MTCARS MPG Analysis using Regression Models & Statistical Inference

Executive Summary This report will examine the relationships between variables contained in the 'mtcars' data set and the MPG performance of the cars. Specifically, the report will attempt to answer the following questions:

1. Is an automatic or manual transmission better for MPG?
2. Quantify the MPG difference between the automatic and manual transmission.

About the data The mtcars dimension includes 32 observations and 11 variables. See the help(mtcars) in R for details of the dataset variables.

There are 3 variables (**mpg**, **hp** and **qsec**) which I assume are **outcomes of all other variables**. The variable list also has variables (**cyl**, **vs**, **am**, **gear** & **carb**) in numeric form which should be treated as factors. So prior to any analysis, we convert these into factors.

Exploratory Analysis From a previous analysis in class, we know weight will be a factor in the **mpg** performance. The **am** variable will be included since the point of this report is to investigate the impact of the transmission type.

See **Appendix A** for the relationships between these 2 regressors against **mpg**. The quantiles suggest some difference between the transmission types but their mpg range also shows overlaps (**See Appendix B**).

Model Selection Strategy To select the appropriate regressors, we take the following steps:
(1) Use VIF values to filter highly colinear regressors. **(2)** identify variables using corelation that can represent others. **(3)** formulate the models using the final regressor selection.Note: **wt** and **am** will be included regardless of their VIF and colinear results. **(4)** Finally, use the Anova test to determine which models have better fit.

(1) Variable Inflation Test Use **sqrt(vif)** to identify variables with inflationary impact on other variables.

```
##      cyl     disp       hp     drat       wt     qsec       vs       am
## 3.364380 7.769536 5.312210 2.609533 4.881683 3.284842 2.843970 3.151269
##     gear     carb
## 2.670408 1.862838
```

This list suggests **disp** (7.8), **hp** (5.3), **wt**(4.8) have high colinearity against all other variables. Remove disp (highest). Also remove performance indicator types hp and qsec. And check the vif again.

Note on hp and qsec variable context: both variables measure performance as a result of other design inputs. One doesn't specify, for example a 10-second qsec as part of a design. It is rather a goal or a result of the car design. So, it does not have practical predictive value but can indicate performance of another (e.g. mpg, hp), post design. Therefore, both variables are excluded from the model.

```
##       wt      cyl     drat       vs       am     gear     carb
## 2.629214 2.146443 2.603951 2.601238 2.596990 2.083641 1.420435
```

VIF values for the remaining variables have all dropped below 3.

(2) Colinearity Test Use cyl to verify representation of other variables.

```
##        mpg         wt        cyl       drat         vs         am
## -0.8521620  0.7824958  1.0000000 -0.6999381 -0.8108118 -0.5226070
##       gear       carb
## -0.4926866  0.5269883
```

cyl's correlation to **wt, drat, vs, am** and **carb** is above 0.5 ($|\text{cor}| > 0.5$). I contend that using **cyl** can represent these other regressors which allows removing them without significantly increasing the residuals.

(3) Model Selected Build the model with variables, **wt + am + cyl** as regressors. Then compare against models that include other regressors.

```
f1 <- lm(mpg ~ wt + am,data=x)
f2 <- update(f1,mpg ~ wt + am + cyl)
f3 <- update(f1,mpg ~ wt + am + cyl + vs)
f4 <- update(f1,mpg ~ wt + am + cyl + carb + drat)
anova(f1,f2,f3,f4)
```

(4) Anova Test The anova results (see **Appendix C**)indicate that **f2 (mpq ~ wt + am + cyl)** is a minimal adequate model with p-value (0.007196) indicating significant at $\alpha = 0.01$. Adding more regressors negates the significance.

