regression models on mtcars data

zhenyue zhu

April 24, 2016

Executive Summary

In this project, I am looking at a data set of a collection of cars. The variables in this data sets include a data frame with 32 observations on 11 variables. I will focus on the following two particular questions: 1. "Is an automatic or manual transmission better for MPG" 2. "Quantify the MPG difference between automatic and manual transmissions" The following analysis shows that transmission type has a strong effect on mpg. The difference of MPG for these two types is 1.81mpg, with manual car performs better.

Exploratory analysis

Before I do the regression of the model, I will transform some variables into factors. These includes cyl, vs, am, gear, carb, because these vaiables labeled are integers, to identify different classes, instead of continuous variables.

```
data(mtcars)
mtcars$cyl <- factor(mtcars$cyl)
mtcars$vs <- factor(mtcars$vs)
mtcars$am <- factor(mtcars$am, labels = c("Automatic", "Manual"))
mtcars$gear <- factor(mtcars$gear)
mtcars$carb <- factor(mtcars$carb)
```

I first use boxplot to show the difference between automatic and manual in terms of MPG. In figure 1, it is clear that the manual transmission has better MPG than automatic one. Next, I plot a pairwise graph (figure 2) to get a greater intuition of what variables has greater infuence on mpg. There is a linear relationship between MPG and each of cyl, disp, hp, drat, wt, qsec, and am.

Statistical inferences

At this step, we make the null hypothesis that transmision has no effect on MPG. We then use the two sample T-test to show it.

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

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

As we see above that p value is 0.0014, so we reject the null hypothesis and claims that **transmission type has a strong effect on MPG**.

Fit with various models

I first fit the model for mpg only use predictor am. Then try a different model to fit use all the predictors. Finally use step function to automatically select the best fit model.

```r
fit1 <- lm(mpg ~ am, data = mtcars)
fit2 <- lm(mpg ~ ., data = mtcars)
fit3 <- step(lm(mpg ~ ., data = mtcars), trace = 0)
summary(fit1)$adj.r.squared
```

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

```r
summary(fit2)$adj.r.squared
```

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

```r
summary(fit3)$adj.r.squared
```

```
## [1] 0.8400875
```

As shown above, the most significant predictors in determining the MPG are cyl, hp, wt and am. When comparing the fitting with predictor am, the best fitting has p value around 1.69e-9. (comparison in appendix) It means that adding predictors like cyl, hp and wt significantly improves the models accuracy. On the contrary when compares the model with all the variables are included. The p value is only 0.0018, which means that all variables as predictors are not as good as the model with only cyl, hp, wt and am. Also the model fit2 has the highest ajusted R-squared. Finally we choose fit2 which has predictors cyl+hp+wt+am as our final model.

```r
summary(fit3)$coeff
```

```
##                Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 33.70832390 2.60488618 12.940421 7.733392e-13
## cyl6        -3.03134449 1.40728351 -2.154040 4.068272e-02
## cyl8        -2.16367532 2.28425172 -0.947214 3.522509e-01
## hp          -0.03210943 0.01369257 -2.345025 2.693461e-02
## wt          -2.49682942 0.88558779 -2.819404 9.081408e-03
## amManual     1.80921138 1.39630450  1.295714 2.064597e-01
```

The coefficient means that for average weight 4cyl car, the mpg is around 33.7 When increase the cylinder to 6, mpg decrease by 3. When cylinder increase to 8, mpg will decrease again by 2.2. When hp increase by 1, mpg decrease by another 0.03. For 1000lbs weight increase in car, the mpg dcrease by 2.49. **Finally the manual transmission will gain the mpg by 1.81 compared with automatic transmission when other things are fixed.**

The uncertainty in the conclusion is that I do not include the interaction between variables, it will be better to see how different variables are related to each other. Then incooperate a model which has the interaction terms included for the fit.

Residual analysis (appendix Fig.3)

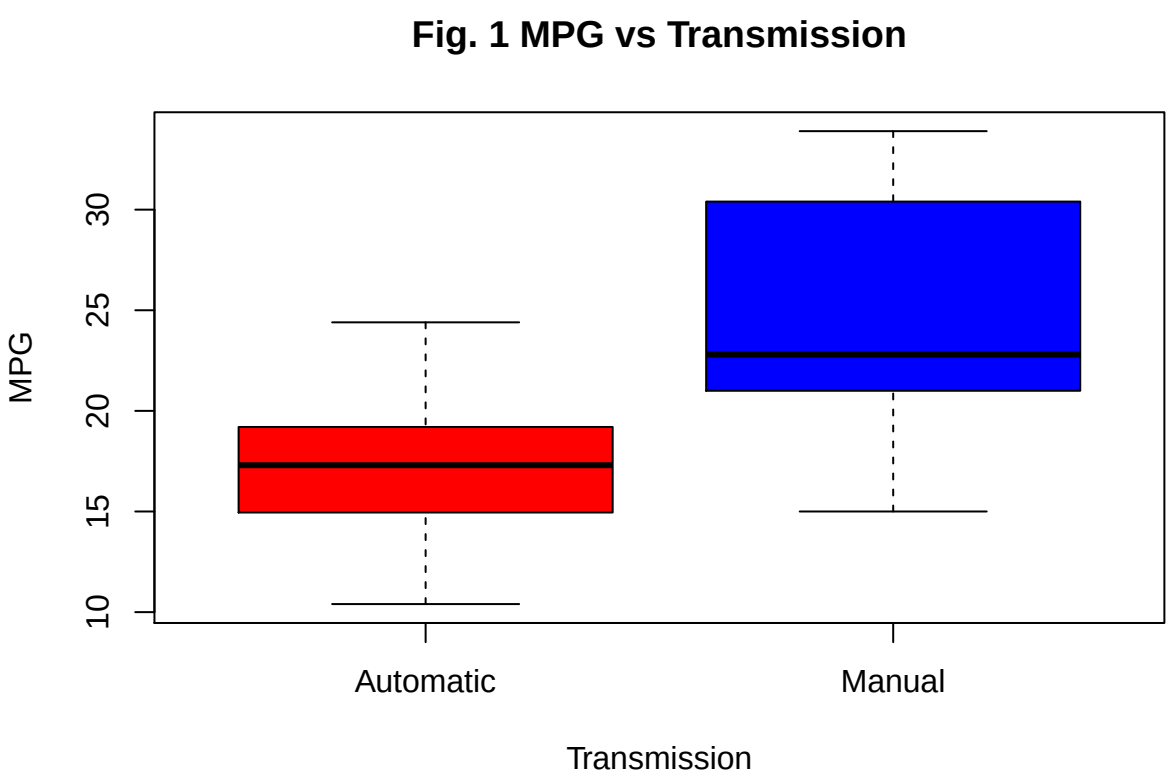
1. The Residuals vs Fitted plot shows no pattern between the residuals and fitted values indicating that this regression model has no other hidden patterns.
2. The QQ plot shows that the points line up as expected meaning that the distribtion is normal and our model predictions are accurate.
3. In both the Scale-Location plot and the Residuals vs Leverage plots, the points are in a group with none too far from the center indicating no point had too much leverage.

