Car Transmission on Fuel Efficiency

Xingmin Aaron Zhang

Our goal for this report is to explore the effect of car transmission, manual or automatic, on fuel efficiency measurement by miles per gallon. We will use the mtcars dataset in R for this project.

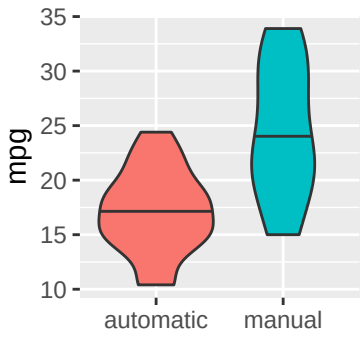
Load Data and Exploratory analysis

The dataset has 11 variables and 32 observations. By default, all variables are considered numeric. We correct some into factors, including cyl, vs, am, gear and carb.

```
attach(mtcars)
mtcars$cyl = as.factor(mtcars$cyl)
mtcars$vs = as.factor(mtcars$vs)
mtcars$am = as.factor(mtcars$am)
mtcars$gear = as.factor(mtcars$gear)
mtcars$carb = as.factor(mtcars$carb)
```

We plot mpg against transmission to develop an initial impression.

```
library(ggplot2)
ggplot(mtcars) + geom_violin(aes(x = factor(am), y = mpg, fill = factor(am)), draw_quantiles = c(0.5)) +
    scale_fill_discrete(breaks = c(0, 1), labels = c("automatic", "manual")) +
    xlab("") + scale_x_discrete(breaks = c(0, 1), labels = c("automatic", "manual")) +
    theme(legend.position = "none")
```

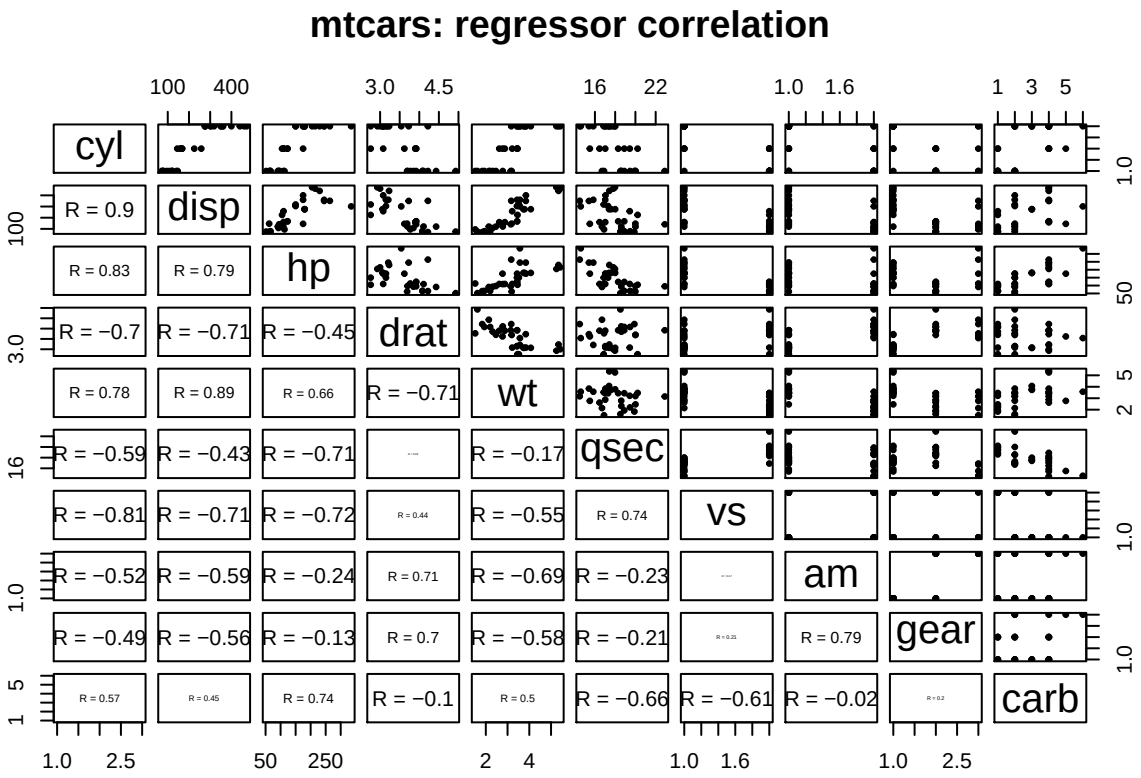


We create pairwise plot to look at the relationship between each variables.

```
# plot inspired by http://www.sthda.com/english/wiki/scatter-plot-matrices-r-base-graphs
panel.cor <- function(x, y){
    usr <- par("usr"); on.exit(par(usr))
    par(usr = c(0, 1, 0, 1))
    r <- round(cor(as.numeric(x), as.numeric(y)), digits=2)
    txt <- paste0("R = ", r)
    cex.cor <- 0.8/strwidth(txt)
    text(0.5, 0.5, txt, cex = cex.cor * r)
```

```
}

upper.panel<-function(x, y){
  points(x,y, pch = 19, cex = 0.5)
}
pairs(mtcars[,-1], gap = 0.5, lower.panel = panel.cor, upper.panel = upper.panel,
      main ="mtcars: regressor correlation")
```



The plot shows there are some variables, such as cyl ~ disp, cyl ~ hp, cyl ~ wt, disp ~ hp, disp ~ wt and drat ~ wt, that show high correlation. We should pay attention to those variables as all might not be required in our linear prediction model.

