REGRESSION MODELS PROJECT *Lawrence Lau*

Summary

Given the mtcars dataset, we use a linear regression and verify the results through a hypothesis test to show that manual transmissions average 7.24 miles per gallon more than automatic transmissions. We then use covariate adjustment and multiple regression models to explore other variables that could be factors, and after finding a best model to hold all else equal we discover manual transmissions are more efficient by 1.49 mpg .

Exploratory Data Analysis

In Appendix 1, Boxplot - Transmission Type vs MPG, we see the mean for automatics is 17.15 mpg and the mean for manuals is 24.39 mpg. Let's run a hypothesis test to verify there really is a difference.

Hypothesis Test And the p-value is…

```
## [1] 0.001374
```

With a p-value of less than .05, we reject the null hypothesis that the average miles per gallon are the same in automatics and manuals. So there really is a difference. But as the R-Squared number from the model shows…

```
summary(lm(mpg~am))$r.squared
```

```
## [1] 0.3598
```

… that only explains ~36% of the variance. We must dig deeper into the other variables.

Models Analyses

```
fit2 <- lm(mpg~., data=mtcars); fit2$coefficients
```

```
## (Intercept)         cyl        disp          hp        drat          wt
##    12.30337    -0.11144     0.01334    -0.02148     0.78711    -3.71530
##        qsec          vs          am        gear        carb
##     0.82104     0.31776     2.52023     0.65541    -0.19942
```

From this list of correlation coefficients of predictor variables fitted individually against MPG as the outcome, we can rank the following variables having a negative effect on mileage efficiency from greatest to least: wt (weight), carb (# of carburetors), cyl (# cylinders), and hp (horsepower). The variables having a possible effect on efficiency from least to greatest are disp (displacement), vs (V or Inline cylinders), gear (# forward gears), drat (rear axle ratio), qsec (1/4 mile time), and am (transmission type).

I'll take the wt, carb, cyl, vs, gear, drat, and qsec variables and after ruling out any variables that have a stronger than .5 correlation with am, fit them into a new regression with am against mpg. We must be careful which variables we choose, because anytime a regressor is added the standard error increases in the other regressors. Accordingly, the more regressors that are added, the higher the variance inflates- and actually

the more correlated the regressors are the higher the inflation level. In other words, regressing two variables that are similar to each other (or have some sort of relationship like # of pistons and # of cylinders) will make the standard error that much higher for those variables than another variable that has no correlation with them. How much higher? The square root of the variance taken as a percentage. However if you underfit a model by precluding a regressor, say in an attempt to avoid inflating the variance, you will have a bias. It is preferable to include all necessary and even unnecessary variables in order to have an unbiased variance estimate.

First, Correlations.

```
##     am/wt am/carb  am/cyl  am/vs am/gear am/drat am/qsec
## 1 -0.6925 0.05753 -0.5226 0.1683  0.7941  0.7127 -0.2299
```

Next, we'll leave out any pairs that have a correlation > .5 (bye bye am/gear and am/drat) and run covariate models.

```
## (Intercept)          am          wt (Intercept)          am        carb
##    37.32155    -0.02362    -5.35281    23.14584     7.65312    -2.19175
## (Intercept)          am         cyl (Intercept)          am          vs
##    34.52244     2.56703    -2.50096    14.59444     6.06667     6.92937
## (Intercept)          am        qsec
##   -18.88928     8.87633     1.98187
```

The two most interesting regressors are transmission type/weight (am+wt) and transmission type/cylinder alignment (am+vs). The coefficients suggest for am+wt that manual transmissions are actually less gas efficient (by .02 miles per gallon) while the weight of the car matters more (for every 1,000 pounds heavier a car is it gets 5.35 less miles per gallon). It's surprising that the shape of the cylinder arrangement (V shaped or Inline) makes as much of a difference as it suggests here (an Inline/Straight cylinder arrangement gets 6.92 miles more per gallon). We'll now finally run an ANOVA (***AN***alysis ***O***f ***VA***riance) test, starting with predictor am and outcome mpg, then add wt, then adding vs, and one by one the remaining regressors and see at which point the p-values tell us not to include variables.

```
fitA <- lm(mpg~am); fitB <- lm(mpg~am+wt); fitC <- lm(mpg~am+wt+vs); fitD <- lm(mpg~am+wt+v
s+cyl); fitE <- lm(mpg~am+wt+vs+cyl+carb); fitF <- lm(mpg~am+wt+vs+cyl+carb+qsec)
anova <- anova(fitA, fitB, fitC, fitD, fitE, fitF)
anova[6]
```

```
##   Pr(>F)
## 1
## 2 <2e-16 ***
## 3 0.0045 **
## 4 0.0513 .
## 5 0.0749 .
## 6 0.2428
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The p-value crosses above the .05 threshold after Model 3, which as you can see in the code above is fitC. In other words, including the variables in and only up to those found in fitC, am+wt+vs (transmission type, weight, and cylinder alignment) gives us our best model. A residual plot of that model (reference APPENDIX 2, Residual Plot, lm(mpg~am+wt+vs)) shows us that everything looks right. Is variance constant throughout and not dependent on x (no heteroskedasticity)? check. Is there a normal distribution? Check. Looks like a good model fit.

