Regression Models - Course Project

Executive Summary

This report evaluates and attempts to quantify the impact of an automatic or manual transmission on a car's fuel efficiency (mpg). It is based on an analysis of the mtcars dataset, using multivariate regression.

The findings of this analysis were as follows:

- By itself, the type of transmission is not a good predictor of fuel efficiency.
- Car weight (wt) and performance, as expressed by qsec, must also be taken into consideration.
- For a given weight and qsec, we would exect a car with manual transmission to get between 0.04573 and 5.825944 more mpg than one with automatic transmission

Exploratory Data Analysis

The mtcars data set is relatively small, having 32 observations of 11 measures, with each observation corresponding to a different model of car. See http://stat.ethz.ch/R-manual/R-devel/library/datasets/html/mtcars.html for a description of the dataset. The dataset is complete, with no N/As.

Examining the correlation matrix for mtcars [Figure A1, in the appendix], we see that mpg is most strongly and negatively correlated with wt , cyl and disp . There is only moderate correlation between mpg and am, the predictor of primary interest to us in this report.

There is also a very strong correlation between cyl and disp, and am is most strongly correlated with wt. [Figure A2] shows a boxplot of wt Vs am

Several transformations were applied to mtcars [Figure A3]. The transformed data frame was called mtcars1.

Model Selection

The strategy adopted for model selection was to add predictors incrementally, comparing iterations for fit (adjusted R-squared), the significance of the predictors, the impact on variance and the distribution of residuals. The selection of individulal predictors was guided by a little knowledge of the domain and some basic Newtonian mechanics (e.g. F=m*a).

On that basis, four initial models were selected for evaluation:

```
fit1 <- lm(mpg ~ am + wt, data=mtcars1)
fit2 <- lm(mpg ~ am + wt + qsec, data=mtcars1)
fit3 <- lm(mpg ~ am + wt + hp, data=mtcars1)
fit4 <- lm(mpg ~ am + wt + qsec+hp, data=mtcars1)
```

Considering model fit, significance of predictors and variance inflation, fit2 looks like the best model. fit2 and fit4 have the highest Adjusted R-squared, but hp is not significant in fit4, which shows much higher variance inflataion than fit2. Some of these statistics are shown in [Figure A4]; Variance Inflation Factors have been ommitted for space considerations.

A second set of models where then compared, each adding one of cyl, disp drat,vs,gear and carb to fit2. None of these models provided a better fit than fit2.

A third set of models were compared, exploring interactions in the predictors of fit2. The one chosen, based on the criteria above was:

```
bestfit<-lm(mpg~wt+am:wt+qsec,data=mtcars1)
```

bestfit has an Adjusted R-squared of 0.8834 and all of its coefficeints are highly significant, with the maximum P value being 0.000209 for wt:am1. Some diagnostic plots were then created for bestfit [Figure A5]. The residuals do not show any strong pattern and are fairly normally distributed.

Conclusion

Although we expect bestfit to provide better predictions of mpg than fit2, it is more difficult to interpret in terms of the research question. It maybe beneficial to interpret both models to see if they cast light on different aspects of the question. The coefficients of both models are:

```
summary(fit2)$coefficients; summary(bestfit)$coefficients;
```

```
##             Estimate Std. Error t value  Pr(>|t|)
## (Intercept)   18.898     0.7194  26.271 2.856e-21
## am1            2.936     1.4109   2.081 4.672e-02
## wt            -3.917     0.7112  -5.507 6.953e-06
## qsec           1.226     0.2887   4.247 2.162e-04
```

```
##             Estimate Std. Error t value  Pr(>|t|)
## (Intercept)  18.6320     0.4997  37.284 2.064e-25
## wt           -3.0907     0.5988  -5.162 1.782e-05
## qsec          0.9415     0.2101   4.480 1.147e-04
## wt:am1       -4.4532     1.0458  -4.258 2.094e-04
```