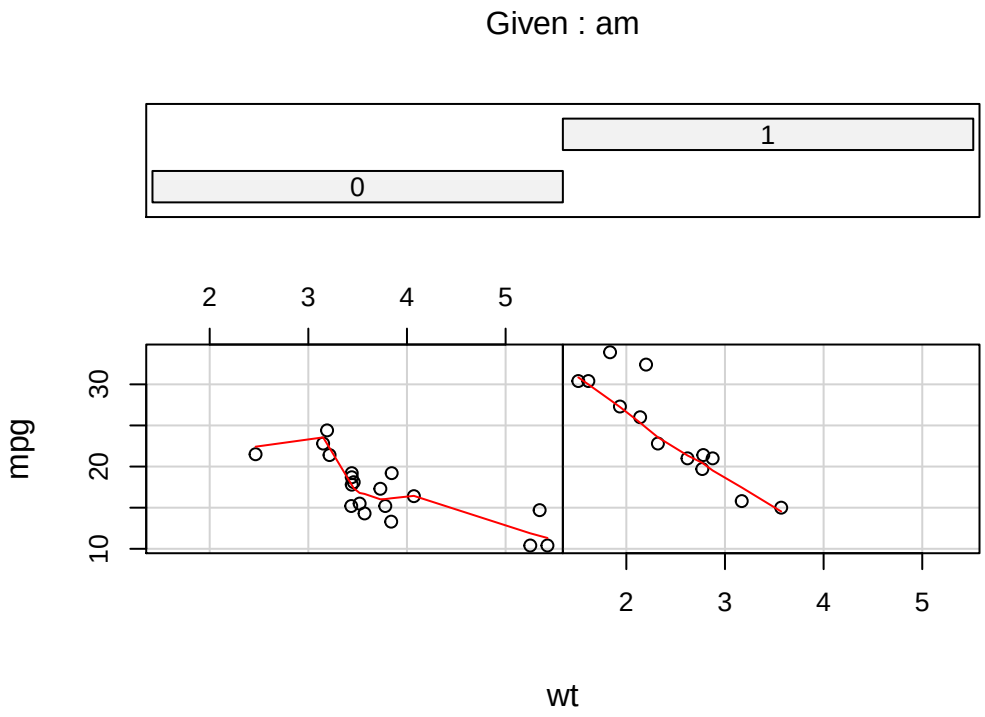
Validating the Model Fitness First, the model's R^2 explains 83.75% of the mpg performance (**see Appendix F**). Second, the residual plot (see **Appendix D**) shows adequate coverage with no discernable pattern that may indicate a missing confounding variable. And third, the residual QQ plot (**Appendix E**) follows the diagonal line suggesting a nearly normal residual distribution.

Interpreting the Model See **Appendix F** for the Model Summary values.