```
## (Intercept)          am          wt          vs
##      30.079       1.491      -3.784       3.615
```

R-Squared

```
## [1] 0.8079
```

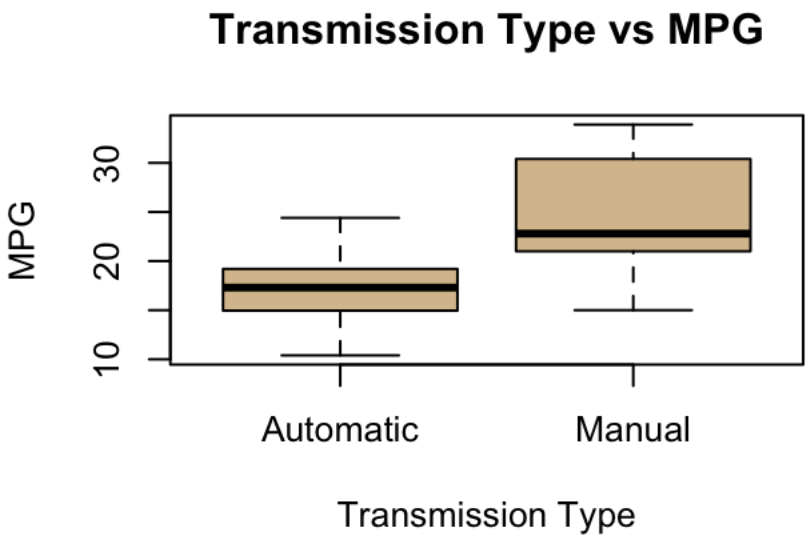
Our best model coefficient shows us that on average we can expect a car with a manual transmission to average 1.49 miles per gallon better than its counterpart with an automatic transmission. The R-squared number of .8079 means this model explains 81% of the variance- in other words, we still have about a 19% rate of uncertainty.

Conclusion

Our exploratory data analysis suggested that cars with manual transmissions are more efficient by 7.24 mpg. A quick hypothesis test proved there really was a difference. To find how much different, each variable was regressed as a predictor variable against mpg, and the variables that had the largest net effect were selected to run a correlation test against am (transmission type). Any pairs that had a correlation > .5 (suggesting the pairs were dependent or otherwise related) were thrown out. The remaining variables were then individually added as predictors to a model with am as the predictor variable against mpg. An ANOVA test was run with those models to see which model was the best (last model for which p-value < .05). The resulting model was a regression with the predictor variables am (transmission type) + wt (weight) + vs (cylinder alignment) against mpg as the outcome. The coefficients 1.49(am), -3.78(wt), and 3.62(vs) give us our best estimates that ***1.)