Motor Trend Analysis

Conner McBride

Executive Summary

This report provides an analysis of the `mtcars` dataset and specifically looks for any significant association between the type of transmission (automatic or standard) a vehicle is equipped with and fuel efficiency as measured in miles/gallon. The methods of analysis primarily use linear regression but also incorporate exploratory plots, investigation of correlation to assist in feature selection, as well as statistical inference (the relevant R code, plots, and calculations can be found in the appendices).

Results from the analysis show that while in the `mtcars` dataset fuel efficiency has an apparently significant association with transmission type, it is only marginal and not directly linked. When controlling for other factors such as vehicle weight, number of cylinders, and displacement the association between fuel efficiency and transmission type all but disappears.

Exploratory Data Analysis

The `mtcars` dataset has 11 features describing 32 unique make and model combinations of vehicles. To get an idea of the interactions between the different features, we'll start the exploratory analysis with a paired plot (*fig. 1*).

There appears to be a strong association between transmission type and fuel efficiency. However, transmission type is also strongly correlated with several other features (number of cylinders, displacement, and weight among others) all of which may better explain differences in fuel efficiency and need to be addressed as confounders.

A density plot (*fig.2*) gives a better idea of how fuel efficiency measure are distributed by transmission type and shows that, at least marginally, the correlation is significant. We can construct a simple single variable linear model with miles/gallon as the response variable and transmission type as the predictor to see.

Analysis

The naive, single-variable model ($Y_{mpg} = \beta_0 + \beta_1 X_{am}$) confirms what was demonstrated in the density plot that the two features - gas mileage and transmission type - are signficantly associated. The model also quantifies the linear difference in β_1, the slope coefficient, which says that on average automatic transmission vehicles get 7.24 fewer miles/gallon compared to manual transmission cars. The model appears to demonstrate a significant association, but as the pairs plot demonstrated there are several potential confounders.

Two of the features most closely correlated with transmission, number of cylinders and displacement, would intuitively be close covariates since the latter is naturally a function of the former. A simple scatter plot (*fig. 3*) shows a clear pattern and confirms this intuition. While there are some automatic cars with 4 cylinders and small displacements there are far more in with large diplacements and 8 cylinders, both categories correlated with low fuel efficiency.

A similar pattern can be seen with weight. A plot of displacement by weight with points colored by number of cylinders shows how closely the three variables are correlated (*fig. 4*). For this reason, only the one that is most closely associated with fuel efficiency will be used. The correlations for weight, displacement, and number of cylinders respectively are -0.868, -0.848, and -0.852.

Our intuition based on the above is that transmission type will be insignificant when weight is added to the model.

Using an additive error, multivariable linear model ($Y_{mpg} = \beta_0 + \beta_{am}X_{tr} + \beta_{wt}X_{wt}$) with weight added as a predictor the coefficient for automatic transmission is near zero (β_{atr} = -0.024) indicating a neglible difference in the average fuel efficiency between automatic and manual transmission vehicles featured in the dataset when weight is accounted for.

A residuals and qq plot(*fig.5*) reveal a few candidate outliers with residuals approximately equal to or greater than the coefficient β_1 we had in the naive model and some minor issues with normality in the distribution of the residuals.

Table 1: Comparison of Outliers & Grouped Means

	cyl	disp	hp	drat	wt	qsec	gear	carb
Means 4cyl	4	105.136	82.636	4.071	2.286	19.137	4.091	1.545
Means 8cyl	8	353.100	209.214	3.229	3.999	16.772	3.286	3.500
Chrysler Imperial	8	440.000	230.000	3.230	5.345	17.420	3.000	4.000
Fiat 128	4	78.700	66.000	4.080	2.200	19.470	4.000	1.000
Toyota Corolla	4	71.100	65.000	4.220	1.835	19.900	4.000	1.000

Comparing the values for the three outliers with the grouped summaries, a few things stand out. The three cars vary from the mean in their displacement (and consequently horsepower) and weight most notably. Adding displacement to the model might draw these outliers in, but as we've already detected, that variable is a covariate of weight already. Instead, we'll look at quarter-mile time (*qsec*) to see if this measure of acceleration captures some of the variation not explained by weight.