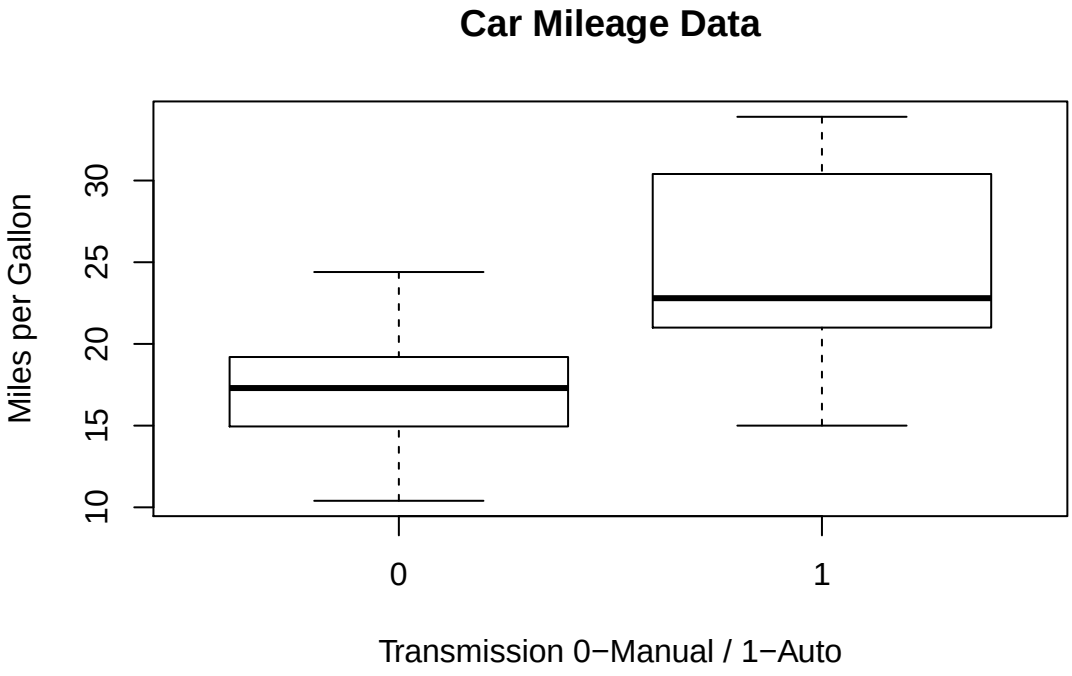
The model calculates the combined intercept for factors, Automatic Transmission (am0) and 4 cylinders (cyl4), at 33.75mpg. This is the empirical mean for both am0 and cyl4. **Manual transmission improves the mpg performance by only 0.15mpg. But this is not significant and may be due to chance**. 6-cylinder (cyl6) and 8-cylinder (cyl8) factors on the other hand, decreases mpg performance by 4.26mpg and 6.08mpg respectively. And every 1,000 lbs increase in weight results in a 3.15mpg loss.

So, does a manual transmission result in better mileage than an automatic one? In this model, the answer is no. Other confounding variables explains mpg performance more than the transmission type.

Appendix A - Data Exploration - Relationship of wt + am to mpg



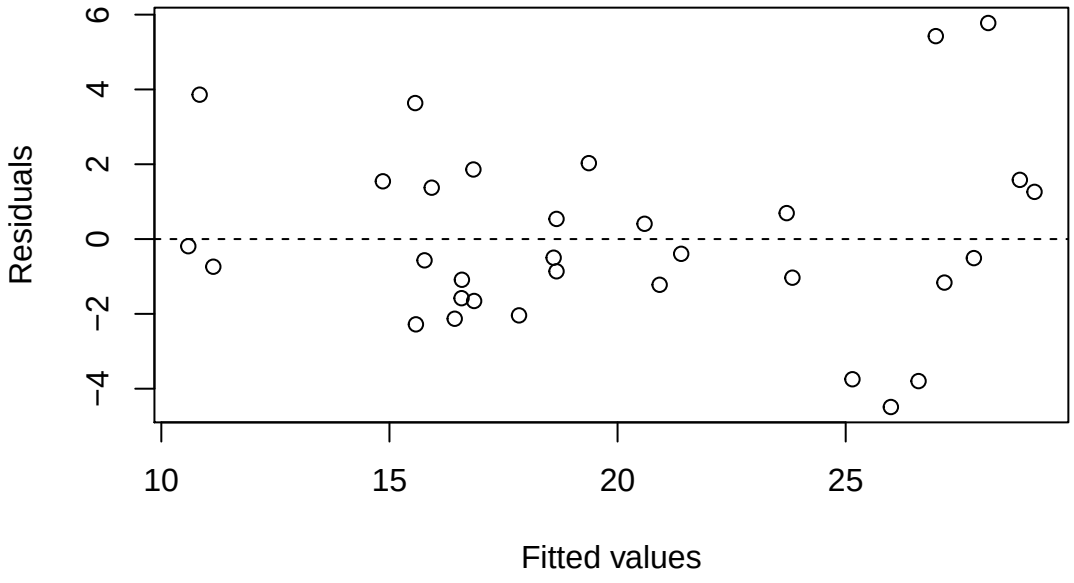
Appendix B - Data Exploration - Transmission Box plot against MPG



