Motor Trend Analysis

Morozov Gleb

Sunday, June 14, 2015

Executive Summary

Do cars with manual transmission behave more favorably than automatic transmission cars with respect to fuel efficiency? It is a common belief that changing gears manually results in better fuel management To do this analysis we will use a data set that examines the fuel efficency and 10 aspects of automobile design and performance for 32 automobiles (all 1973 - 1974 models). Out of the 32 cars, 13 have manual transmissions and 19 have automatic transmissions.

In this data set on average there is a difference in fuel efficency depending on transmission type such that on average manual vehicles achieve a fuel effiency of 7.2 miles per gallon more than automatic vehicles.

However, transmission type is not a particularly good predictor of fuel efficiency. By applying analysis of variance (ANOVA) to the dataset, and building a number of models, we were able to identify that the number of cylinders and the weight of the automobile are good predictors of fuel efficiency, achieving an adjusted R squared of 0.82. If we add transmission type to this model, then the difference in fuel effiency for a manual transmission is much smaller, just 0.18 miles per gallon for a vehicle with the same weight and number of cylinders.

Therefore we conclude that number of cylinders and weight are good predictors of fuel efficiency, but transmission type is not.

The data set

```
library(ggplot2)
library(datasets)
data(mtcars)
attach(mtcars)
```

The data set was extracted from the 1974 edition of Motor Trend US Magazine and it deals with 1973 - 1974 models. It consists of 32 observations on 11 variables:

- `mpg`: Miles per US gallon
- `cyl`: Number of cylinders
- `disp`: Displacement (cubic inches)
- `hp`: Gross horsepower
- `drat`: Rear axle ratio
- `wt`: Weight (lb / 1000)
- `qsec`: 1 / 4 mile time

- `vs`: V/S
- `am`: Transmission (0 = automatic, 1 = manual)
- `gear`: Number of forward gears
- `carb`: Number of carburetors

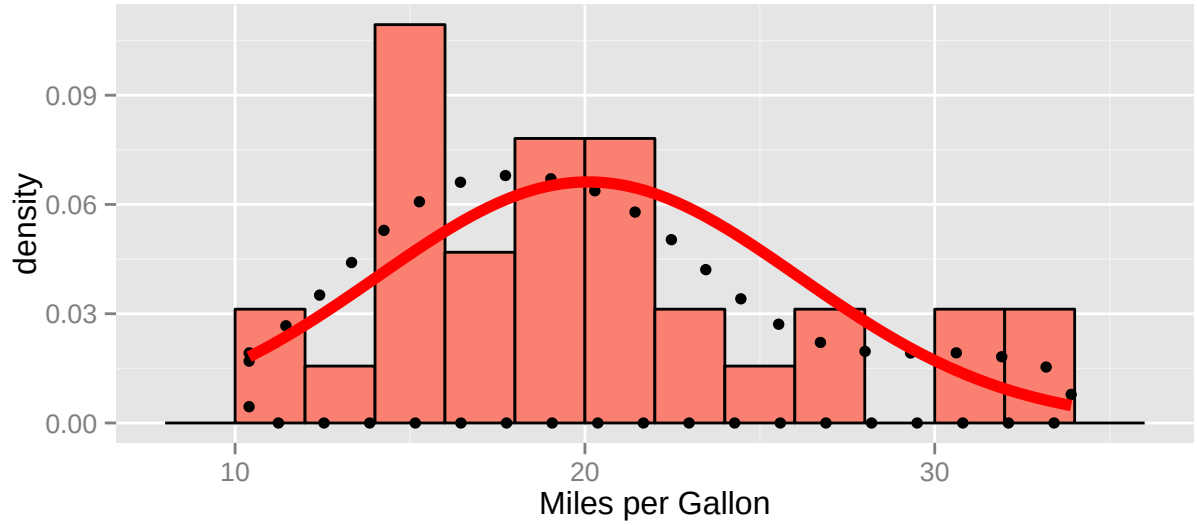
Here we see that our predictor variable of interest, `am`, is a numeric class. Since we are dealing with a dichotomous variable, let's convert this to a factor class and label the levels as `Automatic` and `Manual` for better interpretability.

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

Exploratory Data Analysis

Since we will be running a linear regression, we want to make sure that its assumptions are met. Let's plot the dependent variable `mpg` to check its distribution.

```
ggplot(mtcars, aes(x = mpg)) +
        geom_histogram(binwidth = 2, colour = "black", fill ="salmon",
                       aes(y = ..density..)) +
        geom_density(size = 2, colour = "black", linetype = "dotted") +
        stat_function(fun = dnorm, colour = "red",
                      size = 2, arg = list(mean=mean(mpg), sd= sd(mpg))) +
        labs(x = "Miles per Gallon")
```