A scatter plot (*fig.6*), as a preliminary diagnostic, shows the relationship between fuel efficiency and acceleration with weight (rounded to the nearest integer) encoded with color. The regression line runs parallel to a single weight class, suggesting that it's nearly orthogonal to the regression line found in the model that includes weight. We'll contruct a third model that includes acceleration.

An analysis of variance test (ANOVA) shows a significant difference between our second model and the new model that includes acceleration ($F_{av} = 18.034$, $p_{av} = 0.0002$). A plot of the new residuals (*fig.7*) shows an improvement in all model assessment metrics, although the three cars that motivated the new model remain outliers. A plot of the standardized residuals (*fig.7*) shows a pattern hinting at a missing component in the model and may point to a explanatory feature missing from the data.

Outliers - Leverage & Influence

Table 2: Coefficients of Finalized Model

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	13.668	6.701	2.040	0.051
factor(am)1	2.198	1.350	1.629	0.115
wt	-4.640	0.731	-6.349	0.000
qsec	1.136	0.271	4.188	0.000

```
##
## Call:
## lm(formula = mpg ~ factor(am) + wt + qsec, 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
## factor(am)1   2.9358     1.4109   2.081 0.046716 *
## wt           -3.9165     0.7112  -5.507 6.95e-06 ***
## qsec          1.2259     0.2887   4.247 0.000216 ***
## ---
## 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
```

As a final diagnotic, we'll look once more at the three outliers to see if they have an undue influence on the regression line. After running three tests, The *Chrysler Imperial* has the maximum value in two of them *dfbetas* and *dffits*. The other two don't appear to influence the regression line as significantly. For this reason, we'll eliminate the *Chrysler* and refit our last model. This will be the finalized model.

Conclusions

The findings suggest that, for reasons that aren't clear in the data, transmission type is also closely related to the three confounders. **In response to the motivating questions posed by *Motor Trend*, within the context of the `mtcars` dataset and without controlling for other variables, automatic transmission is associated with worse gas mileage not better. However, transmission type has almost no predictive power when other variables are included in the model.**

Table 3: Predictive Sum of Squares by Model

Model	Predictors	PRESS
Model 1	am	830.332
Model 2	am + wt	351.892
Model 3	am + wt + qsec	231.303
Rmv Outlier	am + wt + qsec	188.596

Using predictive sum of squares as a comparative measure of model performance, the progressive improvement in models is clear.

