Regression Models Course Project

Executive Summary

The purpose of this project is to quantitatively determine whether an automatic or manual transmission is better for gas mileage (mpg). The data is from the 1974 Motor Trend US magazine, and covers 32 cars' design, performance and fuel consumption. By use of linear regression, we are able to determine that, on average, automatic transmissions equate to lighter and ultimately more fuel efficient (i.e. lower MPG) vehicles than their manual counterparts.

Initial Project Setup & Exploratory Data Analysis

Let's begin by loading the appropriate mtcars data.

```r
data("mtcars")
```

Using ?mtcars, we are able to see that the 'am' (Transmission, 0 = automatic, 1 = manual) and 'vs' (Engine, 0 = V-shaped, 1 = straight) variables should be factors. We'll take a look at the data and transform the necessary variables in order to compute our exploratory graphs & later regression models accurately.

```r
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
```

```r
mtcars$vs <- as.factor(mtcars$vs)
mtcars$am <- factor(mtcars$am)
```

Now that our project is set up, we can visualize the data to explore potential relationships. First, we'll run a correlation matrix plot across all variables (**Appendix Figure 1**). We see that the the wt variable is highly correlated with both mpg and am so this will probably be a variable of interest for us. We'll also take a direct look at transmission type and MPG (**Appendix Figure 2**), since this is the relationship we're most interested in. The resulting boxplot suggests automatics have a lower mean mpg than manual transmissions, but we'll do some regression modeling to take a deeper look at this pattern.

Regression Models & Inference

To further investigate this relationship, we'll use a nested approach, meaning that, with each regression model, an additional predictor will be added. First, we'll run a model with all predictors against the mpg outcome (**Appendix Figure 3**). No relationships were significant (all p-values >.05). However am, wt and qsec showed closer correlations than others predictors (p-values <.3), suggesting potential dependencies.

Nested regression models were then computed with these 3 variables and compared using ANOVA inference testing to check for model parsimony and justification of complexity (i.e. more predictors in the model increases complexity, bias and/or risk of overfitting).

While running nested models, we find that holding weight (wt) constant (summary(mod2)), the relationship between transmission (am) and mpg is no longer significant. When graphing this relationship (**Appendix Figure 4**), we can see that manual cars are typically heavier than automatics, therefore leading to poorer gas mileage. With this in mind, we'll run a 2nd model with an interaction between am and wt as well.

```
mod1 <-lm(mpg ~ am, data=mtcars) #only mpg and am variables
mod2 <- lm(mpg ~ am + wt, data=mtcars) #adding wt
mod3 <- lm(mpg ~ am + wt + am:wt, data=mtcars) #adding interaction
mod4 <- lm(mpg ~ am + wt + qsec + wt:am, data=mtcars) #adding qsec

#ANOVA of the four models
anova(mod1, mod2, mod3, mod4)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ am + wt
## Model 3: mpg ~ am + wt + am:wt
## Model 4: mpg ~ am + wt + qsec + wt:am
##   Res.Df    RSS Df Sum of Sq       F    Pr(>F)
## 1     30 720.90
## 2     29 278.32  1    442.58 101.892 1.161e-10 ***
## 3     28 188.01  1     90.31  20.792 9.938e-05 ***
## 4     27 117.28  1     70.73  16.284  0.000403 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

All of our models have significant outputs. Also, because each p-value in the ANOVA table is significant, we can infer that each added predictor adds valuable information to the model. Before making a final conclusion about whether or not the most complex model (mod4) best fits the data, let's run some diagnostics to make sure it doesn't violate regression assumptions.

Residual Analyses & Diagnostics

We will first run a dfbeta test to determine if any observations significantly influence the model's coefficient estimates (cutoff for small datasets = 1).

```
sum(dfbetas(mod4) > 1)
```

```
## [1] 0
```

There are four standard diagnostic plots looking at the model's residuals we should also run. Looking at the results (**Appendix Figure 5**), the Residual vs Fitted plot has no pattern, the points on the Normal Q-Q plot lie close to the line, the Scale-Location plot has randomly distributed points and the points of the Residuals vs Leverage plot fall within .5 cutoffs, all of which means (respectively) that the model's residuals are independent, normally distributed, have constant variance and have no significant outliers.

The model passes the necessary assumptions! Let's look at the model's output directly for a final interpretation of the results.

