Regression Models Course Project

Lena Horsley

Executive Summary

The purpose of this report is to examine the relationship between transmission type and gas mileage (mpg). Due a t test p value of p 0.001374, it was determined cars with manual transmissions had better gas mileage. Although the initial regression model showed an increase of 7.245 mpg, the R-squared value of 0.3395 led to the assumption of other variables affecting gas mileage. Despite using models with additional variables (such as weight, horsepower, and number of cylinders) and an R-squared value of 0.8267 for the best model, there was a slight increase of 1.47805 mpg for manual transmissions.

Setup

```
library(datasets)
library(ggplot2)
library(knitr)
myCarData <- data.frame(mtcars)
myCarData$am <-gsub("0","AUTO",myCarData$am)
myCarData$am <-gsub("1","MAN",myCarData$am)
mpg <- as.factor(myCarData$mpg)
cyl <- as.factor(myCarData$cyl)
hp <- as.factor(myCarData$hp)
wt <- as.factor(myCarData$wt)
gear <- as.factor(myCarData$gear)
drat <- as.factor(myCarData$drat)
```

Analysis

Initial exploratory analysis

	Manual	Automatic
Mean	24.392308	17.147368
StDev	6.166504	3.833966
Range (low)	15.000000	10.400000
Range (high)	33.900000	24.400000

```
# HYPOTHESIS: Transmission type does not affect gas mileage
t.test(myCarData$mpg~myCarData$am, paired = FALSE)

##
##  Welch Two Sample t-test
##
## data:  myCarData$mpg by myCarData$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 AUTO  mean in group MAN
##           17.14737           24.39231
```

Because α (0.05) is greater than p (0.001374), the result is statistically significant. The hypothesis "Transmission type does not affect gas mileage" can be rejected.

Finding an Appropriate Model

```
# MPG vs Transmission
linRegFit1 <- lm(mpg~am,myCarData)
summary(linRegFit1)

##
## Call:
## lm(formula = mpg ~ am, data = myCarData)
##
## 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 ***
## amMAN          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
```

Although initial analysis showed a strong correlation between miles per gallon and transmission (and increase of 7.245 mpg), an R-squared value of 0.3395 indicates the model is a poor fit for the data. By visualizing the data with a few simple plots (see Figures 1 - 5 in the Appendix), it can be inferred other variables affect gas mileage. We can see this by looking at the coefficients for each variable in the following models.

```
# Confounding variables (all variables considered)
linRegAll <- lm(mpg ~ ., myCarData)
summary(linRegAll)

##
## Call:
## lm(formula = mpg ~ ., data = myCarData)
##
## 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
## amMAN        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
```

With all of the variables as confounders, the R-squared value increases to 0.8066 but the gas mileage decreases to 2.52023. There are no p values less than 0.05.

```
# Confounding variables (transmission, horsepower, drat, weight, # of cylinders)
linRegFit2 <- lm(mpg ~ am + hp + drat + wt + cyl, myCarData)
summary(linRegFit2)

##
## Call:
## lm(formula = mpg ~ am + hp + drat + wt + cyl, data = myCarData)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.3566 -1.7230 -0.6182  1.1543  5.6463
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 34.51085    7.47862   4.615  9.3e-05 ***
## amMAN        1.34202    1.57153   0.854   0.4009
## hp          -0.02557    0.01413  -1.810   0.0818 .
## drat         0.36034    1.49320   0.241   0.8112
## wt          -2.57272    0.94670  -2.718   0.0115 *
## cyl         -0.68421    0.64474  -1.061   0.2983
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.554 on 26 degrees of freedom
## Multiple R-squared:  0.8494, Adjusted R-squared:  0.8204
## F-statistic: 29.32 on 5 and 26 DF,  p-value: 6.633e-10
```

When am, hp, drat, wt, and cyl are confounders, the R-squared value increases to 0.8204. The gas mileage again decreases to 1.34202. The p value for wt is 0.0115 (less than 0.05).

```
# Confounding variables (transmission, horsepower, weight, # of cylinders)
linRegFit3 <- lm(mpg ~ am + hp + wt + cyl, myCarData)
summary(linRegFit3)

##
## Call:
## lm(formula = mpg ~ am + hp + wt + cyl, data = myCarData)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4765 -1.8471 -0.5544  1.2758  5.6608
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 36.14654    3.10478  11.642 4.94e-12 ***
## amMAN        1.47805    1.44115   1.026   0.3142
```

```
## hp          -0.02495    0.01365  -1.828   0.0786 .
## wt          -2.60648    0.91984  -2.834   0.0086 **
## cyl         -0.74516    0.58279  -1.279   0.2119
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.509 on 27 degrees of freedom
## Multiple R-squared:  0.849,  Adjusted R-squared:  0.8267
## F-statistic: 37.96 on 4 and 27 DF,  p-value: 1.025e-10
```

