MT Cars Analysis

Jayesh Gokhale

5/1/2021

Analysis on Motor Trend Cars Dataset

We have to answer two questions -

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

I looked up the help file of mtcars data set and based on data, I gathered that following are the numeric variables

mpg (Miles/(US) gallon), disp (Displacement), hp (Gross Horsepower), drat (Gear Axle Ratio), wt (Weight in 1000 lbs) and qsec (1/4 mile time).

Following are the categorical variables -

cyl (Number of cylinders), vs (Engine: 0 = v-shaped, 1 = straight), **am (Transmission: 0 = automatic, 1 = manual)**, gear (Number of forward gears), carb (Number of Carburetors)

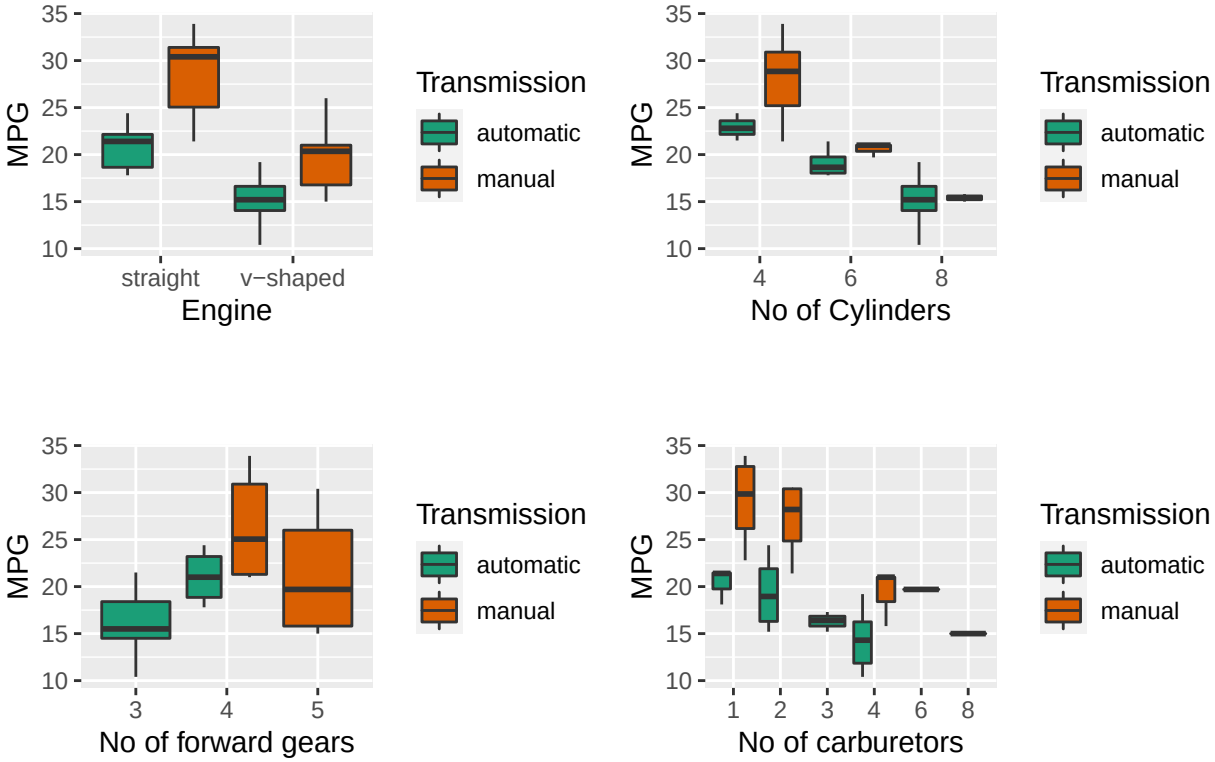
Let us begin with a box plot of mpg vs all factor variables across am (Transmission)

```
getBoxPlot <- function(df,xvar,yvar,fillvar,xlabel,ylabel,filllabel)
{
  mybox <- ggplot(data=df, aes(x=xvar,y=yvar,fill=fillvar)) +
  geom_boxplot(outlier.colour="black",
               outlier.size=2,position=position_dodge(1)) +
  labs(title = "",
       subtitle = "",
       y = ylabel, x = xlabel) +
    scale_fill_brewer(name = filllabel,palette="Dark2")  +
  theme(plot.title = element_text(hjust = 0.5))

  return(mybox)
}

plotVS <- getBoxPlot(mt,mt$vs,mt$mpg,mt$am,"Engine","MPG","Transmission")
plotCYL <- getBoxPlot(mt,mt$cyl,mt$mpg,mt$am,"No of Cylinders","MPG","Transmission")
plotGEAR <- getBoxPlot(mt,mt$gear,mt$mpg,mt$am,"No of forward gears","MPG","Transmission")
plotCARB <- getBoxPlot(mt,mt$carb,mt$mpg,mt$am,"No of carburetors","MPG","Transmission")

figure.box <- ggarrange(plotVS, plotCYL, plotGEAR, plotCARB, ncol = 2, nrow = 2)
figure.box
```



