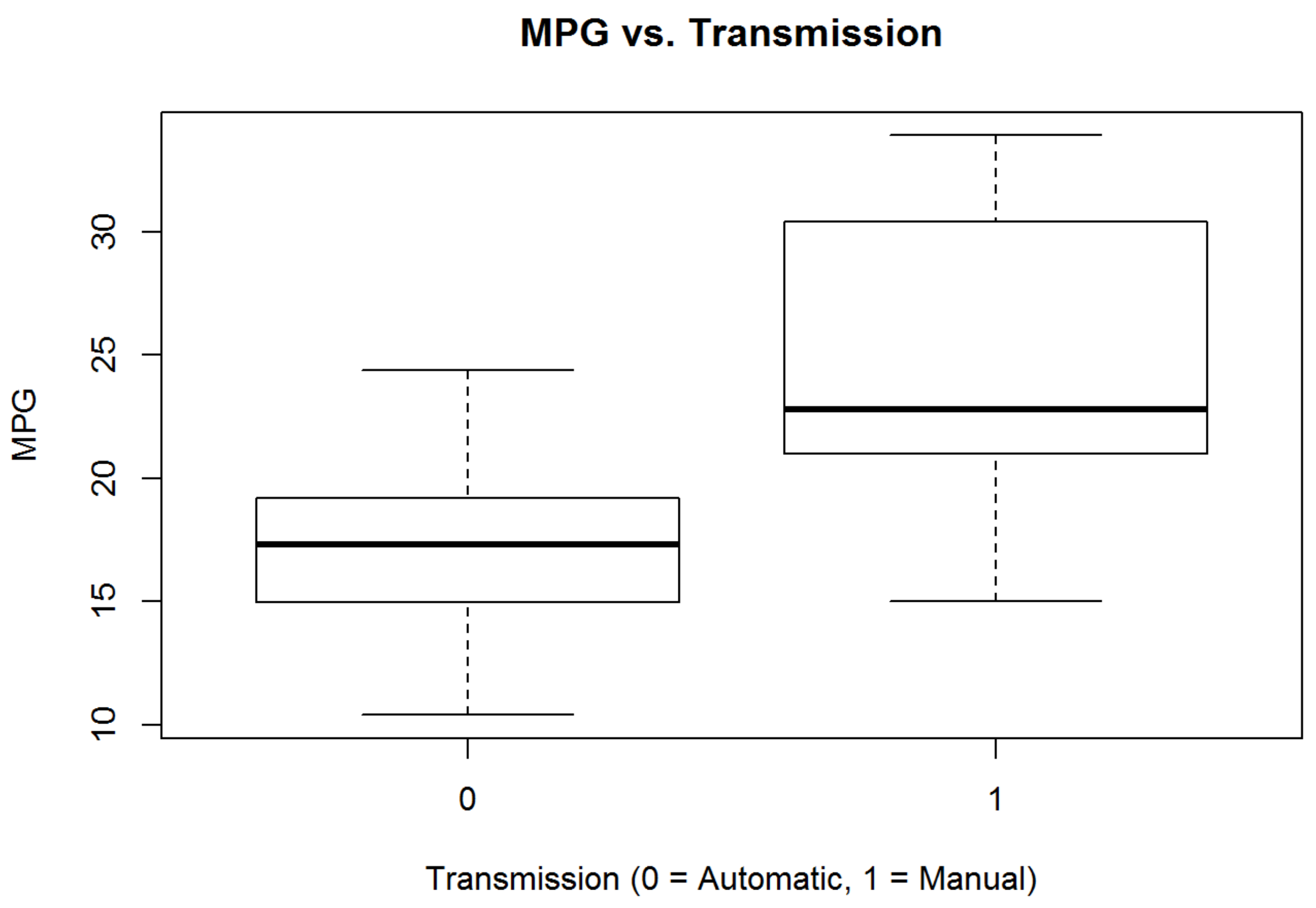
We originally set out to answer two questions: “Is an automatic or manual transmission better for MPG?” and “Quantify the MPG difference between automatic and manual transmissions”

