Miles per gallon in automatic vs manual cars

Gloria Q

Summary

The principle objective of this assignment is to explore the relationship between a set of variables and miles per gallon (mpg) using the mtcars data set for a fictional assignment for *Motor Trend* magazine. The principle questions to answer are: 1) determine if automatic or manual transmission is better for mpg and 2) quantify the difference between manual and automatic transmission. Initial analysis determined that manual car tranmission has a better mpg compared to automatic cars. Further multivariable analysis and anova testing demonstrated that cylinder size and car weight further impacted mpg.

Exploratory Analysis

Basic inspection of mtcars data frame can be found in the appendix

1. Examine how variables interact with each other:

```
#upload necessary packages and data, conduct t-test
library(datasets); data(mtcars); library(ggplot2); library(GGally); library(gridExtra);
library(dplyr); library(car)
#change cyl and am to factors
mtcars <- mtcars %>% mutate(am=as.factor(if_else(mtcars$am==0, "auto", "manual")),
cyl=as.factor(mtcars$cyl))
#calc standard deviation of variance influence factor (VIF) of each variable against mpg
fit=lm(mpg ~ ., data=mtcars); sqrt(vif(fit))[,1]
```

```
##      cyl     disp       hp     drat       wt     qsec       vs       am
## 6.028697 4.650961 4.009813 1.933020 3.896435 2.782022 2.470153 2.157224
##     gear     carb
## 2.317650 3.282641
```

These results indicate that cyl, disp, hp and wt coefficients vary the greatest amount of times when correlated to the other variables compared to when they are orthogonal. Next step is therefore to visually display their relationship with mpg and transmission as a factor. (**graph in appendix**)

The graph (appendix) indicates a decrease in mpg with increase in cyl, disp, hp, wt in varying degrees. A regression analysis can further quantify the data behavior.

Variable Ananlysis

1. Determine difference in mpg between auto and manual cars

```
fit1=summary(lm(mpg ~ am, mtcars))$coef; fit1
```

```
##              Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 17.147368   1.124603 15.247492 1.133983e-15
## ammanual     7.244939   1.764422  4.106127 2.850207e-04
```

This regression fit calculated that automatic cars run on average 17.1473684 mpg, while manual cars run on average $\beta_o + \beta_1 = 24.3923077$ mpg. For hypothesis testing, where $H_o : \mu_{manual} = \mu_{auto}$ and $H_\alpha : \mu_{auto} \neq \mu_{manual}$, the p-value is 2.8502074×10^{-4}, therefore the null hypothesis can be rejected. Without considering other variables, this analysis states that manual cars have a better mpg.

