Regression Models Course Project

Executive Summary

- Exploratory Data Analysis show that mpg is distributed normally with no outliers
- hp,disp,wt are negatively correlated with mpg and drat is positively correlated
- Box plot of MPG reveals that MPG is higher for Manual transmission than Automatic trasmission
- Hypothesis testing confirms that there is significant difference in mean mpg between Manual and Automatic transmission
- Mutlivariate model built on am,wt & qsec to predict mpg explains 84% of the variance
- manual transmitted car that weighs 2000 lbs have 5.797 more MPG than an automatic transmitted car with same weight and qsec

Loading required packages

```
library("ggplot2")
library(dplyr)
library(Hmisc)
library(reshape2)
library(gridExtra)
```

Loading mtcars dataset and coverting the required variables to factors

```
data(mtcars)
mtcars <- mutate(mtcars,am = as.factor(am),cyl = as.factor(cyl),gear = as.factor(g
ear),
                 carb = as.factor(carb),vs = as.factor(vs))
```

Lets check if mpg is different for automatic and manual transimission cars using a T-test Null hypothesis will be no difference in mean MPG between the two groups p-value is less than 0.05,null hypothesis is rejected and we can confirm that mean mpg for automatic and manual cars are different Estimated mean MPG of manual transmission cars is 7 more than automatic transmission cars

```
ttest_am <- t.test(mpg ~ am,data=mtcars)
ttest_am$p.value
```

```
## [1] 0.001373638
```

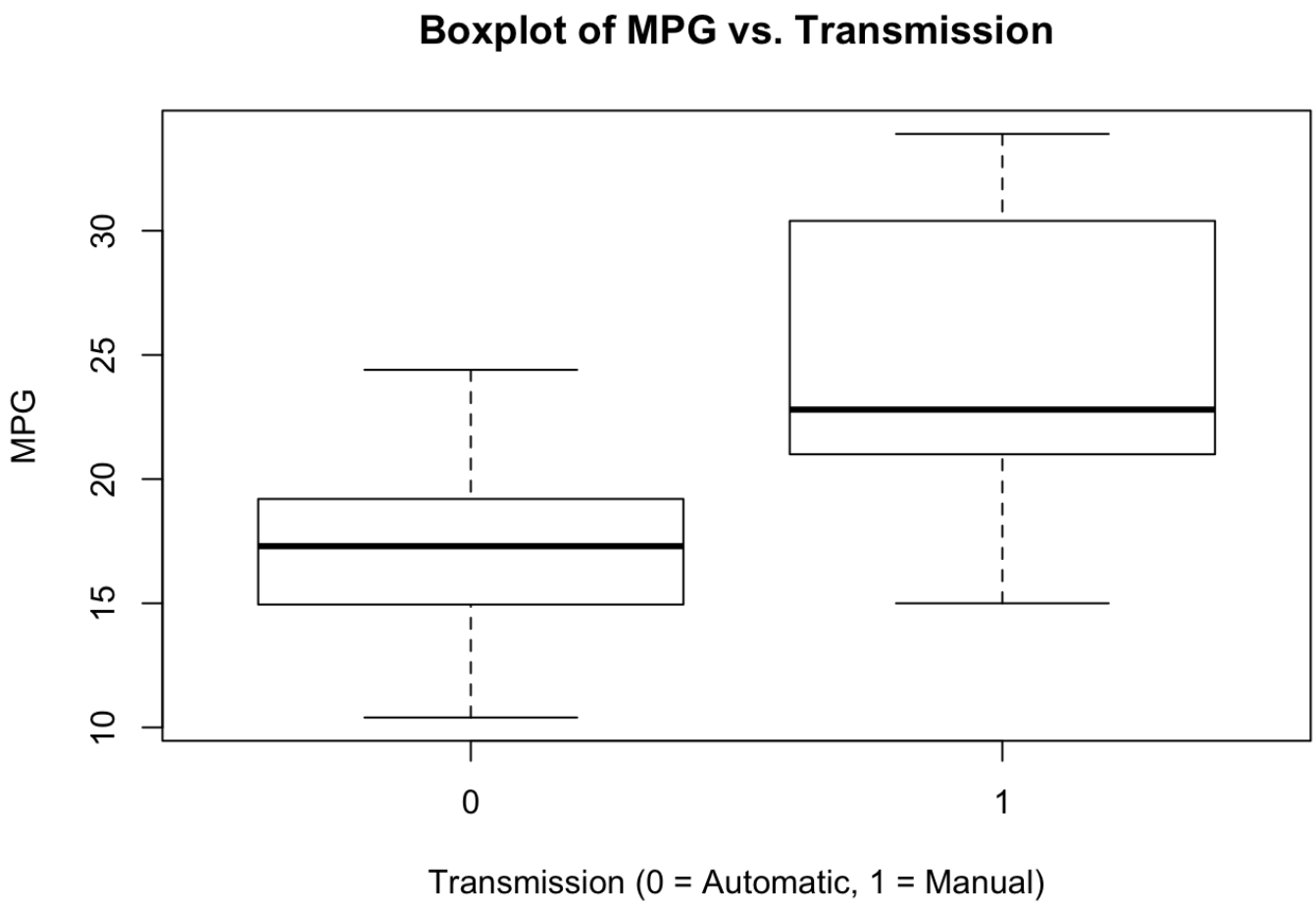
```
ttest_am$estimate
```

```
## mean in group 0 mean in group 1
##        17.14737        24.39231
```

