Regression Models- Course Project

Nidhi Shrivastava

Saturday, May 16, 2015

Coursera Course Project on Regression Models

Executive Summary

Motor Trend is a autombile trend magazine. It is interested in understanding in exploring the relationship between a set of variables and miles per gallon (MPG) (outcome). In this report, we will use a dataset from the 1974 Motor Trend US magazine to answer the following questions:

(A) Is an automatic or manual transmission better for MPG?
(B) Quantify the MPG difference between automatic and manual transmissions?

Using hypothesis testing and simple linear regression, we can conslude that there is a signficant difference between the mean MPG for automatic and manual transmission cars. To adjust for other confounding variables such as the weight and quarter mile time (acceleration) of the car, multivariate regression analysis was run to understand the impact of transmission type on MPG. The final model results indicates that weight and quarter mile time (acceleration) have signficant impact in quantifying the difference of mpg between automatic and manual transmission cars.

Data Processing

Reading in the mtcars dataset in **RSudio** for analysis Data was obtained in R-CRAN and its documentation can be found on : http://stat.ethz.ch/R-manual/R-devel/library/datasets/html/mtcars.html

```
data(mtcars)
names(mtcars)
```

```
##  [1] "mpg"  "cyl"  "disp" "hp"   "drat" "wt"   "qsec" "vs"   "am"   "gear"
## [11] "carb"
```

Here we see that our predictor variable of interest, am, is a numeric class. Since we are dealing with a dichotomous variable, let's convert this to a factor class and label the levels as Automatic and Manual for better interpretability.

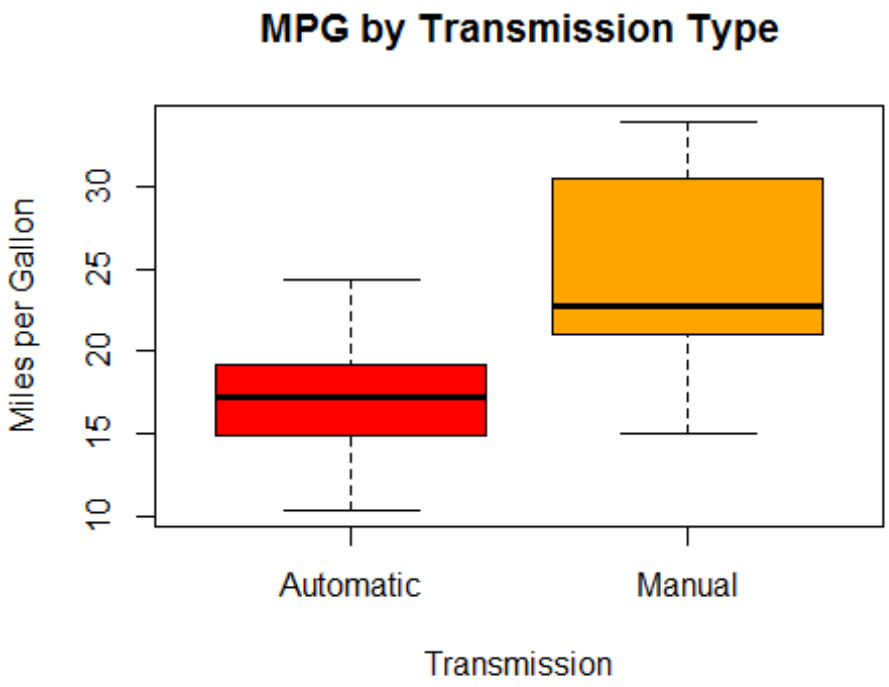
```
mtcars$am <- as.factor(mtcars$am)
levels(mtcars$am) <- c("Automatic", "Manual")
```

Exploratory Data Analysis

Since we will be running a linear regression, we want to make sure that its assumptions are met. The distribution of mpg is approximately normal and there are no apparent outliers skewing our data (Appendix 1 figure).