*** a standard/manual transmission is more efficient than an automatic by 1.49 mpg, ***2.)*** for every 1,000 pounds heavier a car gets it loses 3.78 mpg and ***3.)*** an inline/straight cylinder shape gets 3.62 mpg more than a v-shaped arrangement. The intercept of 30.079 tells us the average car averages 30.079 miles per gallon. These numbers explain about 81% of the variance and leave 19% uncertain. Generally, manual transmissions are better for gas efficiency than automatics.

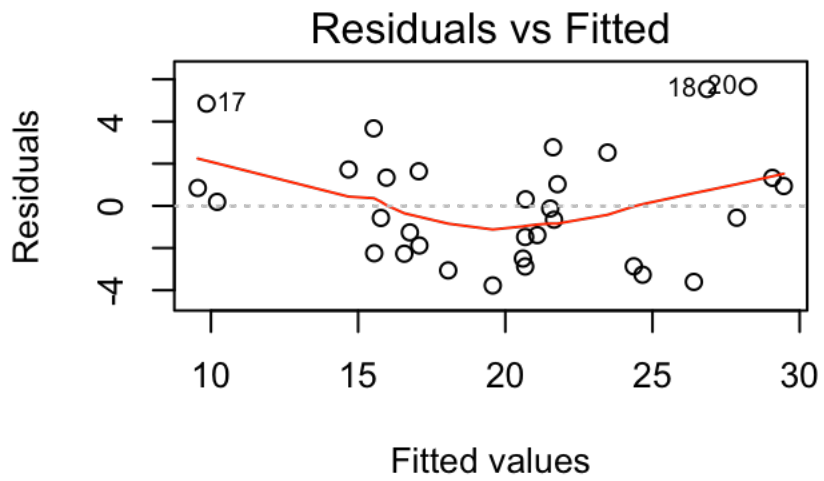
APPENDIX 1 - Boxplot, Transmission Type vs MPG

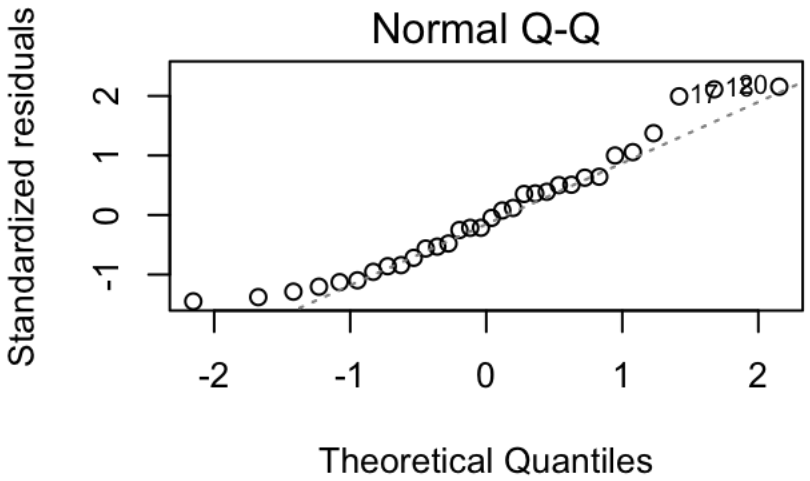
```
par(mfrow=c(2,2))
boxplot(mpg~am, data = mtcars, col = c("tan", "tan"), xlab = "Transmission Type", ylab = "M
PG", main = "Transmission Type vs MPG", xaxt="n")
axis(side=1, at=c(0,1)+1, labels=c("Automatic", "Manual"))
```

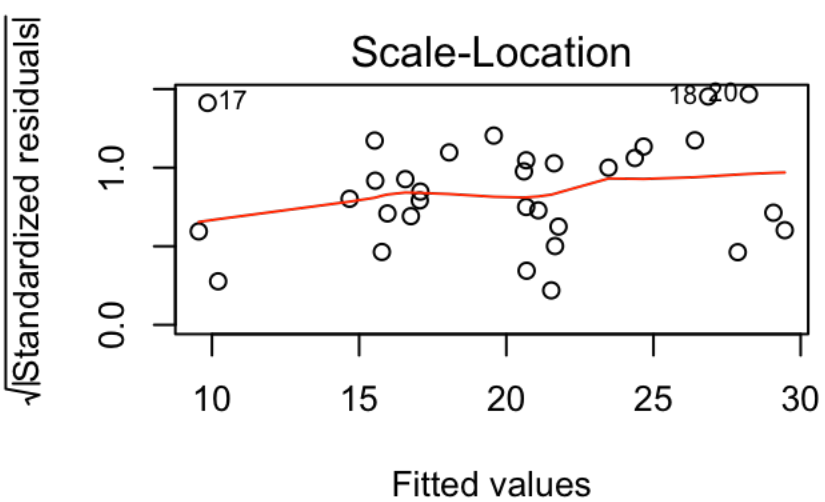


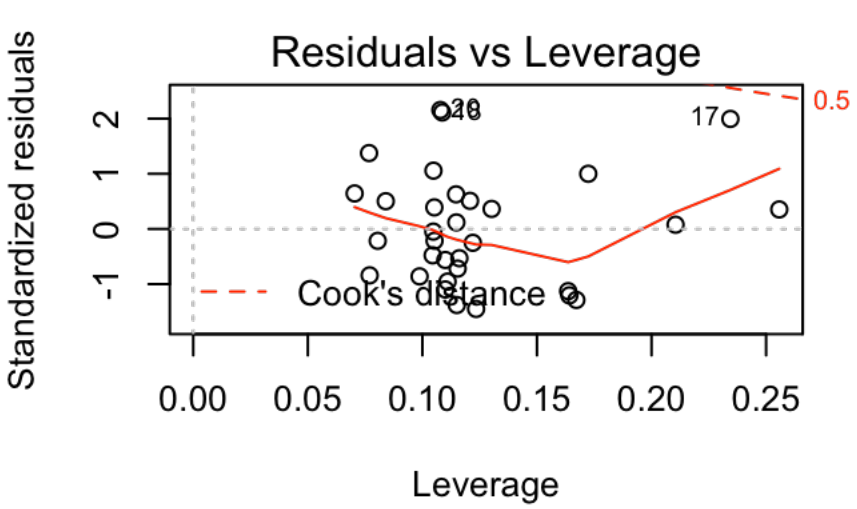
APPENDIX 2 - Residual Plot, lm(mpg~am+wt+vs)

```
par(mfrow = c(2,2))
plot(lm(mpg~am+wt+vs))
```


Residuals vs Fitted
Residuals
Fitted values
17
18
20
10
15
20
25
30
-4
0
4


Normal Q-Q
Standardized residuals
Theoretical Quantiles
17
18
20
-2
-1
0
1
2


Scale-Location
√|Standardized residuals|
Fitted values
17
18
20
10
15
20
25
30
0.0
1.0


Residuals vs Leverage
Standardized residuals
Leverage
Cook's distance
0.5
17
0.00
0.05
0.10
0.15
0.20
0.25
-1
0
1
2