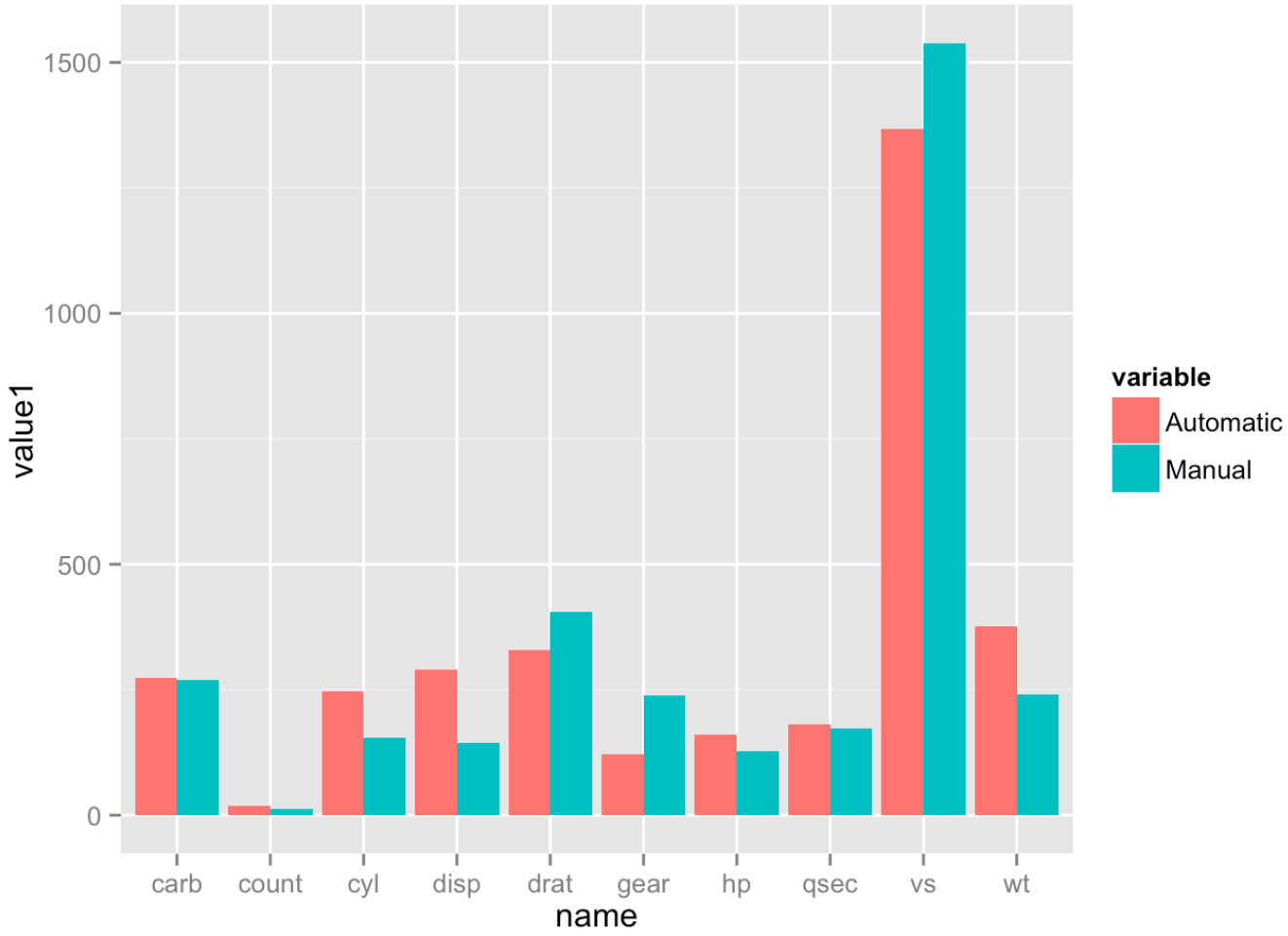
To quantify the MPG difference we need to, - explore the data to understand if there are other variables responsible for variation in mpg - build regression models to estimate MPG difference based on am