Finally, the R-squared value increases to 0.8267 when am, hp, wt, and cyl are confounders. There's a slight increase of gas mileage of 1.47805 from the previous model. The p value for wt is 0.0086 (less than 0.05).

```
# Compare the two models (linRegFit1 and linRegFit3 - the model with the highest R-squared value).
anova(linRegFit1,linRegFit3)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ am + hp + wt + cyl
##   Res.Df   RSS Df Sum of Sq      F    Pr(>F)
## 1     30 720.9
## 2     27 170.0  3     550.9 29.166 1.274e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Conclusion

Despite simple analysis showing vehicles with manual transmissions as having better gas mileage, the R-squared value of 0.3385 (from a linear model of mpg and transmission type) suggests other factors affect gas mileage. In addition to transmission type (am), other variables such as hp, wt, cyl, and drat influence a vehicle's gas mileage. With this quick analysis, the best model considered hp, wt, and cyl as confounders and had an R-squared value of 0.8267. There was a slight increase in gas mileage for cars with a manual transmission at a value of 1.47805.

It's worth noting these results may be affected by the size of the data set. A larger data set would result in a more accurate model where correlation between gas mileage and a single variable would be clearer.

Appendix

Figure 1.

```
ggplot(data=myCarData, aes(transmission, mpg, fill = transmission)) + geom_boxplot()
```

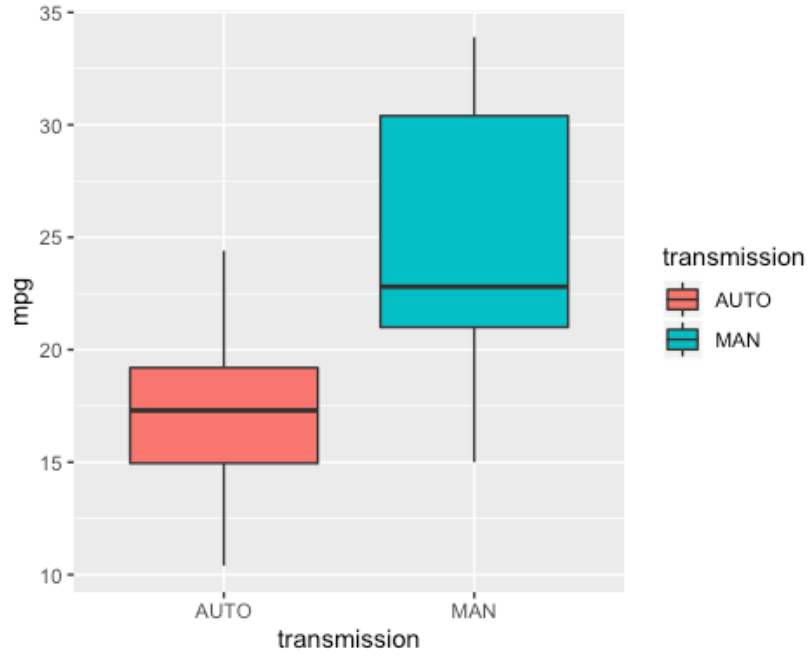

Figure 2. Number of Gears as a Factor

```
coplot(mpg ~ disp | as.factor(gear), data = myCarData, rows = 1)
```

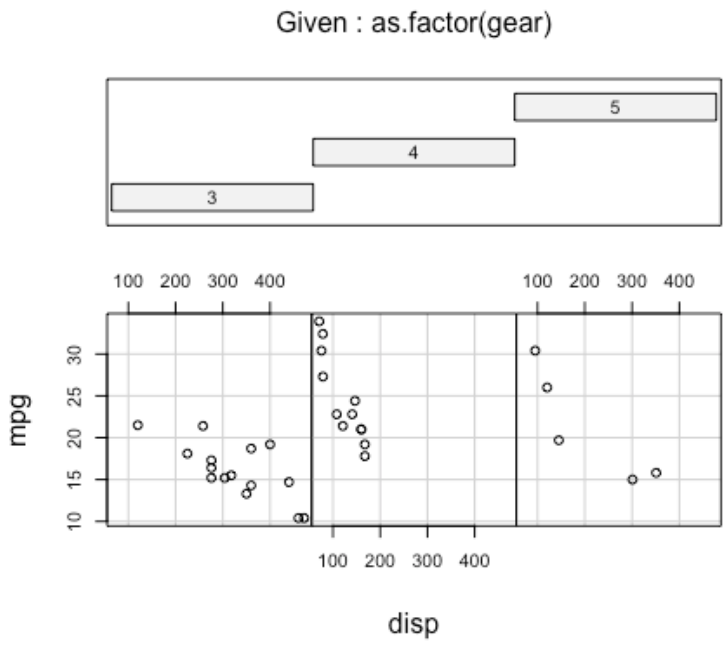

Figure 3. Number of Cylinders as a Factor

