RegressionModelAssignment

Ashwani Gupta

February 26, 2016

## Exploratory Ananlysis

library(ggplot2)

## Warning: package 'ggplot2' was built under R version 3.2.3

data(mtcars)  
mtcars[1:3, ] # Sample Data

## mpg cyl disp hp drat wt qsec vs am gear carb  
## Mazda RX4 21.0 6 160 110 3.90 2.620 16.46 0 1 4 4  
## Mazda RX4 Wag 21.0 6 160 110 3.90 2.875 17.02 0 1 4 4  
## Datsun 710 22.8 4 108 93 3.85 2.320 18.61 1 1 4 1

dim(mtcars)

## [1] 32 11

mtcars$cyl <- as.factor(mtcars$cyl)  
mtcars$vs <- as.factor(mtcars$vs)  
mtcars$am <- factor(mtcars$am)  
mtcars$gear <- factor(mtcars$gear)  
mtcars$carb <- factor(mtcars$carb)  
attach(mtcars)

## The following object is masked from package:ggplot2:  
##   
## mpg

Please see the **Appedix Figure 1 and 2** Clearly there appears to be a linear trend between mpg and weight for both the cases. Furthermore we can see a trend for MAnual and Auto transmission too. It is clear that manulal transmission car tends to have less weight and and high mpg. And as for the pair graph, we can see some higher correlations between variables like "wt", "disp", "cyl" and "hp".

## Inference

At this step, we make the null hypothesis as the MPG of the automatic and manual transmissions are from the same population (assuming the MPG has a normal distribution). We use the two sample T-test to show it.

result <- t.test(mpg ~ am)  
result$p.value

## [1] 0.001373638

result$estimate

## mean in group 0 mean in group 1   
## 17.14737 24.39231

result$conf.int

## [1] -11.280194 -3.209684  
## attr(,"conf.level")  
## [1] 0.95

By looking at the confidence interval we can clearly say there is no possibiliy that transmission is not playing an impact on mpg.And the mean for MPG of manual transmitted cars is about 7 more than that of automatic transmitted cars.

fullModel <- lm(mpg ~ .,data = mtcars)  
summary(fullModel)

This model has the Residual standard error as 2.833 on 15 degrees of freedom. And the Adjusted R-squared value is 0.779, which means that the model can explain about 78% of the variance of the MPG variable. However, none of the coefficients are significant at 0.05 significant level.

Then, we use backward selection to select some statistically significant variables.

stepModel <- step(fullModel, k=log(nrow(mtcars)))  
summary(stepModel) # results hidden

This model is "mpg ~ wt + qsec + am". It has the Residual standard error as 2.459 on 28 degrees of freedom. And the Adjusted R-squared value is 0.8336, which means that the model can explain about 83% of the variance of the MPG variable. All of the coefficients are significant at 0.05 significant level.

amIntWtModel<-lm(mpg ~ wt + qsec + am + wt:am, data=mtcars)  
summary(amIntWtModel) # results hidden

This model has the Residual standard error as 2.084 on 27 degrees of freedom. And the Adjusted R-squared value is 0.8804, which means that the model can explain about 88% of the variance of the MPG variable. All of the coefficients are significant at 0.05 significant level. This is a pretty good one.

At last lets fit a simple model

amModel<-lm(mpg ~ am, data=mtcars)  
summary(amModel) # results hidden

It shows that on average, a car has 17.147 mpg with automatic transmission, and if it is manual transmission, 7.245 mpg is increased. Clearly this model is worst among created.

Finally lets try to find out the best among we created

anova(amModel, stepModel, fullModel, amIntWtModel)   
confint(amIntWtModel) # results hidden

We end up selecting the model with the highest Adjusted R-squared value, "mpg ~ wt + qsec + am + wt:am".

summary(amIntWtModel)$coef

## Estimate Std. Error t value Pr(>|t|)  
## (Intercept) 9.723053 5.8990407 1.648243 0.1108925394  
## wt -2.936531 0.6660253 -4.409038 0.0001488947  
## qsec 1.016974 0.2520152 4.035366 0.0004030165  
## am1 14.079428 3.4352512 4.098515 0.0003408693  
## wt:am1 -4.141376 1.1968119 -3.460340 0.0018085763

Thus, the result shows that when "wt" (weight lb/1000) and "qsec" (1/4 mile time) remain constant, cars with manual transmission add 14.079 + (-4.141)\*wt more MPG (miles per gallon) on average than cars with automatic transmission. That is, a manual transmitted car that weighs 2000 lbs have 5.797 more MPG than an automatic transmitted car that has both the same weight and 1/4 mile time.

## Residual Analysis and Diagnostics

Please refer to the **Appendix: Figures** section for the plots. According to the residual plots, we can verify the following underlying assumptions:  
1. The Residuals vs. Fitted plot shows no consistent pattern, supporting the accuracy of the independence assumption.  
2. The Normal Q-Q plot indicates that the residuals are normally distributed because the points lie closely to the line.  
3. The Scale-Location plot confirms the constant variance assumption, as the points are randomly distributed.  
4. The Residuals vs. Leverage argues that no outliers are present, as all values fall well within the 0.5 bands.

As for the Dfbetas, the measure of how much an observation has effected the estimate of a regression coefficient, we get the following result:

sum((abs(dfbetas(amIntWtModel)))>1)

## [1] 0

Therefore, the above analyses meet all basic assumptions of linear regression and well answer the questions.

## Appendix: Figures

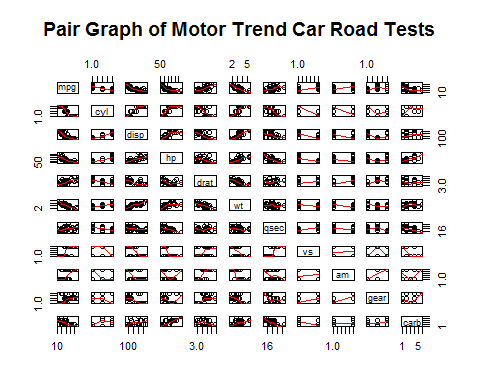
1. Plot of MPG vs Weight for different transmission

qplot(wt,mpg,data = mtcars,facets= .~am)



1. Pair Graph of Motor Trend Car Road Tests

pairs(mtcars,panel = panel.smooth,main="Pair Graph of Motor Trend Car Road Tests")



1. Residual Plots

par(mfrow = c(2, 2))  
plot(amIntWtModel)