BoxPlot Observations

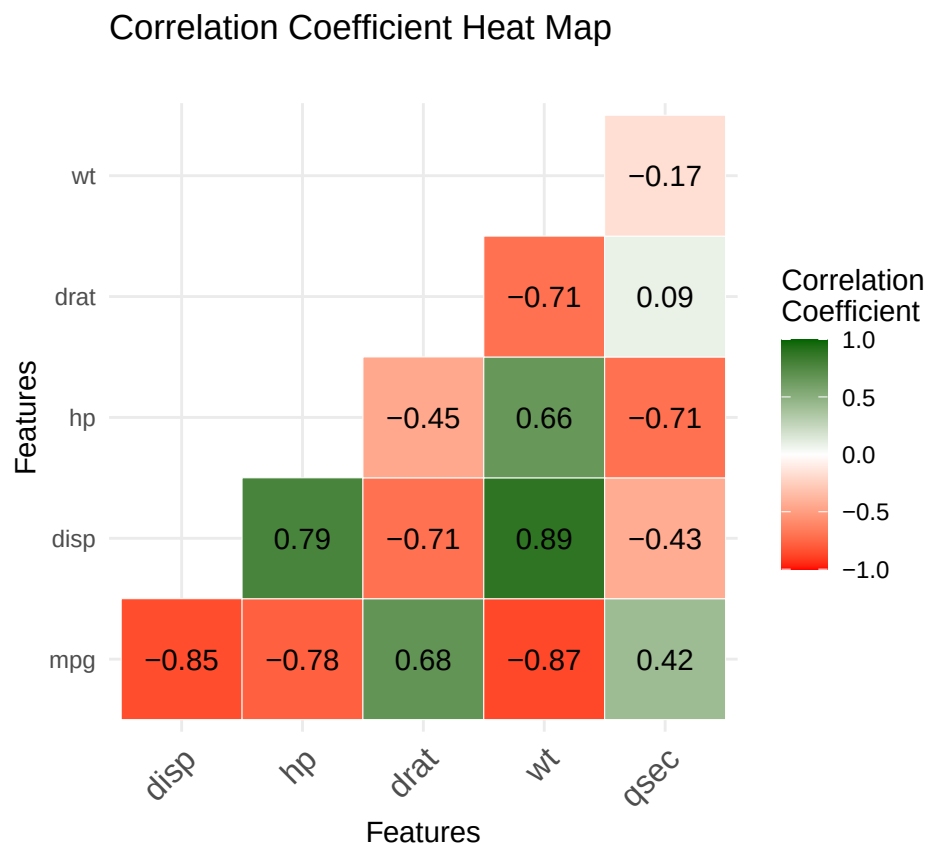
1. Manual transmission does appear to have significantly higher MPG across Engines, Carburetors and No of forward gears wherever applicable
2. For 4 cylinder engines, manual transmission seems to give higher MPG
3. In 6 & 8 cylinder engines, it seems to be too close to call.

The results may be misleading unless we take into account the effect of other numeric variables.

Let us first draw a heatmap of the continuous variables.

```
correlation.matrix <- cor(mt[,c("mpg","disp","hp","drat","wt","qsec")])
correlation.matrix <- round(correlation.matrix,2)
correlation.matrix[upper.tri(correlation.matrix)] <- NA
diag(correlation.matrix) <- NA
row.corr.matrix <- melt(correlation.matrix,na.rm=TRUE)
colnames(row.corr.matrix) <- c("F1","F2","CORR")
ggplot(data = row.corr.matrix, aes(x=F1, y=F2, fill=CORR)) +
  geom_tile(color="white")+
  labs(title = "Correlation Coefficient Heat Map",
       subtitle = "",
       y = "Features", x = "Features") +
  scale_fill_gradient2(low = "red", high = "darkgreen", mid = "white",
   midpoint = 0, limit = c(-1,1), space = "Lab",
   name="Correlation\nCoefficient") +
```