```
coplot(mpg ~ disp | as.factor(cyl), data = myCarData, rows = 1)
```

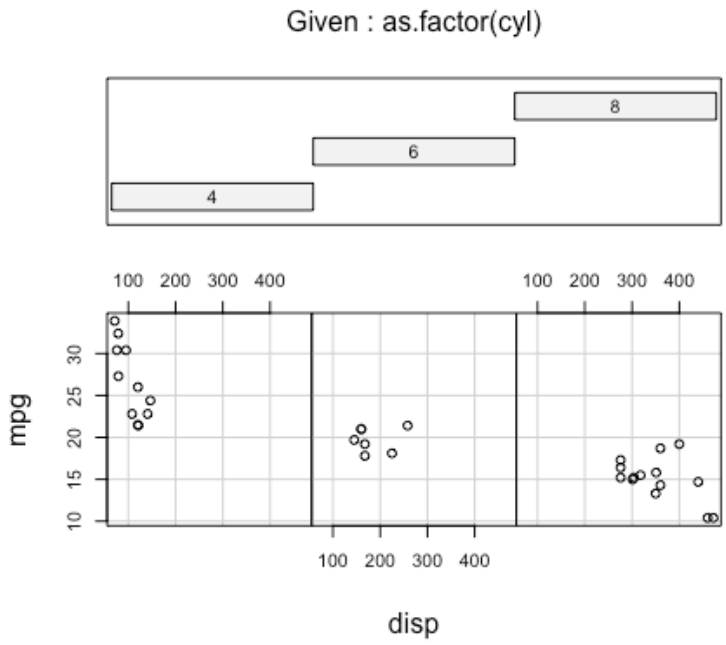

Figure 4. MPG vs Weight

```
plot(myCarData$wt, myCarData$mpg, xlab = "weight", ylab = "mpg")
```

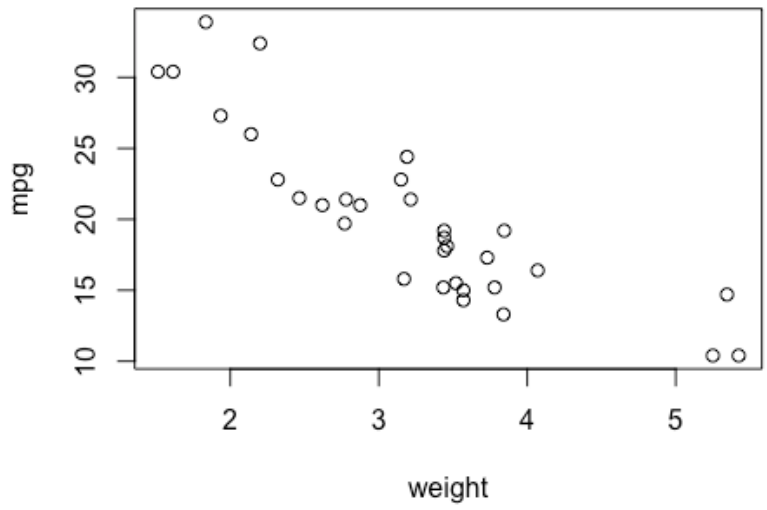

Figure 5. MPG vs Horsepower

```
plot(myCarData$hp, myCarData$mpg, xlab = "horsepower", ylab = "mpg",)
```

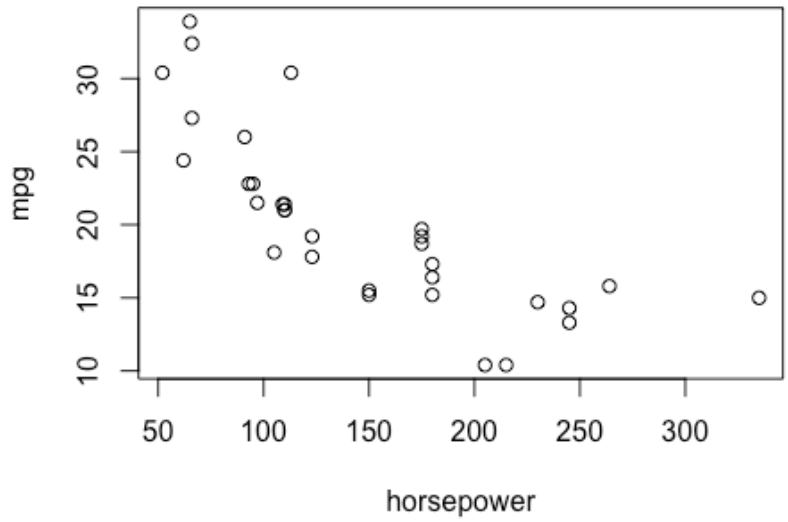

Figure 6. Residuals Plot for linRegFit3

```
par(mfrow = c(2,2))
plot(linRegFit3)
```