# Regression_models_assignment

*Hang_YU*

*February 3, 2017*

## Exploratory data analysis

Conduct exploratory data analysis using pair plot for a global view of the mtcars dataset.

```
library(ggplot2);library(GGally);data(mtcars);
ggpairs(mtcars,upper=list(continuous="smooth"),lower=list(continuous="cor"))
```

From the pair plot (see attachment), there are some relationships between mpg and those 10 predictors. Roughly, the current plot shows that automatic transmission (0) results in fewer mpg than manual gear (1).

## Hypothesis testing

The conclusion is validated through the single-sided student test with alternative hypothesis that automatic transmission has fewer mpg than manual cars:p value of the student test is smaller than 0.05 so we choose the alternative hypothesis.

```
t.test(subset(mtcars,am==0)$mpg,subset(mtcars,am==1)$mpg,alternative = "less")$p.value
```

```
## [1] 0.0006868192
```

Then we quantify the mpg difference using a single-variable linear model:

```
model1 <- lm(mpg~am,data=mtcars);model1$coefficients
```

```
## (Intercept)          am
##    17.147368    7.244939
```

From the coefficients, the mean of mpg increases by 7.244939 units if the transmission changes from auto to manual. However it is incorrect to conclude the relationship yet. We need to check if this relationship is the result of interference from residuals of other predictors. In other words, to eliminite the bias due to omitting significant variables.

## Model selection

At this point, we select appropriate predictors to build the linear regression for mpg prediction. This is done by using anova for nested models. We focus on the p values of anova test to determine the variable significance.

```
model2 <- update(model1,mpg~am+cyl,data=mtcars)
model3 <- update(model1,mpg~am+cyl+disp,data=mtcars)
model4 <- update(model1,mpg~am+cyl+disp+hp,data=mtcars)
model5 <- update(model1,mpg~am+cyl+disp+hp+drat,data=mtcars)
model6 <- update(model1,mpg~am+cyl+disp+hp+drat+wt,data=mtcars)
```

```
model7 <- update(model1,mpg~am+cyl+disp+hp+drat+wt+qsec,data=mtcars)
model8 <- update(model1,mpg~am+cyl+disp+hp+drat+wt+qsec+vs,data=mtcars)
model9 <- update(model1,mpg~am+cyl+disp+hp+drat+wt+qsec+vs+gear,data=mtcars)
model10 <- update(model1,mpg~am+cyl+disp+hp+drat+wt+qsec+vs+gear+carb,data=mtcars)
anova(model1,model2,model3,model4,model5,model6,model7,model8,model9,model10)$Pr
```

```
## [1]           NA 8.231226e-08 1.124136e-01 3.493497e-02 6.112088e-01
## [6] 1.275295e-02 1.826036e-01 8.621415e-01 7.136533e-01 8.121787e-01
```

According to the p values of anova test, am,cyl,hp, and wt and selected for the regression model due to their high variability contribution. Next, the variance inflation of the model with those four predictors is checked:

```
model <- lm(mpg~am+cyl+hp+wt,data=mtcars);vif(model)
```

```
##       am      cyl       hp       wt
## 2.546159 5.333685 4.310029 3.988305
```

The variable cyl has a high variance inflation factor of 5.333686 (generally the threshold is 5) so we remove this variable since it is correlated with at least one of the other variables (maybe hp and wt). After doing so the variance inflation becomes acceptable.

```
model <- lm(mpg~am+hp+wt,data=mtcars);vif(model)
```

```
##       am       hp       wt
## 2.271082 2.088124 3.774838
```

### Model diagnosis

Next, we check the coefficient of the final model:

```
model$coefficients
```

```
## (Intercept)          am          hp          wt
## 34.00287512  2.08371013 -0.03747873 -2.87857541
```

In comparison with model1, the mpg difference between auto and manual is dramatically dicreased, which means the initial conclusion is not that significant. The mpg difference is contributed by multiple variables rather than only transmission type. At last, we plot the final model (see attachment) to check the residuals and high-influence points, which shows that our model is basically reliable however with non-uniform residuals for fitted values caused by omitted variables.

```
par(mfrow=c(4,1));plot(model)
```

### Conclusions

Basically we can conclude that manual transmission has slightly higher mpg (2.08371013) than automatic transmission. However transmission type is not the pure causation of mpg values. Mpg values are influenced by multiple variables such as horse power and weight or even some other potential variables not shown in the dataset.

# Appendix