Appendices

Appendix i: Figures

Figure 1

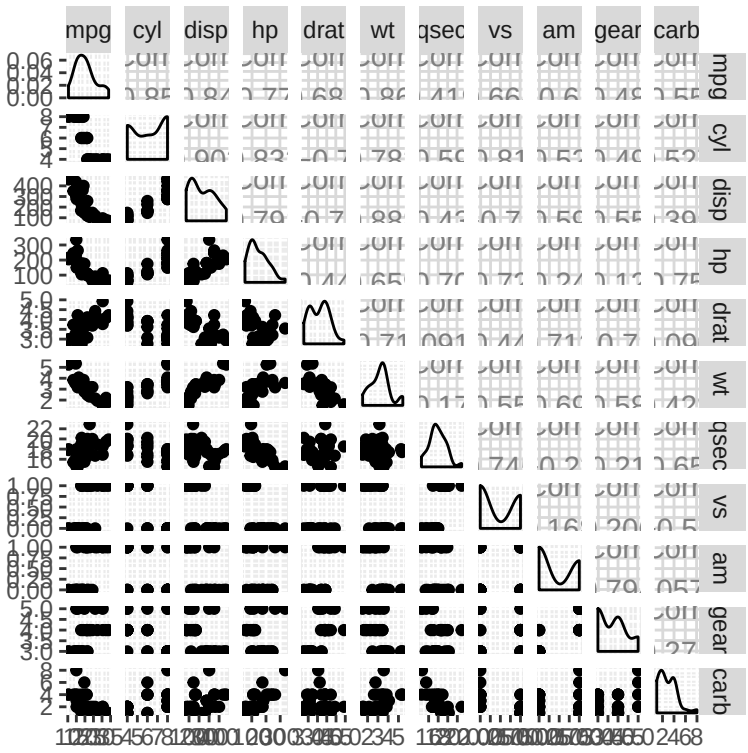

mpg
cyl
disp
hp
drat
wt