95% confidence intervals were calculated for the coefficients of primary interest:

```
c(fit2.am1=qt(0.975, 28)*summary(fit2)$coefficients[2,2], bestfit.wtam1=qt(0.975, 28)*summary(bestfit)$
```

```
##      fit2.am1 bestfit.wtam1
##         2.890         2.142
```

- fit2 suggests that for a given weight and qsec, a car with a manual transmission will get between 0.04573 and 5.825944 more mpg than one with automatic transmission.
- bestfit suggests that for a car of average weight and qsec, there is no difference between having a manual or automatic transmission. It also suggests that, for a given qsec, for each 1000lb above the average weight, manual cars will get between -6.595357 and 2.311071 less mpg than automatic cars, but that for each 1000lb below the average weight they will get the same amount more mpg than automatic cars.

The interpretations of fit2 and bestfit seem somewhat at odds with each other. It is worth looking again at [a2]. In the mtcars dataset, automatic cars were 1358 lbs heavier than manual ones on average. Only one automatic car was below the average weight, while the majority of manual cars were. In the light of these facts, the interpretation of bestfit does not seem meaningful for this data set.

Appendix

Figure A1 - mtcars correlation matrix

```
cor(mtcars)
```

```
##            mpg     cyl    disp      hp     drat      wt    qsec      vs      am    gear     carb
## mpg   1.0000 -0.8522 -0.8476 -0.7762  0.68117 -0.8677  0.4187  0.6640  0.59983  0.4803 -0.55093
## cyl  -0.8522  1.0000  0.9020  0.8324 -0.69994  0.7825 -0.5912 -0.8108 -0.52261 -0.4927  0.52699
## disp -0.8476  0.9020  1.0000  0.7909 -0.71021  0.8880 -0.4337 -0.7104 -0.59123 -0.5556  0.39498
## hp   -0.7762  0.8324  0.7909  1.0000 -0.44876  0.6587 -0.7082 -0.7231 -0.24320 -0.1257  0.74981
## drat  0.6812 -0.6999 -0.7102 -0.4488  1.00000 -0.7124  0.0912  0.4403  0.71271  0.6996 -0.09079
## wt   -0.8677  0.7825  0.8880  0.6587 -0.71244  1.0000 -0.1747 -0.5549 -0.69250 -0.5833  0.42761
## qsec  0.4187 -0.5912 -0.4337 -0.7082  0.09120 -0.1747  1.0000  0.7445 -0.22986 -0.2127 -0.65625
## vs    0.6640 -0.8108 -0.7104 -0.7231  0.44028 -0.5549  0.7445  1.0000  0.16835  0.2060 -0.56961
## am    0.5998 -0.5226 -0.5912 -0.2432  0.71271 -0.6925 -0.2299  0.1683  1.00000  0.7941  0.05753
## gear  0.4803 -0.4927 -0.5556 -0.1257  0.69961 -0.5833 -0.2127  0.2060  0.79406  1.0000  0.27407
## carb -0.5509  0.5270  0.3950  0.7498 -0.09079  0.4276 -0.6562 -0.5696  0.05753  0.2741  1.00000
```

Figure A2 - box plot of weight by transmission type

```
par(mfrow=c(2,2));
boxplot(mtcars$wt, xlab="All Cars", ylab="Weight")
boxplot(wt~am,data=mtcars,xlab="Has Manual Transmission", ylab="Weight")
```