Appendix

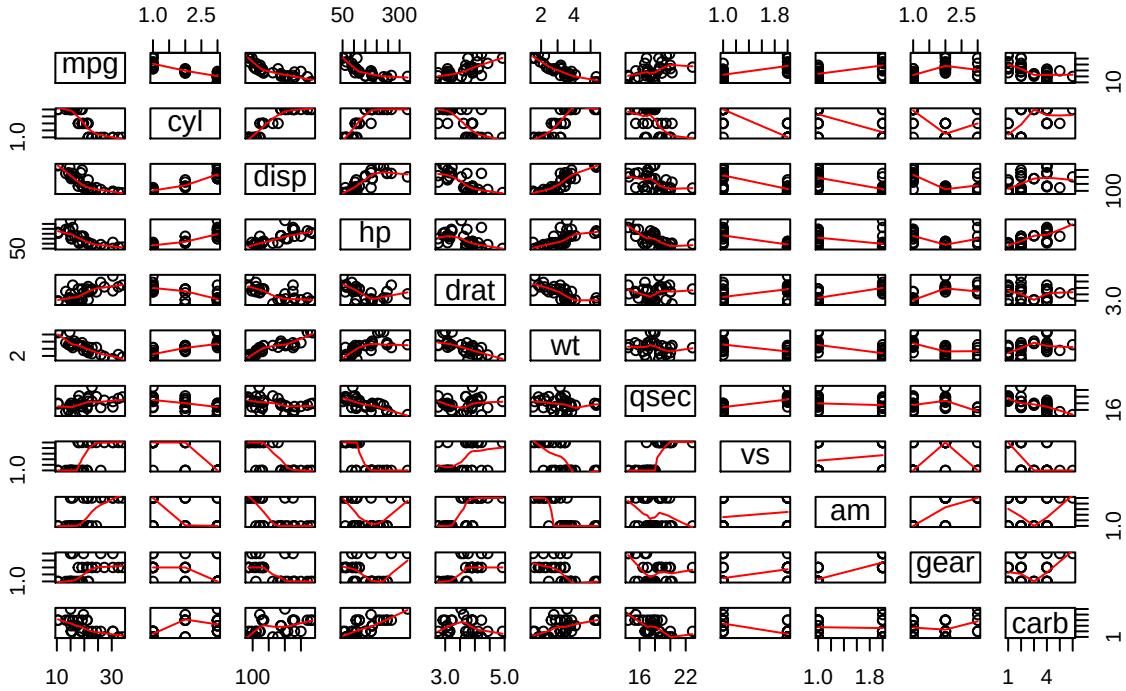
The 11 variables are as follows:

1. mpg - Miles/(US) gallon
2. cyl - Number of cylinders
3. disp - Displacement (cu.in.)
4. hp - Gross horsepower
5. drat - Rear axle ratio
6. wt - Weight (1000 lbs)
7. qsec - 1/4 mile time
8. vs - V/S
9. am - Transmission (0 = automatic, 1 = manual)
10. gear - Number of forward gears
11. carb - Number of carburetors

```
boxplot(mpg ~ am, data = mtcars,
        xlab = "Transmission", ylab = "MPG",
        main = "Fig. 1 MPG vs Transmission", col = c("red", "blue"),
        names = c("Automatic", "Manual"))
```



```
pairs(mtcars, panel = panel.smooth, main = "Fig.2 Pairwise plot of mtcars")
```



Compare of different models for the linear regression.

```
anova(fit1, fit2)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ cyl + disp + hp + drat + wt + qsec + vs + am + gear + carb
##   Res.Df   RSS Df Sum of Sq      F   Pr(>F)
## 1     30 720.9
## 2     15 120.4 15    600.49 4.9874 0.001759 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(fit1, fit3)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ cyl + hp + wt + am
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     30 720.90
## 2     26 151.03  4    569.87 24.527 1.688e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Fig.3 the residuals of the final model that we choose for the fit.

```
par(mfrow=c(2, 2))
title="Fig.3 residual"
plot(fit3)
```