qsec
vs
am
gear
carb

Figure 2

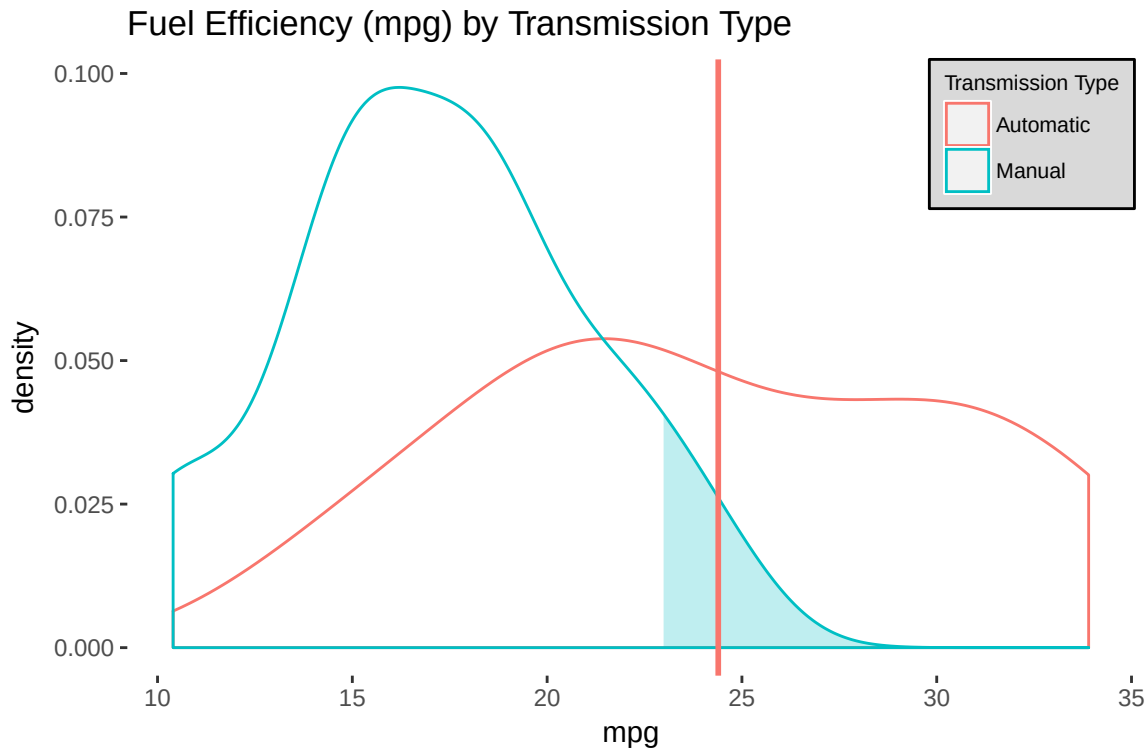


Figure 3

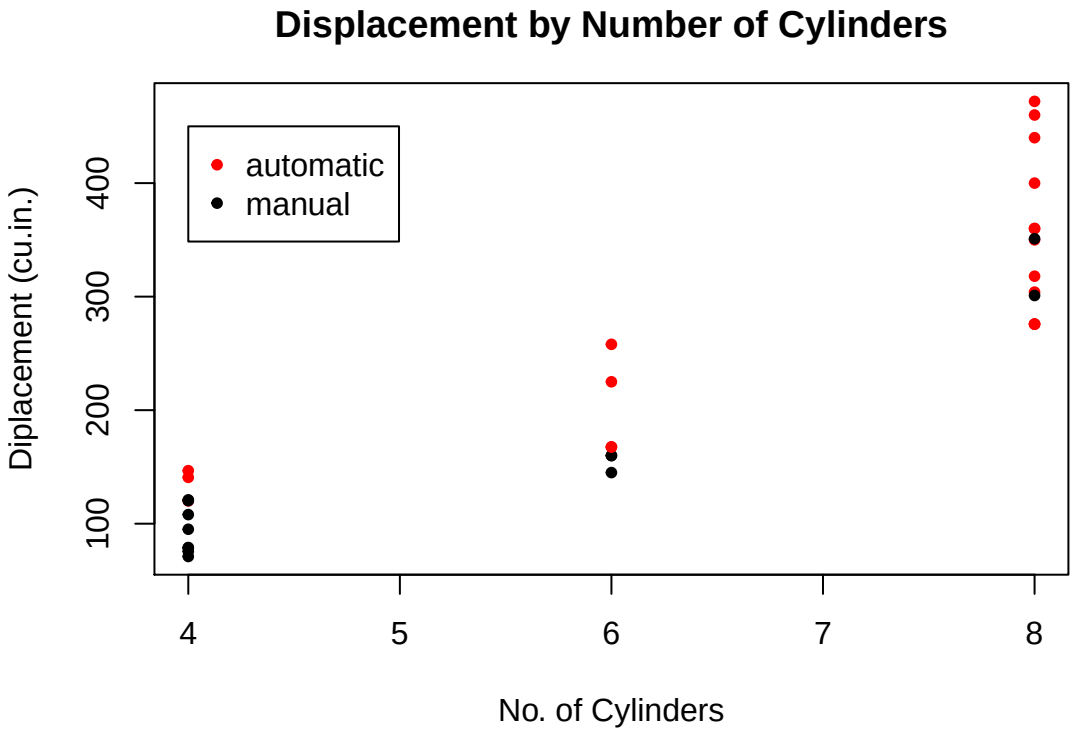