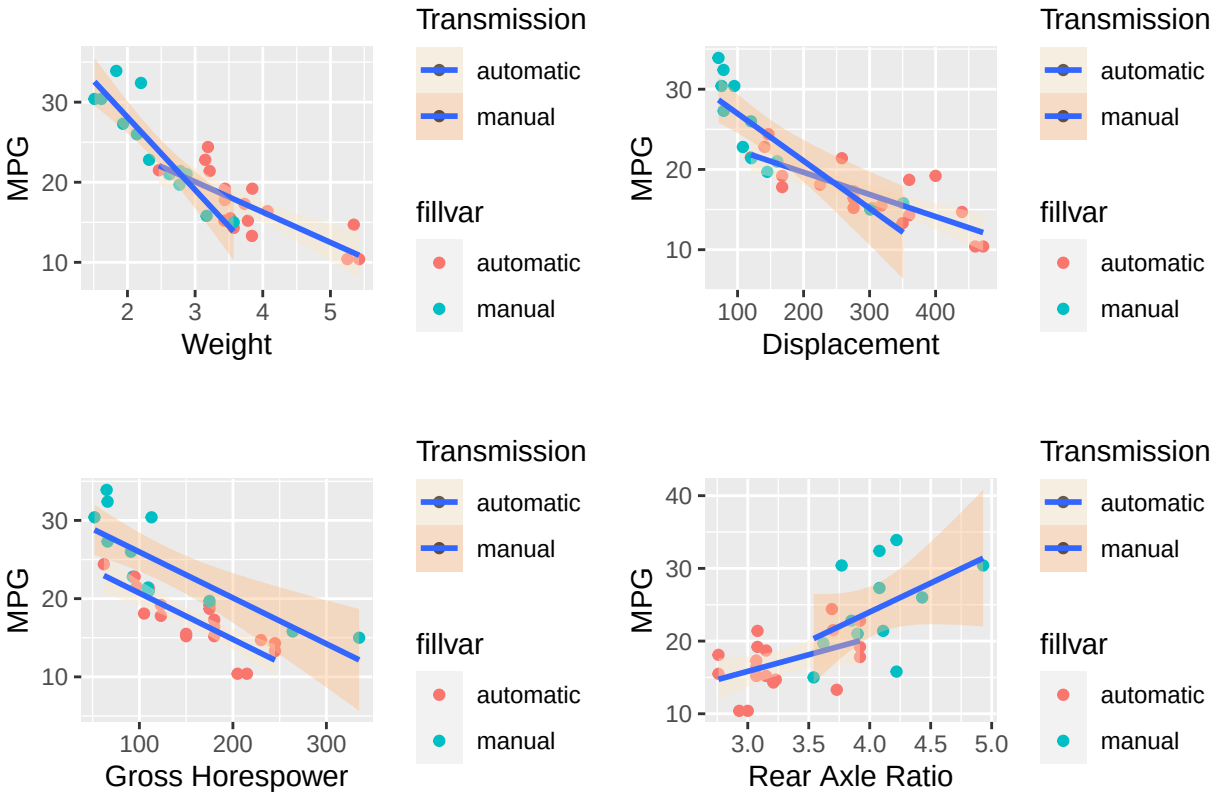
```
 theme_minimal()+
theme(axis.text.x = element_text(angle = 45, vjust = 1,
   size = 12, hjust = 1))+
coord_fixed()+
 geom_text(aes(x=F1, y=F2, label = CORR), color = "black", size = 4)
```



Heatmap Observations

1. We can see that wt, disp and hp are negatively highly correlated to mpg
2. drat is positively correlated to mpg.
3. We can also observe that wt, disp, hp and drat are correlated amongst themselves (either + or -).

Let us now look at the relationship between am (Transmission) and mpg while adjusting for each of the four numeric variables viz. wt, disp, hp and drat.



Adjustment Plot Observations

1. Except horsepower, Weight, Displacement and Gear Axle Ratio explains for a lot of variation in MPG. So does Displacement and Gear Axle Ratio.
2. We need to do here is a residual analysis. We need to study the how much of the residual variation in MPG is explained by Transmission after removing the effect of weight, hp, disp and drat one by one.