Statistical Analysis

We hope to find the relationship between mpg and transmission. Our first model is simply a linear relationship between mpg and all variables. We perform a variance inflation factor analysis to check for linearly related variables.

```
fit <- lm(mpg ~ ., data = mtcars)
library(car)
```

```
## Loading required package: carData
```

```
sqrt(vif(fit))
```

```
##           GVIF       Df GVIF^(1/(2*Df))
## cyl  11.319053 1.414214        1.834225
## disp  7.769536 1.000000        2.787389
## hp    5.312210 1.000000        2.304823
## drat  2.609533 1.000000        1.615405
## wt    4.881683 1.000000        2.209453
## qsec  3.284842 1.000000        1.812413
## vs    2.843970 1.000000        1.686407
## am    3.151269 1.000000        1.775181
## gear  7.131081 1.414214        1.634138
## carb 22.432384 2.236068        1.364858
```

As expected from the exploratory analysis, disp, hp and wt have high VIF due to their correlation. Therefore, we remove them one by one and check the VIF of remain variables. We found that after removing both hp and disp, the VIFs look better.

```
variables <- colnames(mtcars)
fit2 <- lm(mpg ~ ., data = mtcars[,variables != "hp"])
sqrt(vif(fit2))
```

```
##           GVIF       Df GVIF^(1/(2*Df))
## cyl   9.187241 1.414214        1.740990
## disp  7.563500 1.000000        2.750182
## drat  2.608198 1.000000        1.614992
## wt    4.845050 1.000000        2.201147
## qsec  3.269540 1.000000        1.808187
## vs    2.627953 1.000000        1.621096
## am    3.150134 1.000000        1.774862
## gear  6.314643 1.414214        1.585211
## carb 13.323707 2.236068        1.295575
```

```
fit3 <- lm(mpg ~ ., data = mtcars[,variables != "hp" & variables != "disp"])
sqrt(vif(fit3))
```

```
##           GVIF       Df GVIF^(1/(2*Df))
## cyl  5.664030 1.414214        1.542700
## drat 2.607671 1.000000        1.614829
## wt   3.154976 1.000000        1.776225
## qsec 3.247714 1.000000        1.802142
## vs   2.611426 1.000000        1.615991
## am   2.961625 1.000000        1.720937
## gear 5.057345 1.414214        1.499618
## carb 7.120020 2.236068        1.216881
```