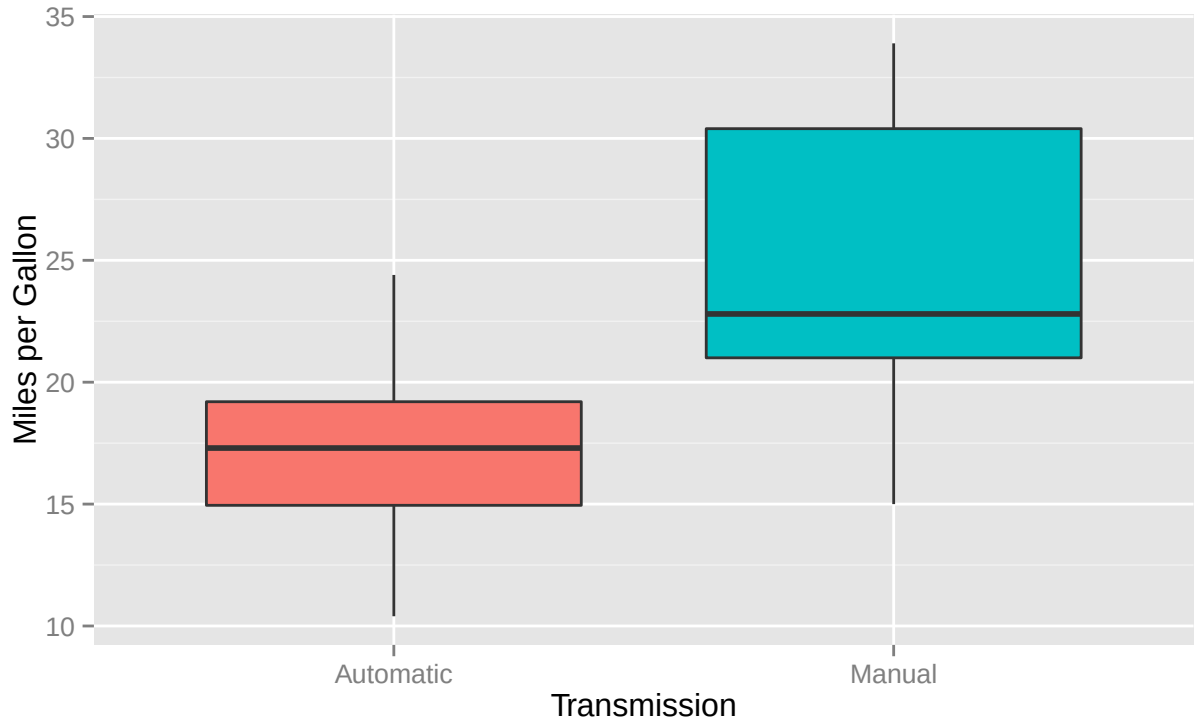
We want to know if our sample for `mpg` is from normaly distributed population. The test result depends on p-value. When $p < 0.05$, then population is likely not normaly distributed. When $p > 0.05$ there is no such evidence.

```
shapiro.test(mtcars$mpg)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  mtcars$mpg
## W = 0.94756, p-value = 0.1229
```

Plot and large p-value indicates that population of `mpg` is likely to be normaly distributed. Now let's check how `mpg` varies by automatic versus manual transmission.

```
ggplot(mtcars, aes(x = am, y=mpg)) + geom_boxplot(aes(fill = am)) +
        labs(x = "Transmission", y = "Miles per Gallon") + guides(fill=F)
```



We want to know if there is any difference at all, in fuel consumption for two types of transmision. Again, p-value will provide an answer. $p < 0.05$ indicates that means are likely different. $p > 0.05$ provides no such evidence.

```
t.test(mtcars$mpg ~ mtcars$am)
```

```
##
##  Welch Two Sample t-test
##
## data:  mtcars$mpg by mtcars$am
## 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 in group Automatic    mean in group Manual
##                17.14737                24.39231
```

Low p-value indicates difference for these two groups. Also the numbers representing the mean fuel consumption for manual and automatic type clearly are diffrent.

Model Selection

In order to select the best model, we need to find out which variables have biggest impact on fuel consumption, beside transmission type. We will use "Backward stepwise regression", which starts with all predictors and removes those which are not statistically significant.

```
base.model <- lm(mpg ~ ., data = mtcars)
fit.model <- step(base.model, direction="backward", trace = 0)
summary(fit.model)
```

```
##
## Call:
## lm(formula = mpg ~ cyl + hp + wt + am, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.9387 -1.2560 -0.4013  1.1253  5.0513
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 33.70832    2.60489  12.940 7.73e-13 ***
## cyl6        -3.03134    1.40728  -2.154  0.04068 *
## cyl8        -2.16368    2.28425  -0.947  0.35225
## hp          -0.03211    0.01369  -2.345  0.02693 *
## wt          -2.49683    0.88559  -2.819  0.00908 **
## amManual     1.80921    1.39630   1.296  0.20646
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.41 on 26 degrees of freedom
## Multiple R-squared:  0.8659, Adjusted R-squared:  0.8401
## F-statistic: 33.57 on 5 and 26 DF,  p-value: 1.506e-10
```

Let's analyse the summary and see if this model is statistically significant. We'll start from the end because there we can find the most important statistics: - Model's p-value of less than 0.05 also indicates that this model likely is significant - R-squared as a measure of model's quality, represents a fraction of outcome's variance explained by the model. In this case the model explains 0.8401 (adjusteed value) or 84.01% of variance – Model is significant if any of coeficientts are non-zero. Clearly this is true, therefore this model is significant. - The model is suggesting "weight", "horsepower" and "cylinder" as significant variables. - However, transmission is marked as not significan in this model

Let's test significance of suggested model comparing it with the basic model:

```r
basic.model <- lm(mpg ~ am, data = mtcars)
anova(basic.model, fit.model)
```

```
## 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
```

Analysis of variance (ANOVA) resulted in p-value much lower than 0.05, which indicates that `fit.model` is statisticaly significant.

Residuals analysis

Points in `Residuals vs Fitted` are randomly scattered with no obvious pattern. Points in Q-Q plot are on the line, indicating that residuals are normaly distributed.

```r
par(mfrow=c(2, 2))
plot(fit.model)
```