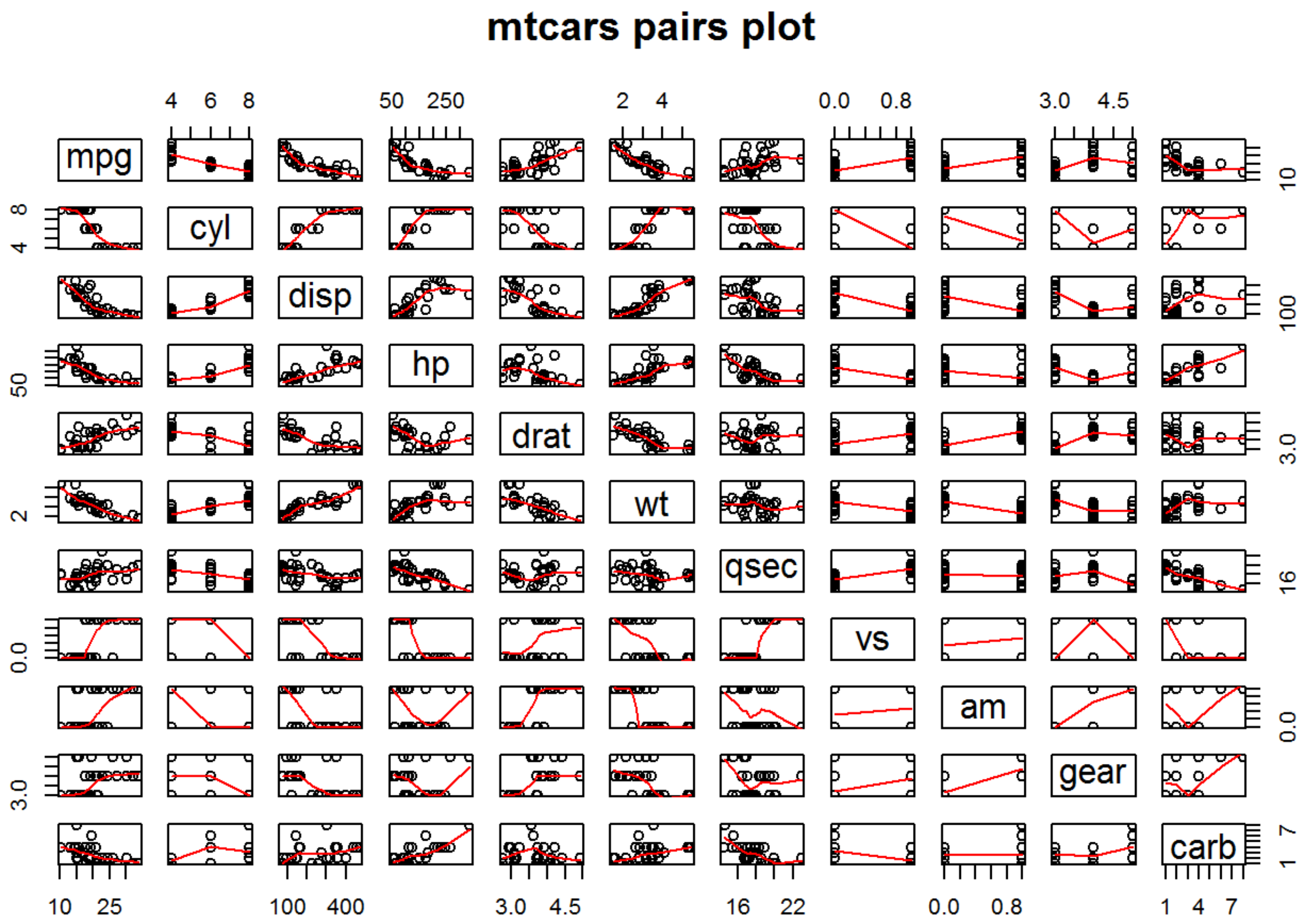
From our analysis we can conclude that on average manual transmissions have greater fuel efficiency, specifically seven mpg greater than automatic cars. At the 95% confidence level manual cars will have anywhere from 3.2 to 11.3 more mpg than automatic cars.

The equation that results from the final regression model(fitfin) is mpg = 9.723 + 14.079am - 2.937wt + 1.017qsec - 4.141(am*wt). For manual cars (use 1 for am) the equation becomes mpg = 23.803 - 7.078wt + 1.017qsec. For automatic cars (substitute 0 for am) the equation is mpg = 9.723 - 2.937wt + 1.017 qsec. Subtracting the automatic equation from the manual equation gives you mpg=14.079 -4.141wt. So manual cars will have this much more mpg than cars with automatic transmission.

Appendix

Exploratory Analysis





Hypothesis Testing

```
##
##  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
```