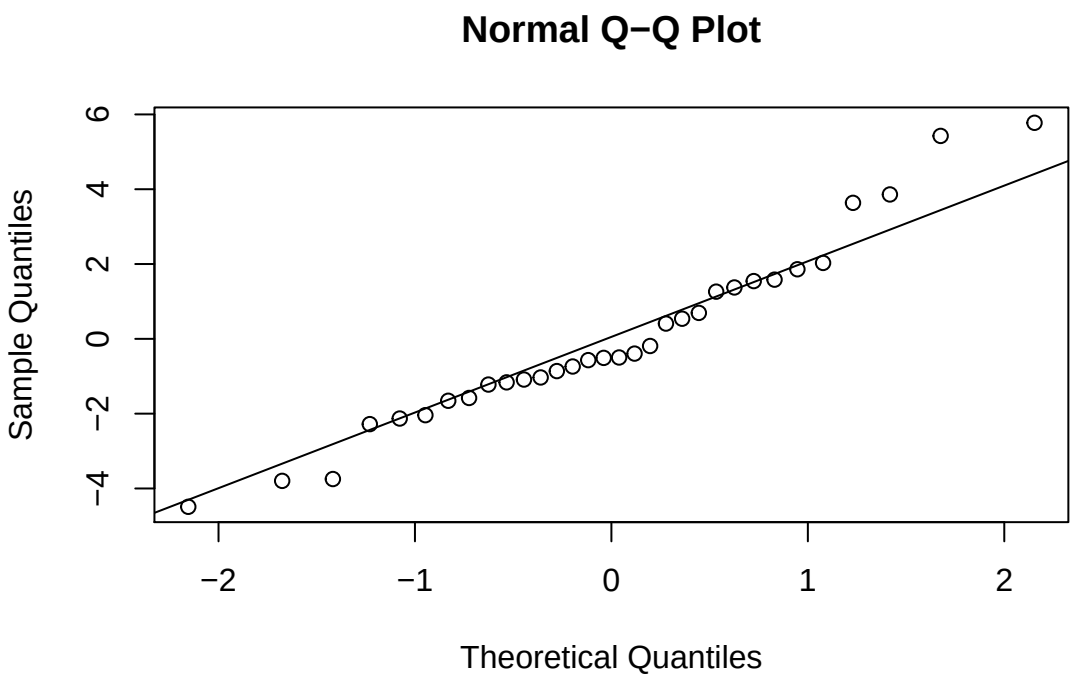
Appendix C - Anova Results

```
## Analysis of Variance Table
##
## Model 1: mpg ~ wt + am
## Model 2: mpg ~ wt + am + cyl
## Model 3: mpg ~ wt + am + cyl + vs
## Model 4: mpg ~ wt + am + cyl + carb + drat
##   Res.Df    RSS Df Sum of Sq      F   Pr(>F)
## 1     29 278.32
## 2     27 182.97  2    95.351 6.2987 0.007196 **
## 3     26 180.02  1     2.945 0.3890 0.539525
## 4     21 158.95  5    21.073 0.5568 0.731662
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Appendix D - Residual Plot for Model mpg ~ wt + am + cyl



Appendix E - QQ Plot of Residuals



Appendix F - Linear Model Summary of mpg ~ wt + am + cyl

```
##
## Call:
## lm(formula = mpg ~ wt + am + cyl, data = x)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.4898 -1.3116 -0.5039  1.4162  5.7758
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  33.7536     2.8135  11.997  2.5e-12 ***
## wt           -3.1496     0.9080  -3.469  0.00177 **
## am1           0.1501     1.3002   0.115  0.90895
## cyl6         -4.2573     1.4112  -3.017  0.00551 **
## cyl8         -6.0791     1.6837  -3.611  0.00123 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.603 on 27 degrees of freedom
## Multiple R-squared:  0.8375, Adjusted R-squared:  0.8134
## F-statistic: 34.79 on 4 and 27 DF,  p-value: 2.73e-10
```