Data exploration - Box plot reveals that manual transmission cars have high MPG - Automatic transmission cars tend to have high mean weight, high mean hp, low mean gear, high number of cylinders - There is relationship b/w mpg and disp,hp,wt,drat - Variation in MPG can not be explained only by transmission type

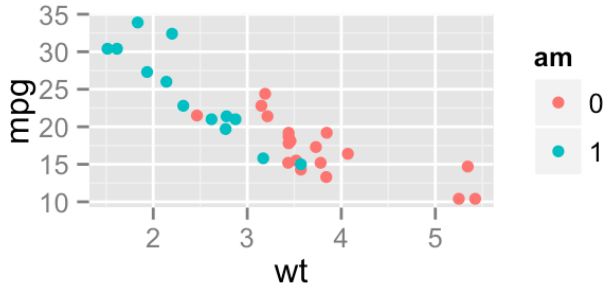
```
am_summary <- mtcars %>%
  group_by(am) %>%
  summarise(count = n(),cyl=mean(cyl)*100,disp=mean(disp),hp=mean(hp),drat=mean(dr
at)*100,wt=mean(wt)*100,
            qsec=mean(qsec)*10,vs=mean(vs)*1000,gear=mean(gear)*100,carb=mean(carb
)*100) # scaling by a factor of 10/100 or 1000 to have all variables in the same r
ange
t <- as.data.frame(t(am_summary))
t <- t[-1, ]
colnames(t) <- c("Automatic","Manual")
t$name <- rownames(t)
mdata <- melt(t,id=c("name"))
mdata$value1 <- as.numeric(as.character(mdata$value))

boxplot(mtcars$mpg ~ mtcars$am, xlab="Transmission (0 = Automatic, 1 = Manual)", y
lab="MPG",
        main="Boxplot of MPG vs. Transmission")
```

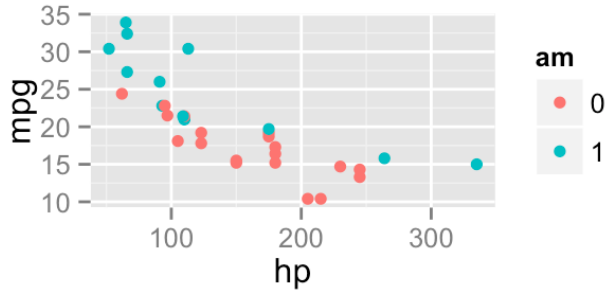


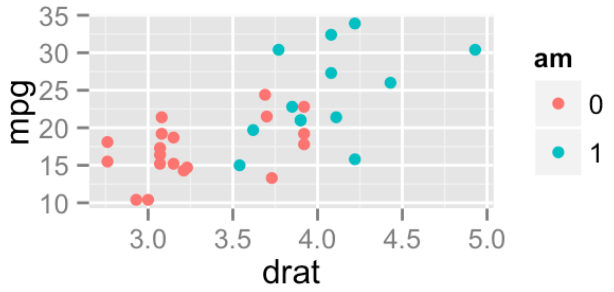
```
ggplot(mdata,aes(name,value1,fill=variable))+
  geom_bar(position="dodge",stat="identity")
```