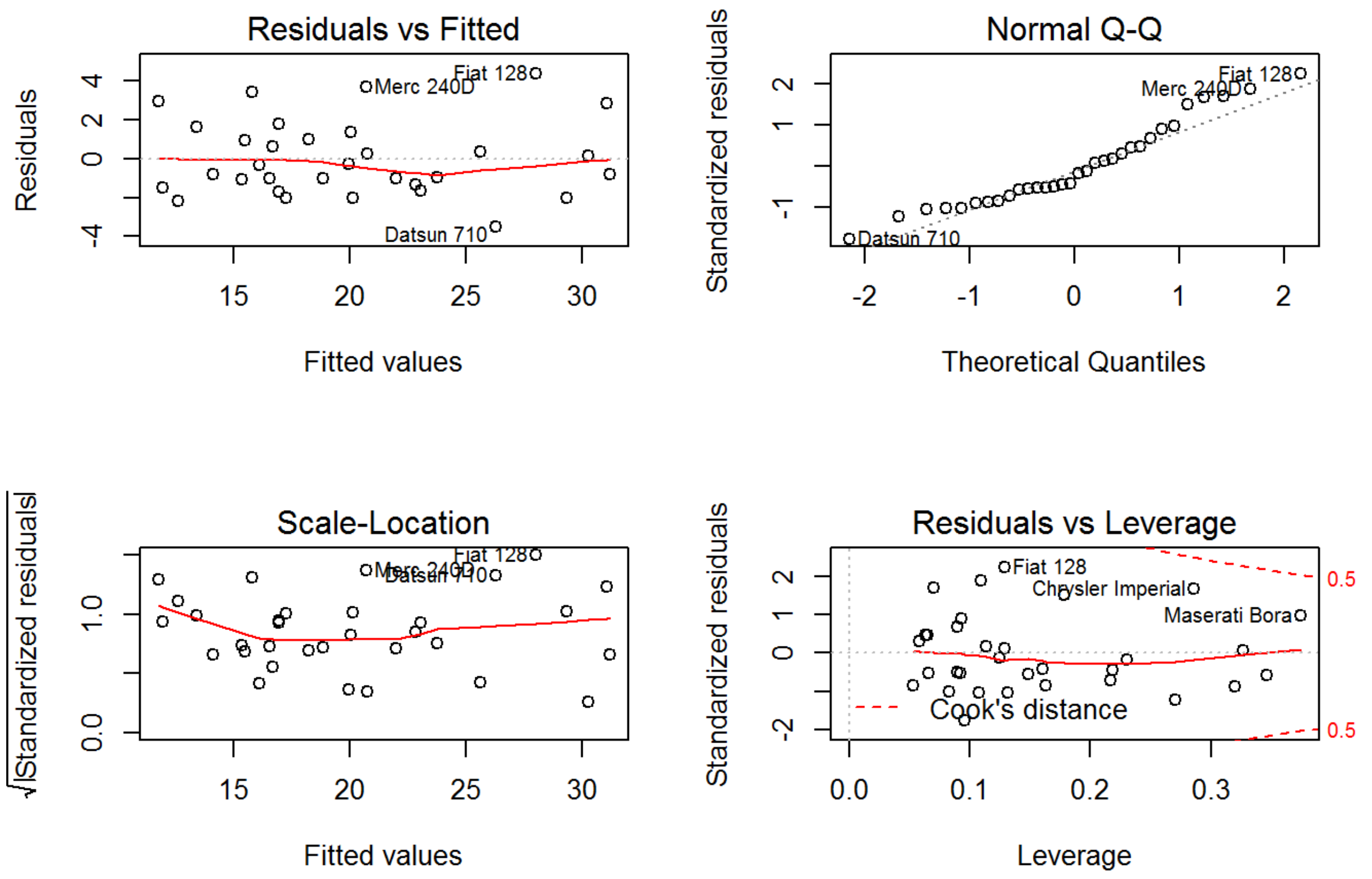
Regression Analysis

```
##
## 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 ***
## am             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
```

```
##
## Call:
## lm(formula = mpg ~ ., data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4506 -1.6044 -0.1196  1.2193  4.6271
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 12.30337   18.71788   0.657   0.5181
## cyl         -0.11144    1.04502  -0.107   0.9161
## disp         0.01334    0.01786   0.747   0.4635
## hp          -0.02148    0.02177  -0.987   0.3350
## drat         0.78711    1.63537   0.481   0.6353
## wt          -3.71530    1.89441  -1.961   0.0633 .
## qsec         0.82104    0.73084   1.123   0.2739
## vs           0.31776    2.10451   0.151   0.8814
## am           2.52023    2.05665   1.225   0.2340
## gear         0.65541    1.49326   0.439   0.6652
## carb        -0.19942    0.82875  -0.241   0.8122
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.65 on 21 degrees of freedom
## Multiple R-squared:  0.869,  Adjusted R-squared:  0.8066
## F-statistic: 13.93 on 10 and 21 DF,  p-value: 3.793e-07
```

```
##
## Call:
## lm(formula = mpg ~ am + cyl + disp + hp + qsec + wt, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.6755 -1.6757 -0.4477  1.2615  4.6289
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 20.05170   13.30486   1.507  0.14432
## am           2.94075    1.71810   1.712  0.09935 .
## cyl         -0.50207    0.78882  -0.636  0.53025
## disp         0.01396    0.01155   1.209  0.23802
## hp          -0.01956    0.01489  -1.314  0.20088
## qsec         0.81018    0.57171   1.417  0.16879
## wt          -3.99773    1.21564  -3.289  0.00299 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.458 on 25 degrees of freedom
## Multiple R-squared:  0.8659, Adjusted R-squared:  0.8337
## F-statistic: 26.91 on 6 and 25 DF,  p-value: 9.29e-10
```

Residual Plots and Diagnostics


Residuals vs Fitted
Residuals
4
2
0
-2
-4
Fiat 128
Merc 240D
Datsun 710
15
20
25
30
Fitted values
Normal Q-Q
Standardized residuals
2
1
-1
Fiat 128
Merc 240D
Datsun 710
-2
-1
0
1
2
Theoretical Quantiles
Scale-Location
√|Standardized residuals|
1.0
0.0
Fiat 128
Merc 240D
Datsun 710
15
20
25
30
Fitted values
Residuals vs Leverage
Standardized residuals
2
1
0
-2
Fiat 128
Chrysler Imperial
Maserati Bora
Cook's distance
0.5
0.5
0.0
0.1
0.2
0.3
Leverage