Figure 4

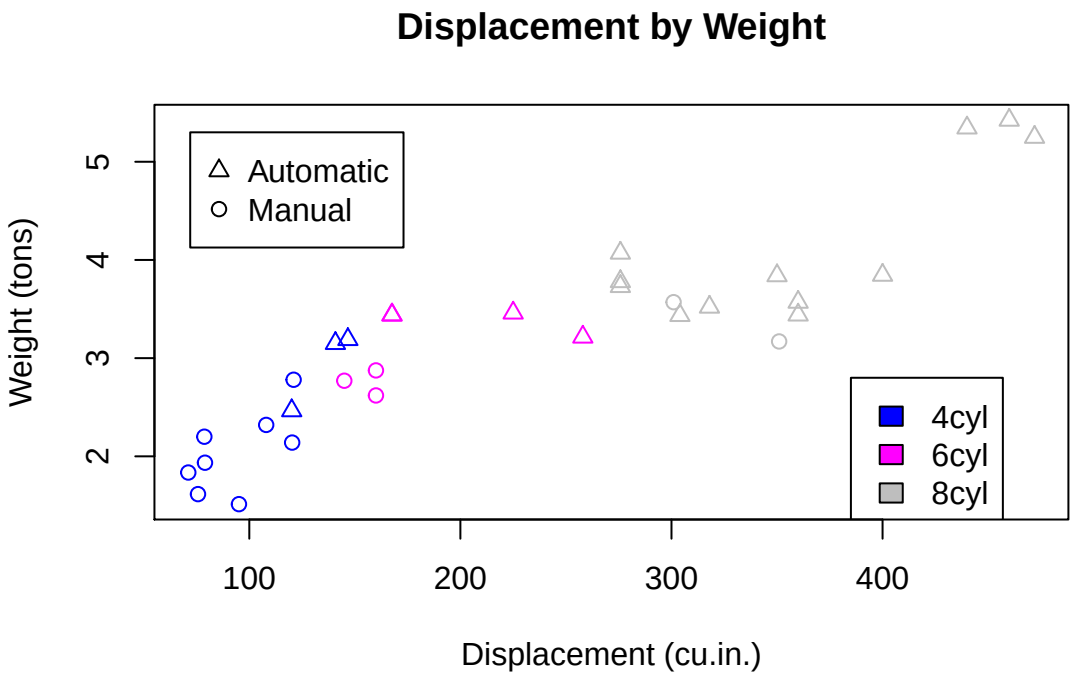


Figure 5

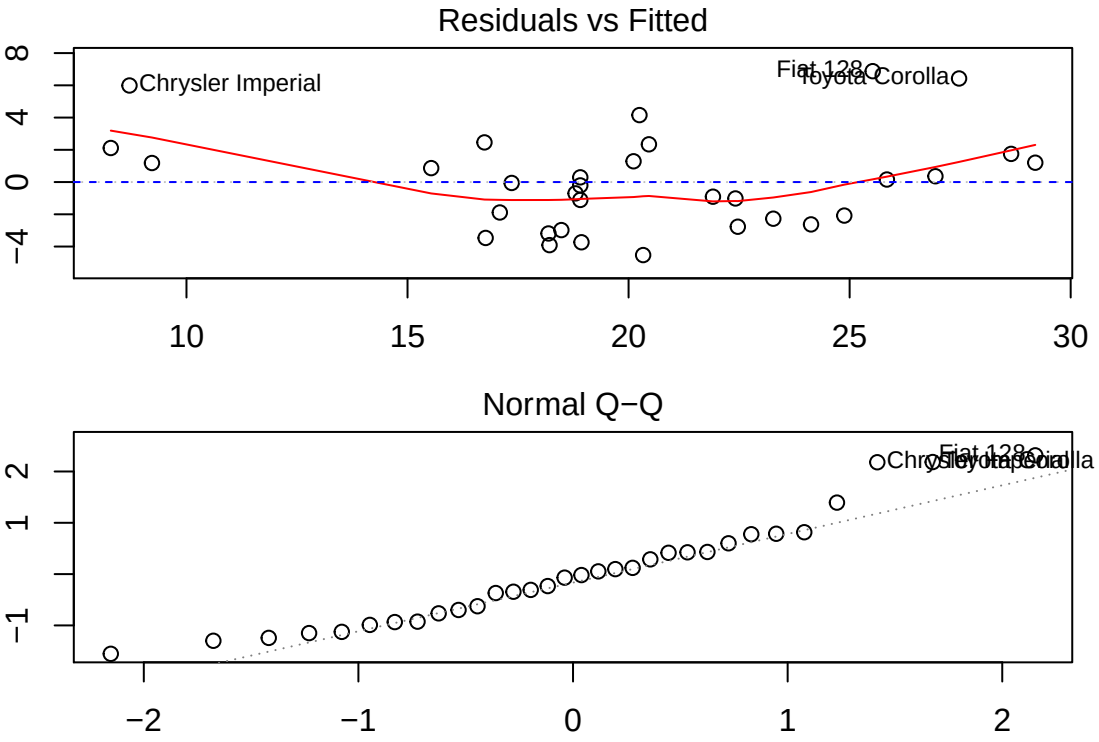