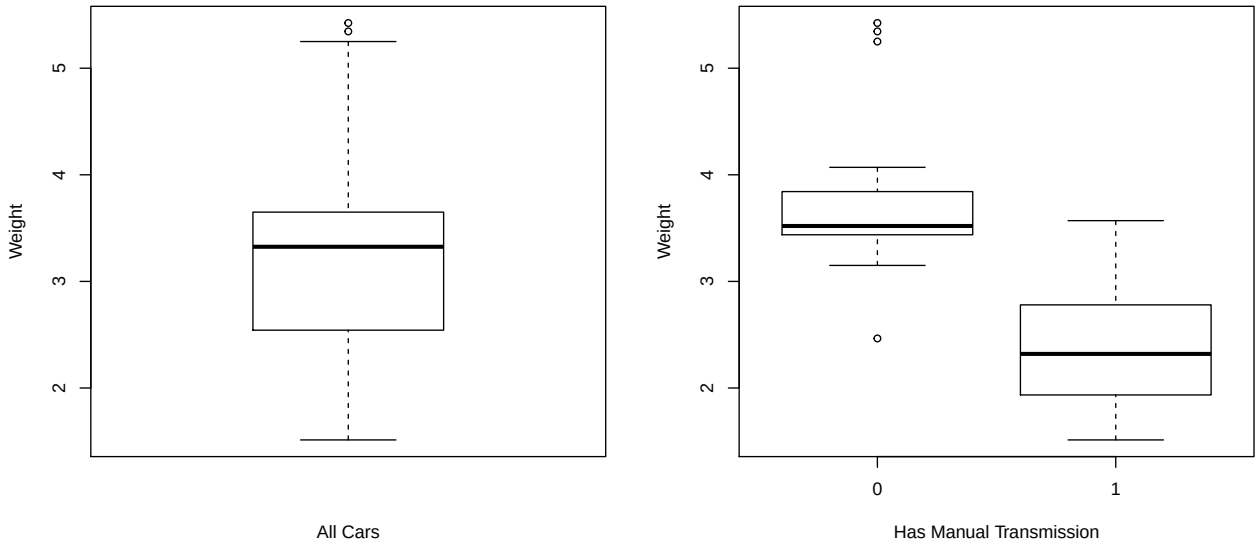

Figure A3 - mtcars data transformations

```
mtcars1<-mtcars
mtcars1$vs<-factor(mtcars$vs)
mtcars1$am<-factor(mtcars$am)
```

```
mtcars1$disp<-mtcars$disp-mean(mtcars$disp)
mtcars1$hp<-mtcars$hp-mean(mtcars$hp)
mtcars1$drat<-mtcars$drat-mean(mtcars$drat)
mtcars1$wt<-mtcars$wt-mean(mtcars$wt)
mtcars1$qsec<-mtcars$qsec-mean(mtcars$qsec)
```

Figure A4 Statistics used to evaluate fits 1-4

Coefficient P Values

```
summary(fit1)$coefficients[,4]; summary(fit2)$coefficients[,4]; summary(fit3)$coefficients[,4];
```

```
## (Intercept)         am1          wt
##   9.640e-21   9.879e-01   1.867e-07
```

```
## (Intercept)         am1          wt        qsec
##   2.856e-21   4.672e-02   6.953e-06   2.162e-04
```

```
## (Intercept)         am1          wt          hp
##   1.594e-21   1.413e-01   3.574e-03   5.464e-04
```

Adjusted R Squared

```
summary(fit4)$coefficients[,4]
```

```
## (Intercept)         am1          wt        qsec          hp
##   7.068e-21   4.579e-02   1.141e-03   7.573e-02   2.231e-01
```

```
c(summary(fit1)$adj.r.squared,summary(fit2)$adj.r.squared,summary(fit3)$adj.r.squared,summary(fit4)$adj
```

```
## [1] 0.7358 0.8336 0.8227 0.8368
```

Variance Inflation Factors

```
vif(fit1);vif(fit2);vif(fit3);vif(fit4)
```

```
##    am    wt
## 1.921 1.921
```

```
##    am    wt  qsec
## 2.541 2.483 1.364
```

```
##    am    wt    hp
## 2.271 3.775 2.088
```

```
##    am    wt  qsec    hp
## 2.542 3.965 3.216 4.922
```

Figure A5 - Diagnostic plots for bestfit

```
par(mfrow=c(2,2)); plot(bestfit)
```