Regression Analysis: Is an Automatic or Manual Transmission better for MPG?

Executive Summary

This report addresses the question of whether an automatic or manual transmission is better for MPG. In answering this question, I attempt to quantify any MPG difference between automatic and manual transmissions. After exploration and model fitting, I conclude that a manual transmission is better for MPG and the difference is signficant. My final model indicates that holding all else constant, I expect a manual transmission to have 2.94 higher MPG that an automatic car. It is important to note that while transmission type is significant, the most significant factor in my model is weight. For every 1000lb increase in weight, I expect that car to drop 3.92 MPG.

Exploratory Analysis

Load All Necessary Packages

```
library(GGally);library(ggplot2);library(printr);library(car)
```

Data Description

A quick glance at the data set reveals the structure and the data available for our model.

```
head(mtcars, n = 2)
```

	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
Mazda RX4	21	6	160	110	3.9	2.620	16.46	0	1	4	4
Mazda RX4 Wag	21	6	160	110	3.9	2.875	17.02	0	1	4	4

```
dim(mtcars)
```

```
## [1] 32 11
```

First, I plot MPG by transmission and see if there is a difference visually.

```
g <- ggplot(mtcars, aes(x = factor(am), y = mpg, fill = factor(am)))+
  geom_boxplot()+ coord_flip()+ xlab("Transmission Type")+ ylab("MPG")+
  ggtitle("Miles per Gallon by Transmission Type")+
  scale_fill_discrete(name = "Transmission Type", labels = c("Automatic", "Manual"))
```

Using the boxplot(Figure A), my initial hypothesis is that a significant difference exists, with manual transmissions having higher MPG than automatic transmissions.

I perform a Student's T-Test to test for a statistically significant different between the means of the two groups–in this case automatic and manual transmission. Our null hypothesis is that there is no difference between the means.

```
t <- t.test(mtcars$mpg ~ factor(mtcars$am))
```

With a p value of of 0.001 it is very unlikely a difference this large is due to chance and I reject the null hypothesis.

Model Fitting

Single Variable Linear Model

My first model is a single variable model using transmission as the only variable.

```
fit_am <- lm(mpg~ factor(am), mtcars)
fit_am
```

```
##
## Call:
## lm(formula = mpg ~ factor(am), data = mtcars)
##
## Coefficients:
## (Intercept)  factor(am)1
##      17.147        7.245
```

This simple model uses automatic transmissions as the baseline. The intercept of 17.147 is the mean mpg of cars with automatic transmissions. An am1 coefficient of 7.245 means a car, holding all else in the model constant, gets 7.245 mpg higher with a manual transmission.

With a p-value of .0003, this is a highly significant result and seems to indicate transmission is a key variable in fuel efficiency. Looking at our adjusted r-squared, however, this model only explains 0.338 of the variation. With a single variable, this model could be confounding the effects of other variables and attributing them to transmission. More exploration is needed.

Multiple Variable Regression Model

```
mfit1 <- lm(mpg ~ cyl+disp+hp+drat+wt+qsec+factor(vs)+factor(am)+gear+carb , mtcars)
mfit1
```

```
##
## Call:
## lm(formula = mpg ~ cyl + disp + hp + drat + wt + qsec + factor(vs) +
##     factor(am) + gear + carb, data = mtcars)
##
## Coefficients:
## (Intercept)          cyl         disp           hp         drat
##    12.30337     -0.11144      0.01334     -0.02148      0.78711
##          wt         qsec  factor(vs)1  factor(am)1         gear
##    -3.71530      0.82104      0.31776      2.52023      0.65541
##        carb
##    -0.19942
```

I start by fitting a model with all variables. Looking at the summary, none of the variables are significant at the .05 level. However, this is probably due to correlation amongst the variables. It makes intuitive sense that variables like cyl and hp would be highly correlated. Looking at a correlation table we can see the correlation between variables.

```
round(cor(mtcars), 3)
```