In the next step, we build nested models with the remaining variables and check whether adding each variables improves prediction.

```r
mtcars_subset <- mtcars[,variables != "hp" & variables != "disp"]
model1 <- lm(mpg ~ am, data = mtcars_subset)
model2 <- update(model1, mpg ~ am + cyl)
model3 <- update(model2, mpg ~ am + cyl + drat)
model4 <- update(model3, mpg ~ am + cyl + drat + wt)
model5 <- update(model4, mpg ~ am + cyl + drat + wt + qsec)
model6 <- update(model5, mpg ~ am + cyl + drat + wt + qsec + vs)
model7 <- update(model6, mpg ~ am + cyl + drat + wt + qsec + vs + gear)
model8 <- update(model7, mpg ~ am + cyl + drat + wt + qsec + vs + gear + carb)
anova(model1, model2, model3, model4, model5, model6, model7, model8)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ am + cyl
## Model 3: mpg ~ am + cyl + drat
## Model 4: mpg ~ am + cyl + drat + wt
## Model 5: mpg ~ am + cyl + drat + wt + qsec
## Model 6: mpg ~ am + cyl + drat + wt + qsec + vs
## Model 7: mpg ~ am + cyl + drat + wt + qsec + vs + gear
## Model 8: mpg ~ am + cyl + drat + wt + qsec + vs + gear + carb
##   Res.Df    RSS Df Sum of Sq       F    Pr(>F)
## 1     30 720.90
## 2     28 264.50  2    456.40 25.8134 7.057e-06 ***
## 3     27 264.32  1      0.17  0.0195  0.890559
## 4     26 182.75  1     81.57  9.2274  0.007433 **
## 5     25 159.14  1     23.61  2.6709  0.120580
## 6     24 159.14  1      0.00  0.0000  0.995586
## 7     22 158.86  2      0.27  0.0155  0.984625
## 8     17 150.29  5      8.58  0.1940  0.960629
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

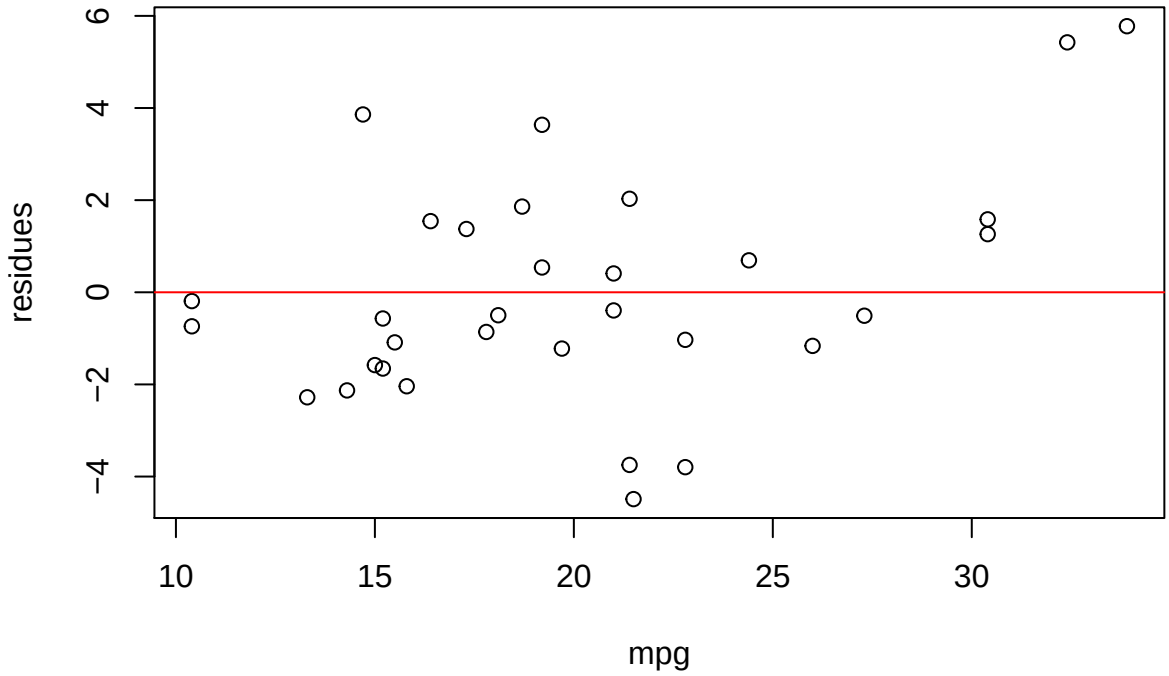
So it appears we should include am, cyl and wt in our final model. Add interactions among them did not improve residue reduction, so we choose the model without interactions.

```r
model <- lm(mpg ~ am + cyl + wt, data = mtcars_subset)
model_int <- update(model, mpg ~ am * cyl * wt )
anova(model, model_int)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am + cyl + wt
## Model 2: mpg ~ am + cyl + wt + am:cyl + am:wt + cyl:wt + am:cyl:wt
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     27 182.97
## 2     20 116.91  7    66.059 1.6144 0.1886
```

We check residules for this model. They look randomly spreaded, indicating that our linear model is fine.

```r
residues <- resid(model)
plot(residues ~ mtcars_subset$mpg, xlab = "mpg")
abline(h = 0, col = "red")
```



We check the coefficients of our final model. The transmission type variable, am, has a coefficient of 0.15, but its variance is so large that it did not pass the t test (p value = 0.91). Instead, increasing the number of cylinders or weight, significantly reduce fuel efficiency. Therefore, we can conclude with high confidence (99%) that the type of transmission did not significantly affect fuel efficiency.

```r
summary(model)$coefficients
```

```
##               Estimate Std. Error    t value     Pr(>|t|)
## (Intercept) 33.7535920  2.8134831 11.9970836 2.495549e-12
## am1          0.1501031  1.3002231  0.1154441 9.089474e-01
## cyl6        -4.2573185  1.4112394 -3.0167231 5.514697e-03
## cyl8        -6.0791189  1.6837131 -3.6105432 1.227964e-03
## wt          -3.1495978  0.9080495 -3.4685309 1.770987e-03
```

Conclusion

We conclude that fuel efficiency is not significantly impacted by the type of transmissions.

Did the student interpret the coefficients correctly? Did the student do some exploratory data analyses? Did the student fit multiple models and detail their strategy for model selection? Did the student answer the questions of interest or detail why the question(s) is (are) not answerable? Did the student do a residual

plot and some diagnostics? Did the student quantify the uncertainty in their conclusions and/or perform an inference correctly? Was the report brief (about 2 pages long) for the main body of the report and no longer than 5 with supporting appendix of figures? Did the report include an executive summary?