Before residual analysis let us do a quick ANOVA to see whether we should add no of cylinders in our model or not since that does not strictly follow the fixed pattern.

```
fit1 <- lm(mpg~wt,data=mt)
fit2 <- update(fit1,mpg~wt+cyl)
fit3 <- update(fit1,mpg~wt*cyl)
anova(fit1,fit2,fit3)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ wt
## Model 2: mpg ~ wt + cyl
## Model 3: mpg ~ wt + cyl + wt:cyl
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1     30 278.32
## 2     28 183.06  2    95.263 7.9443 0.00203 **
## 3     26 155.89  2    27.170 2.2658 0.12386
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Anova Observation

Thus we can see that cylinder is significant here and we will add it to our residual analysis for weight. Let us add it for others too (but not its interaction as is evident from figures above).

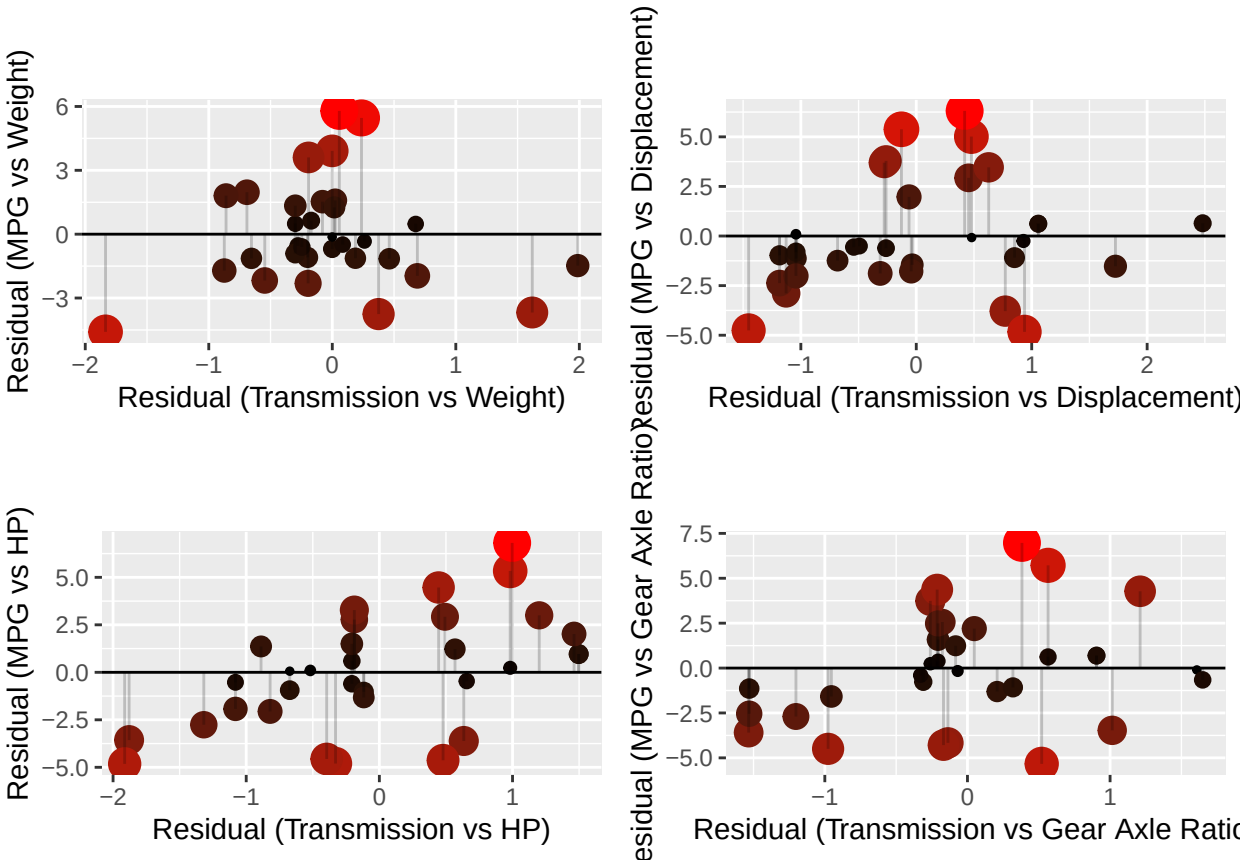
```
getResidPlot <- function(df,xvar,yvar,xlab,ylab)
{
  residPlot <- ggplot(data=df, aes(x=xvar,y=yvar)) +
  geom_point(aes(color=abs(yvar),size=abs(yvar))) +
  labs(title = "",
       subtitle = "",
       y = ylab, x = xlab) +
  theme(plot.title = element_text(hjust = 0.5)) +
  guides(color = FALSE, size = FALSE) +
  scale_color_continuous(low = "black", high = "red") +
  geom_segment(aes(xend = xvar, yend = 0, color=), alpha = .2) +
  geom_hline(yintercept=0)