2. Model selection based on pre-determined significant variables

Based on previously identified principle variables that impact mpg, will now examine nested models to determine which variable contributes best model between mpg and am.

```
#build nesting models and running ANOVA
fit2=lm(mpg ~ am, mtcars); fit3=update(fit2, mpg ~ am + cyl); fit4=update(fit2,
mpg ~ am + cyl + wt); fit5=update(fit2, mpg ~ am + cyl + wt + disp);
fit6=update(fit2, mpg ~ am + cyl + wt + disp + hp); anova(fit2, fit3, fit4, fit5, fit6)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ am + cyl
## Model 3: mpg ~ am + cyl + wt
## Model 4: mpg ~ am + cyl + wt + disp
## Model 5: mpg ~ am + cyl + wt + disp + hp
##   Res.Df    RSS Df Sum of Sq       F    Pr(>F)
## 1     30 720.90
## 2     28 264.50  2    456.40 37.9300 2.678e-08 ***
## 3     27 182.97  1     81.53 13.5510  0.001118 **
## 4     26 182.87  1      0.10  0.0165  0.898954
## 5     25 150.41  1     32.46  5.3954  0.028621 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

For this anova test, each p-value indicates whether the new added variable is significant to the model or not (aka above zero or not). These results indicate the model fit3(am+cyl) and fit4(am+cyl+wt+am) have the greatest impact on mpg.

Conclusion

Linear regression of just mpg and am indicate that mpg between automatic and manual cars are statistically different, with manual cars having a better mpg compared to automatic. When examining the variance of coefficients for all variables, cyl/disp/wt/hp have the greatest influence on mpg. When comparing nested models via ANOVA, models that included am/cyl/wt had the greatest impact on mpg. To check regression assumptions, diagnostic plotting was conducted to compare residuals and predicted values from the ANOVA test (**see appendix**). When examining fit residuals vs predicted values, the density distribution about 0 was slightly better in fit3(mpg~am+cyl), further confirming this to be the optimal regression model.

Appendix

1. mtcars data frame structure:

From ?mtcars, following variable description:

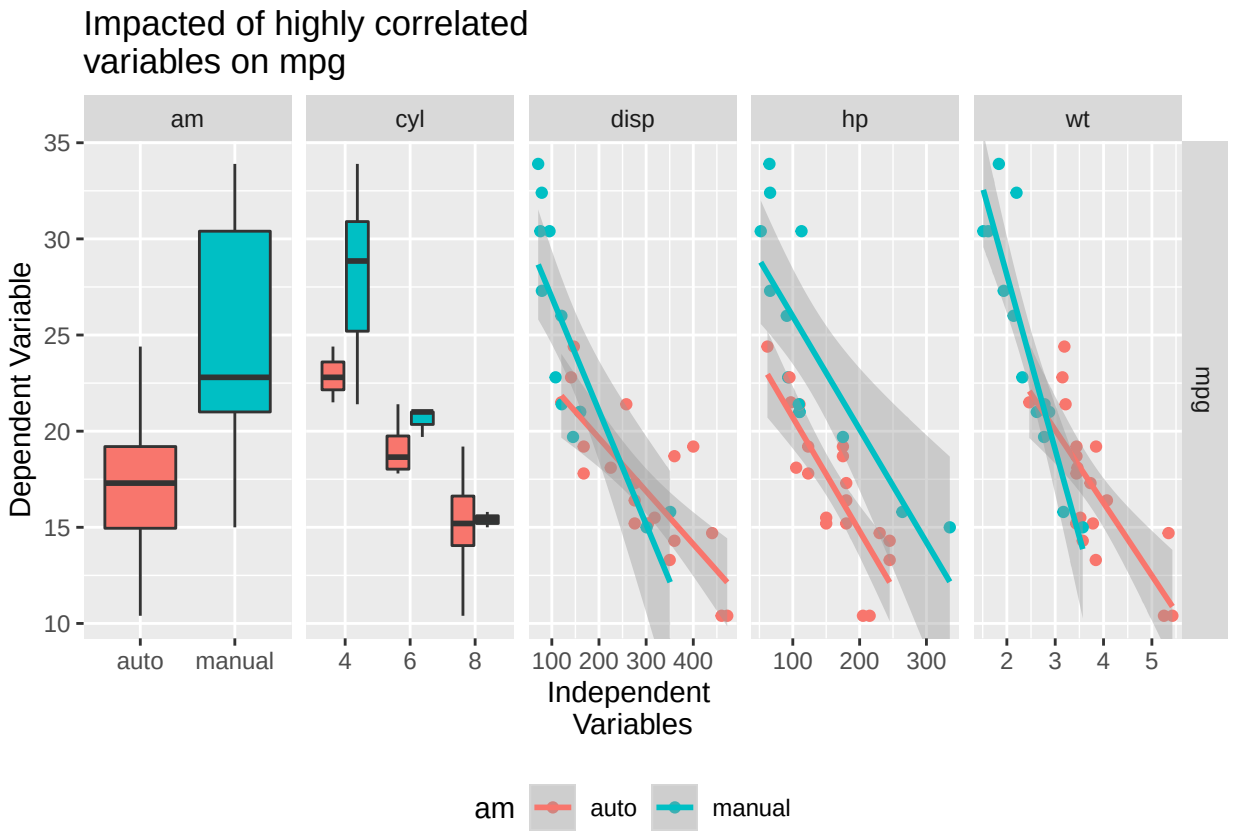
- **mpg**=miles per gallon (US)
- **cyl**=number of cylinders
- **disp**=displacement(cu.in.)
- **hp**=gross horsepower
- **drat**=rear axle ratio
- **wt**=weight(1000 Ilbs)
- **qsec**=$\frac{1}{4}$ mile time
- **vs**=engine(0=v-shaped, 1=straight)
- **am**=transmission(0=auto, 1=manual)
- **gear**=num of forward gears
- **carb**=num of carburetors