Figure 6

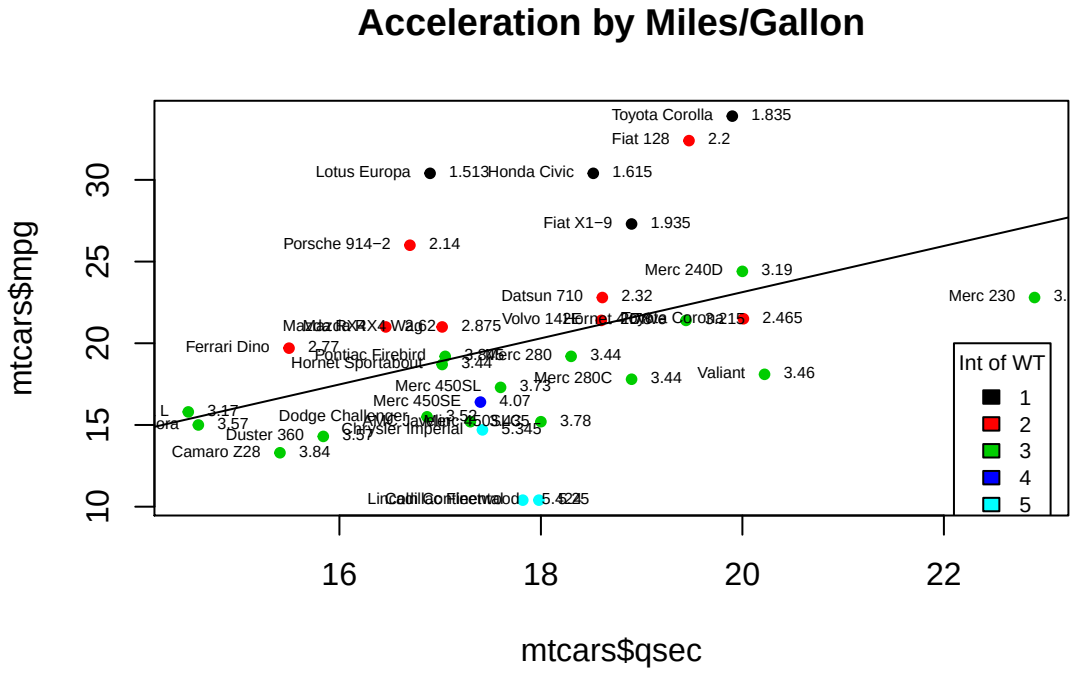
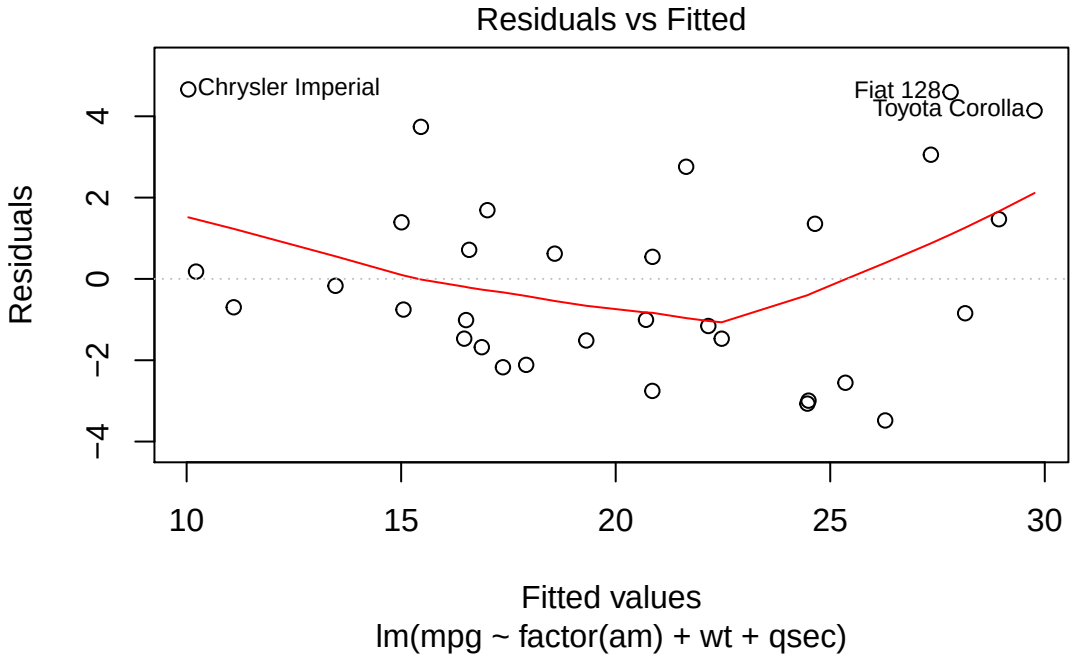