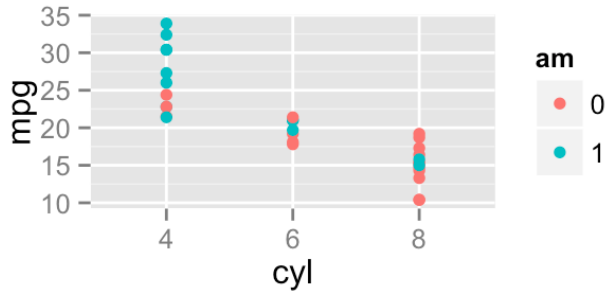


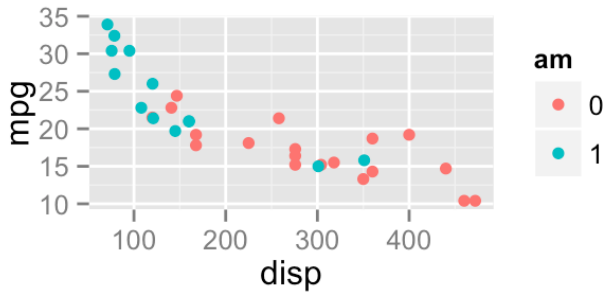

```
plot1 <- qplot(wt,mpg,data = mtcars,color = am)
plot2 <- qplot(hp,mpg,data = mtcars,color = am)
plot3 <- qplot(drat,mpg,data = mtcars,color = am)
plot4 <- qplot(cyl,mpg,data = mtcars,color = am)
plot5 <- qplot(disp,mpg,data = mtcars,color = am)
grid.arrange(plot1,plot2,plot3,plot4,plot5,ncol=2)
```












```
mtcars_corr_matrix <- rcorr(as.matrix(mtcars))
corr_mpg <- as.data.frame(mtcars_corr_matrix$r)
corr_mpg$name <- rownames(corr_mpg)
corr_mpg[corr_mpg$name == "mpg",-12]
```

```
##     mpg        cyl       disp         hp      drat         wt     qsec
## mpg   1 -0.8521619 -0.8475513 -0.7761683 0.6811719 -0.8676594 0.418684
##            vs        am      gear       carb
## mpg 0.6640389 0.5998324 0.4802848 -0.5509251
```

Regression Models

- model based on am explains only 33.8% of variantion so more varibles need to be included to quantify mpg variation

```
fit_am <- lm(mpg ~ am,data=mtcars)
summary(fit_am)$adj.r.squared
```

```
## [1] 0.3384589
```

- model based on all varibles explains 78% of variation but none of the variables are statistically

significant in explaining the relationship

```
fit_allvars <- lm(mpg ~ .,data=mtcars)
summary(fit_allvars)$adj.r.squared
```

```
## [1] 0.7790215
```

- using stepwise backward selection we end up on mpg ~ wt + qsec + am and this model explains 83% of variance

```
stepModel <- step(fit_allvars, k=log(nrow(mtcars)))
```

```
stepModel$call
```

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

```
summary(stepModel)$adj.r.squared
```

```
## [1] 0.8335561
```

- But there is an interaction between am and hp/wt/disp so building a model with interaction b/w wt & am

```
fit_am_wt_interaction <- lm(mpg ~ wt + qsec + am + wt:am ,data=mtcars )
summary(fit_am_wt_interaction)$adj.r.squared
```

```
## [1] 0.8804219
```

- Model selection

```
anova(fit_am, fit_allvars, stepModel, fit_am_wt_interaction)
confint(fit_am_wt_interaction) # results hidde
```

- Model with higher explainaibilty based on r-square is selected

```
summary(fit_am_wt_interaction)$coef
```

```
##              Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)  9.723053  5.8990407  1.648243 0.1108925394
## wt          -2.936531  0.6660253 -4.409038 0.0001488947
## qsec         1.016974  0.2520152  4.035366 0.0004030165
## am1         14.079428  3.4352512  4.098515 0.0003408693
## wt:am1      -4.141376  1.1968119 -3.460340 0.0018085763
```

Thus, the result shows that when “wt” (weight lb/1000) and “qsec” (1/4 mile time) remain constant, cars with manual transmission add 14.079 + (-4.141)*wt more MPG (miles per gallon) on average than cars with automatic transmission. That is, a manual transmitted car that weighs 2000 lbs have 5.797 more MPG than an automatic transmitted car that has both the same weight and 1/4 mile time.