```
#look at mtcars data frame
str(mtcars)
```

```
## 'data.frame':    32 obs. of  11 variables:
##  $ mpg : num  21 21 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 ...
##  $ cyl : Factor w/ 3 levels "4","6","8": 2 2 1 2 3 2 3 1 1 2 ...
##  $ disp: num  160 160 108 258 360 ...
##  $ hp  : num  110 110 93 110 175 105 245 62 95 123 ...
##  $ drat: num  3.9 3.9 3.85 3.08 3.15 2.76 3.21 3.69 3.92 3.92 ...
##  $ wt  : num  2.62 2.88 2.32 3.21 3.44 ...
##  $ qsec: num  16.5 17 18.6 19.4 17 ...
##  $ vs  : num  0 0 1 1 0 1 0 1 1 1 ...
##  $ am  : Factor w/ 2 levels "auto","manual": 2 2 2 1 1 1 1 1 1 1 ...
##  $ gear: num  4 4 4 3 3 3 3 4 4 4 ...
##  $ carb: num  4 4 1 1 2 1 4 2 2 4 ...
```

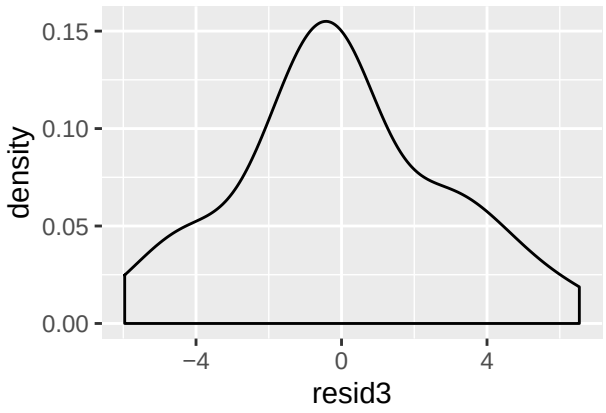
2. Exploratory graph

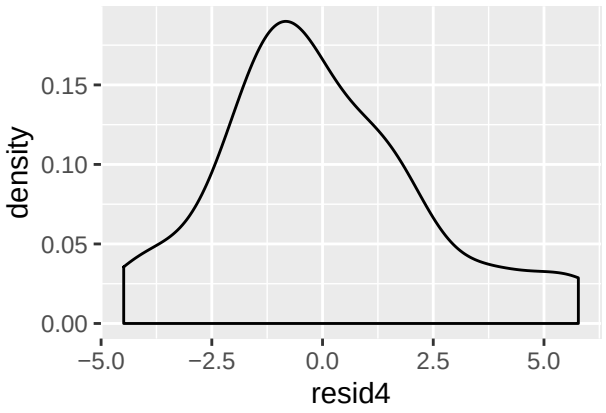
```
#build graphs
ggduo(mtcars, columnsX=c(9,2:4,6), columnsY=1, types = list(continuous =
"smooth_lm"), mapping = aes(color = am), legend = c(1,4), xlab="Independent
Variables", ylab="Dependent Variable", title="Impacted of highly correlated
variables on mpg") + theme(legend.position = "bottom")
```

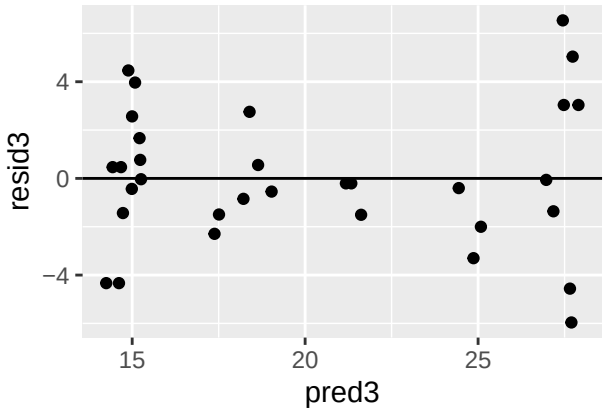


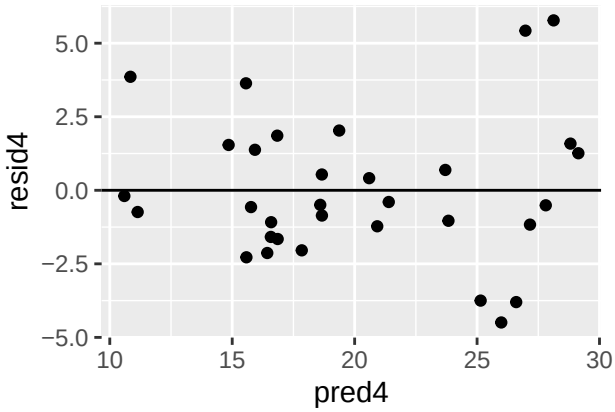
3. diagnostic regression plot:

```
#calc residuals and predicted values
resid3=residuals(fit3);resid4=residuals(fit4);pred3=predict(fit3);pred4=predict(fit4);
#building plots
gResid3=ggplot(data=data.frame(resid3), aes(x=resid3)) + geom_density();
gResid4=ggplot(data=data.frame(resid4), aes(x=resid4)) + geom_density();
gFit3=ggplot(data=data.frame(resid3), aes(x=pred3,y=resid3)) + geom_jitter() +
geom_hline(yintercept=0);
gFit4=ggplot(data=data.frame(resid4), aes(x=pred4,y=resid4)) + geom_jitter() +
geom_hline(yintercept=0);
grid.arrange(gResid3,gResid4,gFit3,gFit4,nrow=2)
```


density
0.15
0.10
0.05
0.00
−4
0
4
resid3


density
0.15
0.10
0.05
0.00
−5.0
−2.5
0.0
2.5
5.0
resid4


resid3
4
0
−4
15
20
25
pred3


resid4
5.0
2.5
0.0
−2.5
−5.0
10
15
20
25
30
pred4