Now let's check how mpg varies by automatic versus manual transmission. A boxplot was created to examine the relationship between mpg and transmission type and it seems that manual transmission has better mpg compared with automatic transmission. But we need to do further analysis to confirm the finding.

```
boxplot(mpg~am, data = mtcars,
        col = c("red", "orange"),
        xlab = "Transmission",
        ylab = "Miles per Gallon",
        main = "MPG by Transmission Type")
```



##Means test (ttest) to analyze mileage of automatic vs. manual transmission

```
aggregate(mpg~am, data = mtcars, mean)
```

```
##          am      mpg
## 1 Automatic 17.14737
## 2    Manual 24.39231
```

The mean MPG of manual transmission cars is 7.24494 MPGs higher than that of automatic transmission cars. Is this a significant difference? We set our alpha-value at 0.5 and run a t-test to find out.

```
autoData <- mtcars[mtcars$am == "Automatic",]
manualData <- mtcars[mtcars$am == "Manual",]
t.test(autoData$mpg, manualData$mpg)
```

```
##
##  Welch Two Sample t-test
##
## data:  autoData$mpg and manualData$mpg
## t = -3.7671, df = 18.332, p-value = 0.001374
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -11.280194  -3.209684
## sample estimates:
## mean of x mean of y
##  17.14737  24.39231
```

The p-value is 0.001374, we may think it is ok to reject the null hypothesis and conclude automatic has low mpg compared with manual cars - however this assumption is based on all other characteristics of auto cars and manual cars are same (e.g: auto cars and manual cars have same weight distribution) - which needs to be further explored in the multiple linear regression analysis.

Correlation analysis

There are 32 observations of 11 variables, since we are interested in the relationshp between mpg and other variables, we first check the correlation between mpg and other variables by using the cor() function

```
data(mtcars)
sort(cor(mtcars)[1,])
```

```
##         wt        cyl       disp         hp       carb       qsec
## -0.8676594 -0.8521620 -0.8475514 -0.7761684 -0.5509251  0.4186840
##       gear         am         vs       drat        mpg
##  0.4802848  0.5998324  0.6640389  0.6811719  1.0000000
```

From the correlation data, we could see cyl, hp, wt and carb are negatively correlated with mpg. In addition to am (which by default must be included in our regression model), we see that wt, cyl, disp, and hp are highly correlated with our dependent variable mpg.

Modeling Procedures

Standard Deviation of MPG by Transmission Type

```
mtcars_vars <- mtcars[, c(1, 6, 7, 9)]
by(mtcars_vars$mpg, mtcars_vars$am, sd)
```

```
## mtcars_vars$am: 0
## [1] 3.833966
```

```
## ------------------------------------------------------
## mtcars_vars$am: 1
## [1] 6.166504
```

Levene's Test for Homogeneity of Variance

```
library(car)

## Warning: package 'car' was built under R version 3.1.3

leveneTest(mpg ~ factor(am), data = mtcars_vars)

## Levene's Test for Homogeneity of Variance (center = median)
##       Df F value  Pr(>F)
## group  1  4.1876 0.04957 *
##       30
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Simple Linear Regression

```
fit <- lm(mpg~am, data = mtcars)
summary(fit)

##
## Call:
## lm(formula = mpg ~ am, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.3923 -3.0923 -0.2974  3.2439  9.5077
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   17.147      1.125  15.247 1.13e-15 ***
## am             7.245      1.764   4.106 0.000285 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.902 on 30 degrees of freedom
## Multiple R-squared:  0.3598, Adjusted R-squared:  0.3385
## F-statistic: 16.86 on 1 and 30 DF,  p-value: 0.000285
```

We do not gain much more information from our hypothesis test using this model. Interpreting the coefficient and intercepts, we say that, on average, manual transmission cars have 7.245 MPGs more than automatic transmission. In addition, we see that the R^2 value is 0.3598. This means that our model only explains 35.98% of the variance.

We need to understand the impact of transmission in conjunction with other factors to quantify the mpg difference between automatic and manual transmission.

Multivariate regression analysis

```
# Here we adopt a stepwise algorithm to choose the best model by using step()
function
stepmodel = step(lm(data = mtcars, mpg ~ .),trace=0,steps=10000)
summary(stepmodel)
```