```
summary(mod4)
```

```
##
## Call:
## lm(formula = mpg ~ am + wt + qsec + wt:am, 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
## am1           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 ***
## am1: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
```

All coefficients have p-values < .05 and are thus significant. This shows that, while holding weight and mile time constant, manual transmissions add 14.079 + (-4.141)*wt more miles per gallon on average than automatic transmission. So, if we compare cars of the same weight and mile time, those with manual transmission will be significantly slower than those with automatic transmission.

Appendix

Figure 1: Correlation plot across all mtcars variables

```
ggcorr(mtcars, nbreaks = 5, palette="Dark2", label = T)
```

```
## Warning in ggcorr(mtcars, nbreaks = 5, palette = "Dark2", label = T): data in
## column(s) 'vs', 'am' are not numeric and were ignored
```

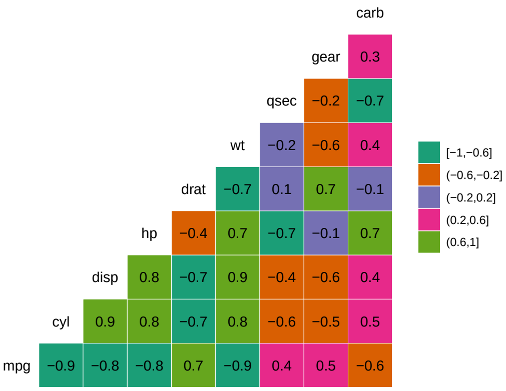


Figure 2: Boxplot of transmission type vs mpg

```r
f2 <- ggplot(mtcars, aes(x = am, y = mpg, color = am)) +
    geom_boxplot(outlier.colour="red", outlier.shape=8, outlier.size=2) + scale_color_brewer(palette="Da
f2
```

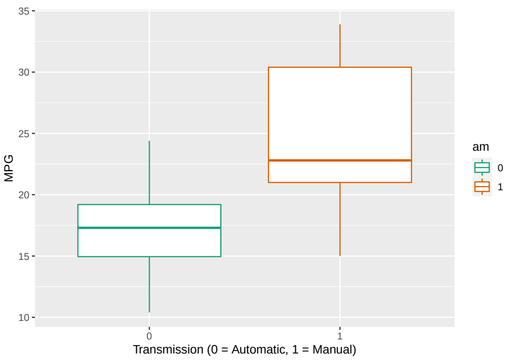


Figure 3: MPG outcome variable vs all predictor variables model

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

```
##
## 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
## vs1          0.31776    2.10451   0.151   0.8814
## am1          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
```

Figure 4: Relationship between mpg, transmission and car weight

```
f4 <- ggplot(mtcars, aes(x=wt, y=mpg, group=am, color= factor(am))) + geom_point() + scale_color_brewer
f4
```

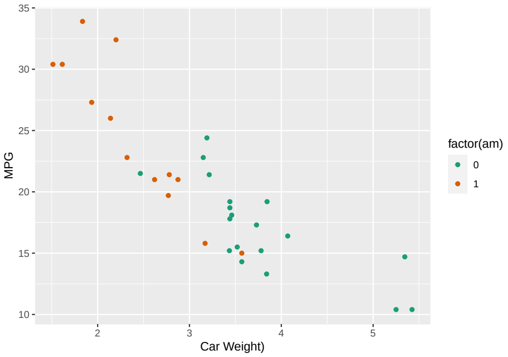
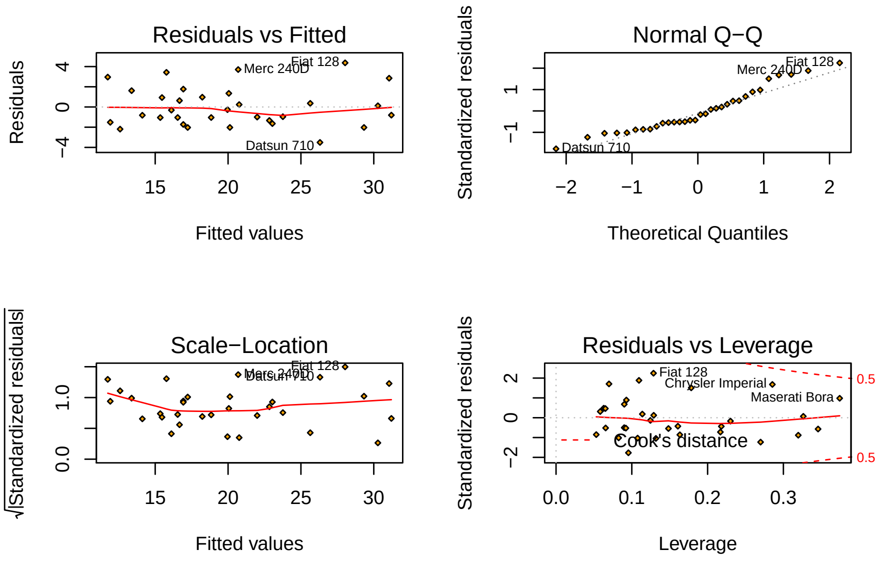


Figure 5: Regression Model Residual Plots

```
par(mfrow=c(2,2))
plot(mod4, pch=23, bg='orange', cex=.5)
```



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