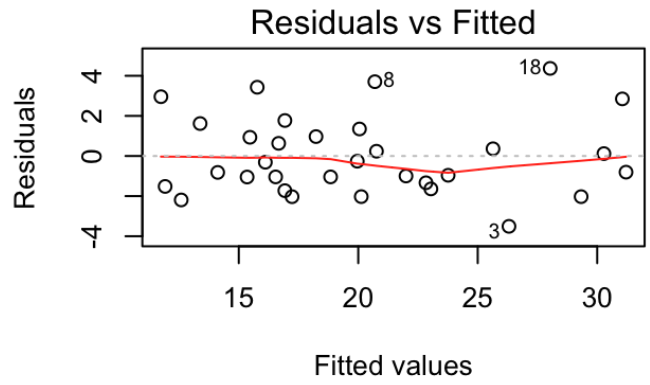
Residual Analysis and Diagnostics Please refer to the Appendix: Figures section for the plots. According to the residual plots, we can verify the following underlying assumptions: - The Residuals vs. Fitted plot shows no consistent pattern, supporting the accuracy of the independence assumption. - The Normal Q-Q plot indicates that the residuals are normally distributed because the points lie closely to the line. - The Scale-Location plot confirms the constant variance assumption, as the points are randomly distributed. - The Residuals vs. Leverage argues that no outliers are present, as all values fall well within the 0.5 bands - As for the Dfbetas, the measure of how much an observation has effected the estimate of a regression coefficient, we get the following result:

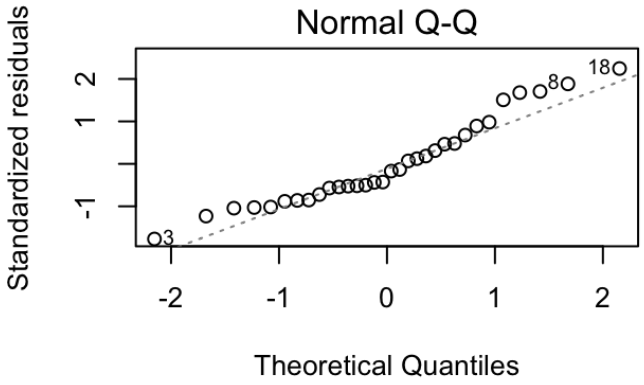
```
sum((abs(dfbetas(fit_am_wt_interaction)))>1)
```

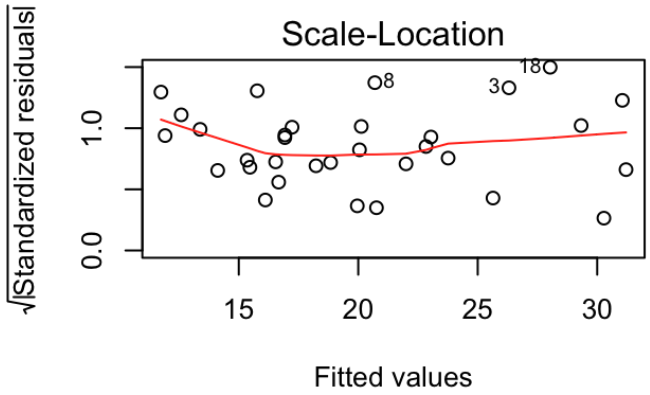
```
## [1] 0
```

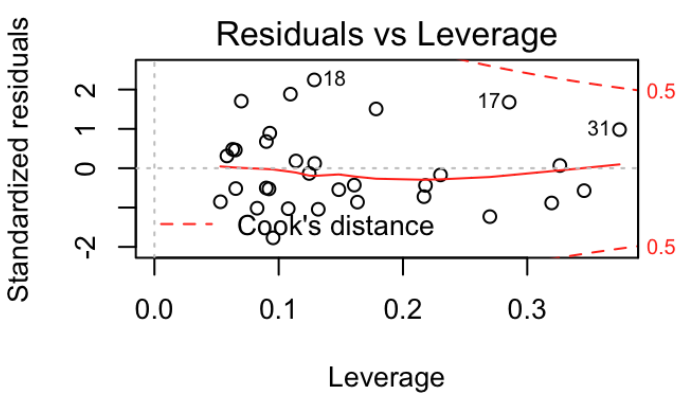
- Residual Plots

```
par(mfrow = c(2, 2))
plot(fit_am_wt_interaction)
```


Residuals vs Fitted
Residuals
Fitted values
18
8
3


Normal Q-Q
Standardized residuals
Theoretical Quantiles
18
8
3


Scale-Location
√|Standardized residuals|
Fitted values
18
8
3


Residuals vs Leverage
Standardized residuals
Leverage
Cook's distance
18
17
31
0.5
0.5