```
##
## Call:
## lm(formula = mpg ~ wt + qsec + am, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4811 -1.5555 -0.7257  1.4110  4.6610
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   9.6178     6.9596   1.382 0.177915
## wt           -3.9165     0.7112  -5.507 6.95e-06 ***
## qsec          1.2259     0.2887   4.247 0.000216 ***
## am            2.9358     1.4109   2.081 0.046716 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.459 on 28 degrees of freedom
## Multiple R-squared:  0.8497, Adjusted R-squared:  0.8336
## F-statistic: 52.75 on 3 and 28 DF,  p-value: 1.21e-11
```

This shows that in adddition to transmission, weight of the vehicle as well as accelaration speed have the highest relation to explaining the variation in mpg. The adjusted R-Squared is 83.4% which means that the model explains 84% of the variation in mpg indicating it is a robust and highly predictive model.

Final model to quantify the mpg difference between automatic and manual transmission

```
# We include 3 variables wt, qsec and am. This model captured 84% of total
variance.
bestfit <- lm(mpg~am + wt + qsec, data = mtcars)
anova(fit, bestfit)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ am + wt + qsec
##   Res.Df    RSS Df Sum of Sq      F   Pr(>F)
## 1     30 720.90
## 2     28 169.29  2    551.61 45.618 1.55e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

This model captured 84% of the overall variation in mpg. With a p-value of 3.745e-09, we reject the null hypothesis and claim that our multivariate model is significantly different from our simple linear regression model.

Before we report the details of our model, it is important to check the residuals for any signs of non-normality and examine the residuals vs. fitted values plot to spot for any signs of heteroskedasticity.The residual diagnostocs show normality and no evidence of heteroskedasticity (Appendix 2).

```
#final model results
summary(bestfit)

##
## Call:
## lm(formula = mpg ~ am + wt + qsec, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4811 -1.5555 -0.7257  1.4110  4.6610
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   9.6178     6.9596   1.382 0.177915
## am            2.9358     1.4109   2.081 0.046716 *
## wt           -3.9165     0.7112  -5.507 6.95e-06 ***
## qsec          1.2259     0.2887   4.247 0.000216 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.459 on 28 degrees of freedom
## Multiple R-squared:  0.8497, Adjusted R-squared:  0.8336
## F-statistic: 52.75 on 3 and 28 DF,  p-value: 1.21e-11
```

Results Summary:

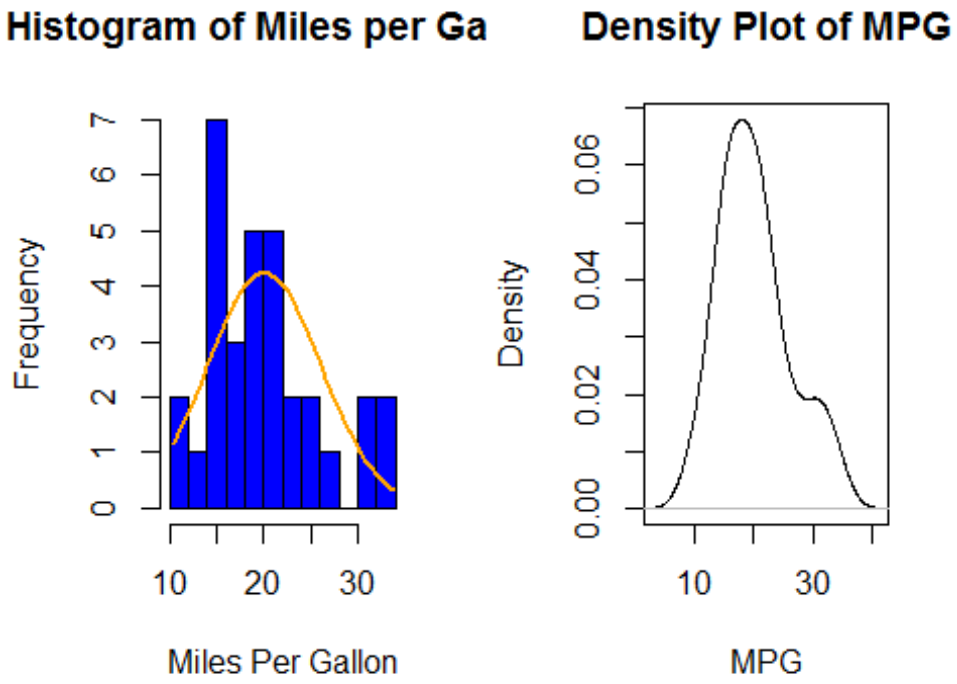
This model explains 84% of the variance in miles per gallon (mpg). Moreover, we see that wt and qsec did indeed confound the relationship between *am* and *mpg* (mostly wt). Now when we read the coefficient for **am**, we say that, on average, manual transmission cars have 2.94 MPGs more than automatic transmission cars. However this effect is much lower than when we did not adjust for weight and qsec.