	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
mpg	1.000	-0.852	-0.848	-0.776	0.681	-0.868	0.419	0.664	0.600	0.480	-0.551
cyl	-0.852	1.000	0.902	0.832	-0.700	0.782	-0.591	-0.811	-0.523	-0.493	0.527
disp	-0.848	0.902	1.000	0.791	-0.710	0.888	-0.434	-0.710	-0.591	-0.556	0.395
hp	-0.776	0.832	0.791	1.000	-0.449	0.659	-0.708	-0.723	-0.243	-0.126	0.750
drat	0.681	-0.700	-0.710	-0.449	1.000	-0.712	0.091	0.440	0.713	0.700	-0.091
wt	-0.868	0.782	0.888	0.659	-0.712	1.000	-0.175	-0.555	-0.692	-0.583	0.428
qsec	0.419	-0.591	-0.434	-0.708	0.091	-0.175	1.000	0.745	-0.230	-0.213	-0.656
vs	0.664	-0.811	-0.710	-0.723	0.440	-0.555	0.745	1.000	0.168	0.206	-0.570
am	0.600	-0.523	-0.591	-0.243	0.713	-0.692	-0.230	0.168	1.000	0.794	0.058
gear	0.480	-0.493	-0.556	-0.126	0.700	-0.583	-0.213	0.206	0.794	1.000	0.274
carb	-0.551	0.527	0.395	0.750	-0.091	0.428	-0.656	-0.570	0.058	0.274	1.000

We have significant collinearity. This can mislead as to the importance of predictors and can lead to imprecise coefficients. The recommended solution is removing variables which are, in effect, trying to do the same job.

The major decision in this analysis is which method to use for model variable selection. Generally, criterion-based methods are recommended over stepwise methods as they search over a wider space and don't use hypothesis testing to choose between models.

Although Mallow's Cp criterion achieved the higest adjusted r-squared, it added two variables with significant collinearity for a very minimal gain in adjusted r-squared. I will use Akaike Information Criterion(AIC) as the criteria for evaluating how well my model is fit. Our goal is to minimize AIC.

```
step <- step(mfit1, direction = 'both')
```

```
## lm(formula = mpg ~ wt + qsec + factor(am), data = mtcars)
```

```
final_model <- lm(mpg ~ wt + qsec + factor(am), data = mtcars)
no_am <- lm(mpg ~ wt + qsec, data = mtcars)
summary(final_model)
```

```
##
## Call:
## lm(formula = mpg ~ wt + qsec + factor(am), data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4811 -1.5555 -0.7257  1.4110  4.6610
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   9.6178     6.9596   1.382 0.177915
## wt           -3.9165     0.7112  -5.507 6.95e-06 ***
## qsec          1.2259     0.2887   4.247 0.000216 ***
## factor(am)1   2.9358     1.4109   2.081 0.046716 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 2.459 on 28 degrees of freedom
## Multiple R-squared:  0.8497, Adjusted R-squared:  0.8336
## F-statistic: 52.75 on 3 and 28 DF,  p-value: 1.21e-11
```

```
anova(no_am, final_model)
```

This model has a respectable adjusted r-squared of 0.834. All values in the model are significant at the .05 level. Here, holding all other variables constant, we expect 2.94 MPG greater from a manual transmission than an automatic. Note, however, that transmission is the least significant variable in our model. Both weight(for obvious reasons) and qsec(likely capturing some aspect of engine power) are far more significant when looking at p-values. The p-value from the analysis of variance between the model with transmission and without is 0.047. This indicates the inclusion of transmission makes the model better, but again, it is a very close call. A larger dataset would be a good starting point for further analysis.

Residual plots in the appendix check for homoscedasticity, linearity, normality and outliers. Plotting our residuals vs our fitted values (Figure B) we see constant variance in our residuals with no fanning or nonlinear trends. I also plot residuals against our predictors and look for the same sort of inconsistant variance or nonlinearity. Overall, our residual plots reveal nothing to be alarmed about

Appendix:

Figure A: MPG by Transmission Type

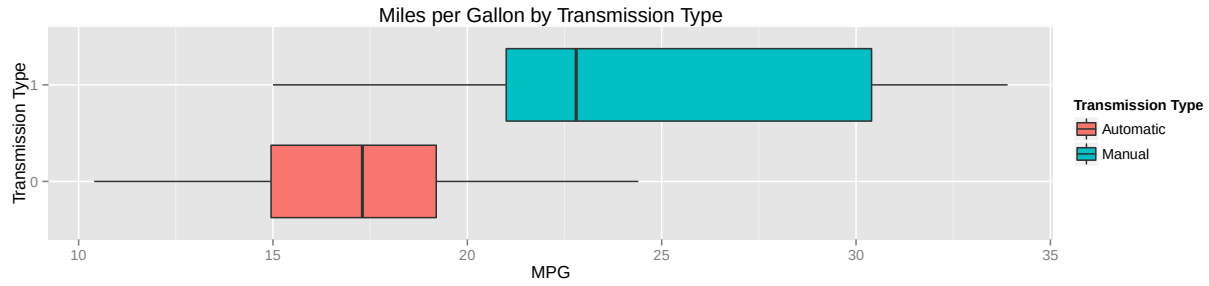

Figure B: Residuals vs Predictors and Residuals vs Fitted Values

```
par(mfrow = c(2,2))
residualPlots(step)
```

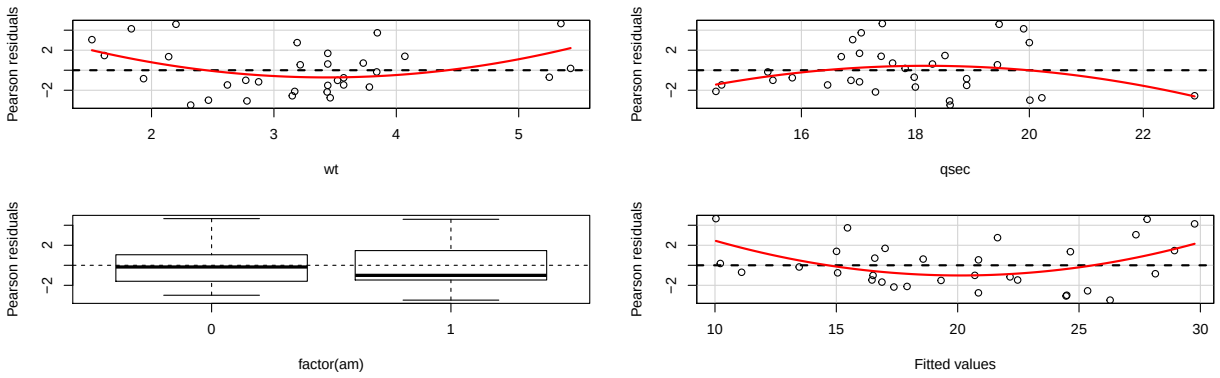

Figure C: Normality of Residuals

```
qqPlot(step)
```

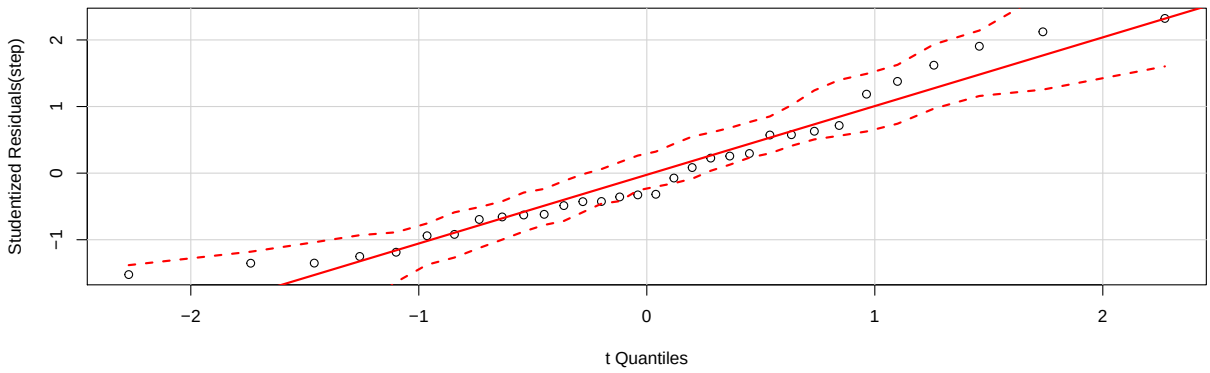

Figure D: Outlier and High Leverage Points

```
influenceIndexPlot(step, id.n = 3)
```