  return(residPlot)
}
mt$resid.mpg.wt <- resid(glm(mpg~wt+cyl,data=mt))
mt$resid.mpg.disp <- resid(glm(mpg~disp+cyl,data=mt))
mt$resid.mpg.hp <- resid(glm(mpg~hp+cyl,data=mt))
mt$resid.mpg.drat <- resid(glm(mpg~drat+cyl,data=mt))

mt$resid.am.wt <- resid(glm(am~wt+cyl,data=mt,family="binomial"))
mt$resid.am.disp <- resid(glm(am~disp+cyl,data=mt,family="binomial"))
mt$resid.am.hp <- resid(glm(am~hp+cyl,data=mt,family="binomial"))
mt$resid.am.drat <- resid(glm(am~drat+cyl,data=mt,family="binomial"))

plotResidWeight <- getResidPlot(mt,mt$resid.am.wt,mt$resid.mpg.wt,"Residual (Transmission vs Weight)","
plotResidDisp <- getResidPlot(mt,mt$resid.am.disp,mt$resid.mpg.disp,"Residual (Transmission vs Displacer
plotResidHP <- getResidPlot(mt,mt$resid.am.hp,mt$resid.mpg.hp,"Residual (Transmission vs HP)","Residual
plotResidDRAT <- getResidPlot(mt,mt$resid.am.drat,mt$resid.mpg.drat,"Residual (Transmission vs Gear Axl

figure.residual <- ggarrange(plotResidWeight, plotResidDisp, plotResidHP, plotResidDRAT, ncol = 2, nrow
figure.residual
```



Residual Plot Observations

All the residual plots are homoscedastic. That is I cannot make out any pattern from these plots. This means that if we take example of weight (first plot), ***If the effect of weight and number of cylinders is removed, merely the fact that an engine is automatic or manual does not seem to have any impact on the MPG. The same holds true for effect of Displacement, HP and Gear Axle Ratio combined with No. of Cylinders***

Question 2 is "Quantify the MPG difference between automatic and manual transmissions"

If we do not account for any other factors then it is the difference of means + or - the pooled variance. Let us calculate that.

```
x1 <- mt[mt$am=="automatic",]$mpg; x2 <- mt[mt$am=="manual",]$mpg
s1 <- sd(x1); s2 = sd(x2); n1 <- length(x1); n2 <- length(x2);
s <- sqrt(((n1-1)*(s1**2) + (n2-1)*(s2**2)) / (n1+n2-2))
mu1 <- mean(x1); mu2 <- mean(x2)
diffMU <- mu2 - mu1
t.statistic <- diffMU / (s * sqrt((1/n1)+(1/n2)))
p.value <- pt(t.statistic,n1+n2-2,lower.tail = FALSE)
```

So if we do not account for any other factors, then mean of manual MPG is higher by 7.2449393 with pooled variance of 4.9020288 and with a p value of 1.4251037×10^{-4} it is statistically higher.

However let us take out the effect of weight and number of cylinders.

```
x1 <- mt[mt$am=="automatic",]$resid.mpg.wt; x2 <- mt[mt$am=="manual",]$resid.mpg.wt
s1 <- sd(x1); s2 = sd(x2); n1 <- length(x1); n2 <- length(x2);
s <- sqrt(((n1-1)*(s1**2) + (n2-1)*(s2**2)) / (n1+n2-2))
mu1 <- mean(x1); mu2 <- mean(x2)
diffMU <- mu2 - mu1
t.statistic <- diffMU / (s * sqrt((1/n1)+(1/n2)))
p.value <- pt(t.statistic,n1+n2-2,lower.tail = FALSE)
```

So if we account for weight and number of cylinders, then mean of manual MPG is higher by 0.0779505 with pooled variance of 2.4698971 and with a p value of 0.4653558 it is statistically insignificant.