Therefore given the above analysis, the question of auto car and manual car is not anwsered and have to be considered in the context of weight and accelaration speed.

Appendix

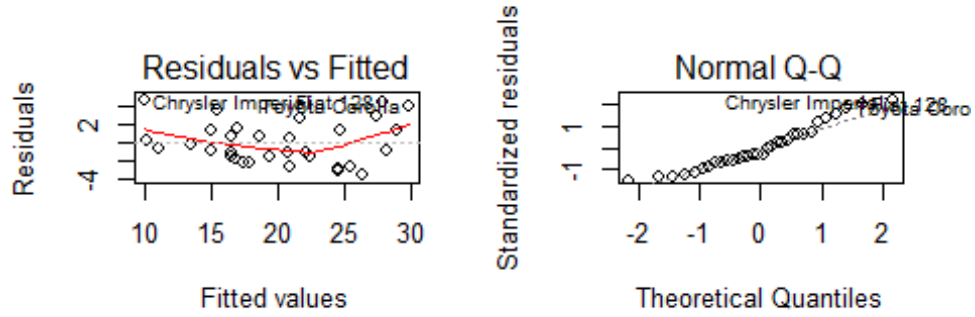
Appendix 1 : Let's plot the dependent variable mpg to check its distribution.

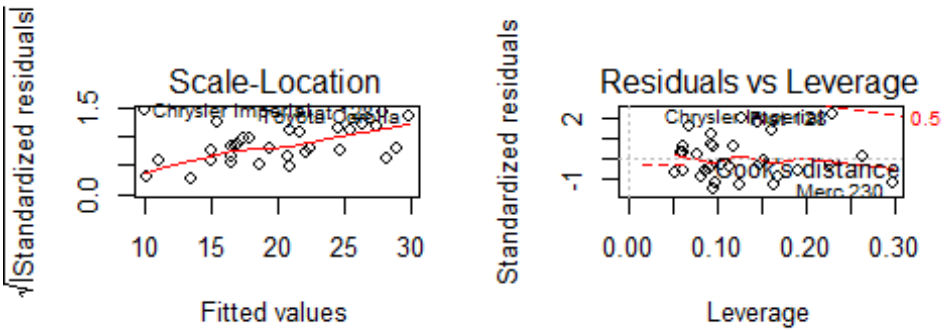
```
par(mfrow = c(1, 2))
# Histogram with Normal Curve
x <- mtcars$mpg
h<-hist(x, breaks=10, col="blue", xlab="Miles Per Gallon",
   main="Histogram of Miles per Gallon")
xfit<-seq(min(x),max(x),length=40)
yfit<-dnorm(xfit,mean=mean(x),sd=sd(x))
yfit <- yfit*diff(h$mids[1:2])*length(x)
lines(xfit, yfit, col="orange", lwd=2)

# Kernel Density Plot
d <- density(mtcars$mpg)
plot(d, xlab = "MPG", main ="Density Plot of MPG")
```



Appendix 2 : Residual diagnostics for final multivariate model

```
par(mfrow = c(2,2))
plot(bestfit)
```


Residuals vs Fitted
Residuals
Fitted values
Chrysler Imperial
Fiat 128
Toyota Corolla
Normal Q-Q
Standardized residuals
Theoretical Quantiles
Chrysler Imperial
Fiat 128
Toyota Corolla


Scale-Location
√|Standardized residuals|
Fitted values
Chrysler Imperial
Fiat 128
Toyota Corolla
Residuals vs Leverage
Standardized residuals
Leverage
Cook's distance
0.5
Chrysler Imperial
Fiat 128
Merc 230