Figure 7



Appendix ii: Code

```
knitr::opts_chunk$set(echo = TRUE)
# load data
library(datasets)
data("mtcars")

# load requisite packages
library(dplyr)
library(GGally)
library(ggplot2)
library(knitr)
library(tibble)
library(xtable)
# initial linear model with single predictor `am`
fit1 <- lm(mpg ~ am, data=mtcars)
# linear model that now includes weight as a controlling factor
fit2 <- lm(mpg ~ factor(am) + wt, data=mtcars)
## create dataframe with means of numeric features grouped by cylinder count
# vector of variables to summarize
num_features <- c("cyl", "disp", "hp", "drat", "wt", "qsec", "gear", "carb")

# group data by cyl and summarize selected features, means
mns <- mtcars[mtcars$cyl != 6,]%>%
  group_by(cyl)%>%
  summarize(disp=mean(disp), hp=mean(hp),
            drat=mean(drat), wt=mean(wt),
            qsec=mean(qsec), gear=mean(gear),
            carb=mean(carb))

# add column with names for each row
mns <- cbind(c("Means 4cyl", "Means 8cyl"), mns)
colnames(mns)[1] <- " "

# vector of outliers from residuals plot, subset on outliers
outlier_names <- c("Chrysler Imperial","Fiat 128", "Toyota Corolla")
outliers <- mtcars[outlier_names, num_features]
outliers <- rownames_to_column(outliers, var=" ")

# bind means dataframe and outliers
outliers_sumry <- rbind(mns, outliers)
kable(outliers_sumry, digits=3,
      caption="Comparison of Outliers & Grouped Means")

# create linear model with weight and acceleration added to predictors
fit3 <- lm(mpg ~ factor(am) + wt + qsec, data=mtcars)

# analysis of variance of second and third models
anvar <- anova(fit2, fit3)

# test change in hat values for each observation if removed from data
hvs_fit3 <- hatvalues(fit3)

# test influence of each observation
```

```r
dff_fit3 <- dffits(fit3)
dfbts_fit3 <- dfbetas(fit3)

# recreate model 3 using revised dataset
fit4 <- lm(mpg~factor(am) + wt + qsec,
           data=mtcars[rownames(mtcars) != "Chrysler Imperial",])

# output pretty table of coefficients
kable(summary(fit4)$coef, digits = 3, caption="Coefficients of Finalized Model")
summary(fit3)
# calculate predictive sum of squares for each model
PRESS <- sapply(list(fit1, fit2, fit3, fit4),
                function(mdl){sum(rstandard(mdl, type="pred")^2)})

# create data frame with PRESS values
mdls <- c("Model 1", "Model 2", "Model 3", "Rmv Outlier")
predictors <- c("am", "am + wt", "am + wt + qsec", "am + wt + qsec")
rdf <- cbind(mdls, predictors, round(PRESS, 3))
colnames(rdf) <- c("Model", "Predictors", "PRESS")
# pairs plot of mtcars features
ggpairs(mtcars)
## density plot of mpg by transmission type
mtcars$am <- relevel(as.factor(mtcars$am), ref=2)
g <- ggplot()+
  geom_density(data=mtcars,aes(x=mpg, color=factor(am))) +
  labs(title="Fuel Efficiency (mpg) by Transmission Type")+
  scale_color_manual(name="Transmission Type",
                     labels=c("Automatic", "Manual"),
                     values=c("#F8766D", "#00BFC4"))+
  theme(panel.background = element_blank(),
        legend.position = c(1,1),
        legend.title = element_text(size=8),
        legend.text = element_text(size = 8),
        legend.justification = c(1,1),
        legend.background = element_rect(color="black", fill="grey85"))

aut_dens <- density(mtcars[mtcars$am==0,]$mpg)
q95 <- quantile(mtcars[mtcars$am==0,]$mpg, probs=0.95)
df_aut_dens <- data.frame(aut_dens$x, aut_dens$y)

g + geom_area(data=subset(df_aut_dens, aut_dens.x >= q95),
              aes(x=aut_dens.x, y=aut_dens.y), fill="#00BFC4", alpha=.25)+
  geom_vline(xintercept = mean(mtcars[mtcars$am==1,]$mpg),
             size=1, color="#F8766D")
# plot of displacement by number of cylinders
par(mfrow=c(1,1), mar=c(4, 4, 3, 2)+0.1)
plot(mtcars$cyl, mtcars$disp, col=factor(mtcars$am), pch=20,
     main="Displacement by Number of Cylinders",
     ylab="Diplacement (cu.in.)", xlab="No. of Cylinders")
legend(4, 450, legend=c("automatic", "manual"), col = c("red", "black"), pch=20)
# plot of weight by displacement colored by number of cylinders
plot(mtcars$disp, mtcars$wt, col=mtcars$cyl, pch=as.integer(mtcars$am),
     main="Displacement by Weight",
```

```
     ylab="Weight (tons)", xlab="Displacement (cu.in.)")
legend(72,5.3, legend=c("Automatic", "Manual"), pch=c(2,1))
legend(385,2.8, legend=c("4cyl", "6cyl", "8cyl"), c("blue", "magenta", "gray"))
# residuals plots
par(mfrow=c(2, 1), mar=c(2, 2, 2, 2))
plot(fit2, which=1)
abline(h=0, lty=2, col="blue")
plot(fit2, which=2)
# scatterplot of covariates
plot(mtcars$qsec, mtcars$mpg, col=mtcars$wt, pch=20,
     main="Acceleration by Miles/Gallon")
abline(lm(mpg~qsec, data=mtcars))
text(mtcars$qsec, mtcars$mpg, rownames(mtcars), cex = 0.5, pos=2)
text(mtcars$qsec, mtcars$mpg, mtcars$wt, cex = 0.5, pos=4)
legend(22.1, 20, legend=c("1", "2", "3", "4", "5"),
       fill=palette()[1:5], title="Int of WT", cex = 0.7)
fit <- lm(mpg~factor(am) + wt + qsec, data=mtcars)
plot(fit, which=1)
##
```