Regression Models Course Project

Kyle

July 14, 2016

Peer Graded Assignment: Regression Models Coursera Project

Answer the Following Questions:

1) "Is an automatic or manual transmission better for MPG"

The manual vehicle is better due to a p-value of .0006 from a two sample t-test in the means of the two samples.

2) "Quantify the MPG difference between automatic and manual transmissions"

We are 95% confident the true difference in the means of manual and automatic transmissions lies between 3.913 MPG and Positive Infinity.

Step 1: Load the Data into R

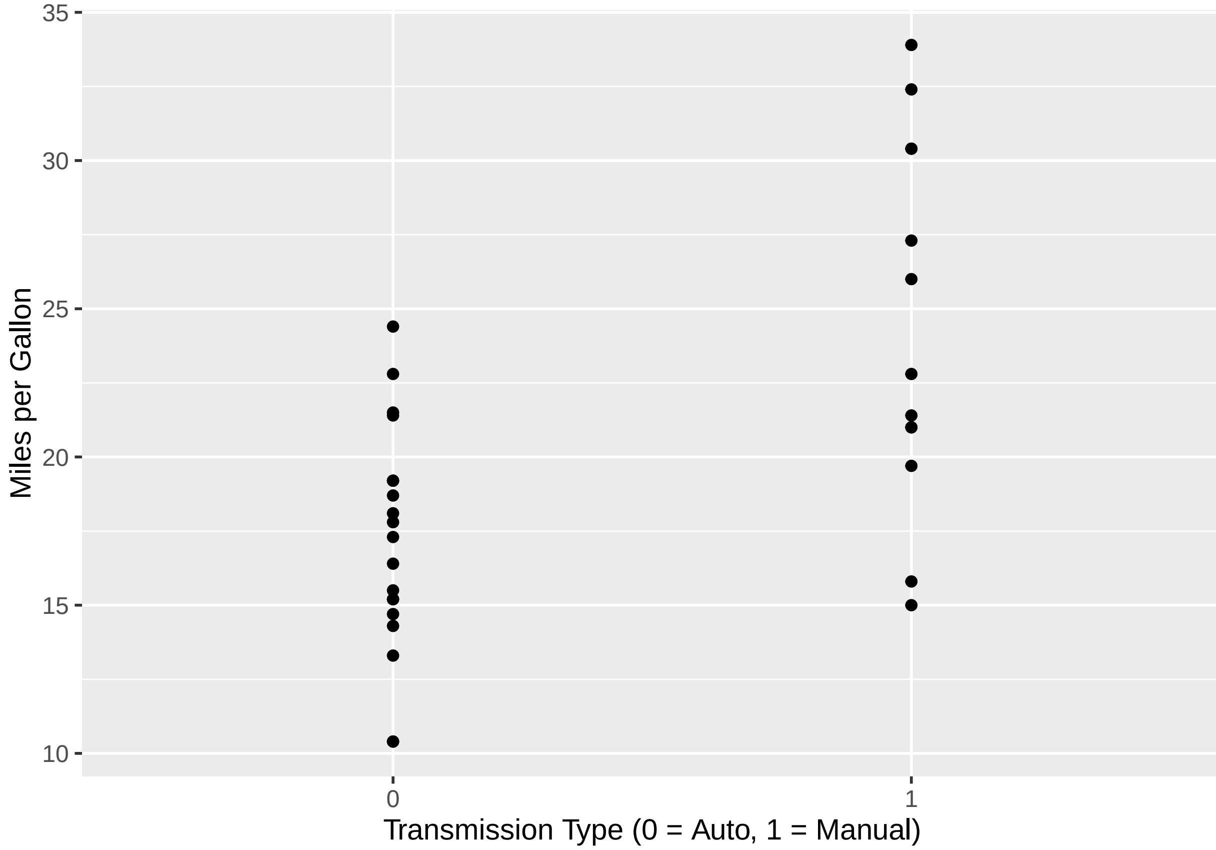
```
library(datasets)
cars <- datasets::mtcars
head(cars)
```

```
##                    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
## Hornet 4 Drive    21.4   6  258 110 3.08 3.215 19.44  1  0    3    1
## Hornet Sportabout 18.7   8  360 175 3.15 3.440 17.02  0  0    3    2
## Valiant           18.1   6  225 105 2.76 3.460 20.22  1  0    3    1
```

Step 2: Exploratory Data Analysis

Now that we have the data loaded into R, it's time to graph the mpg based on automatic and manual transmissions.

```
library(ggplot2)
g <- ggplot(data = cars, aes(x = as.factor(am), y = mpg)) +
     geom_point()+
     labs(x = "Transmission Type (0 = Auto, 1 = Manual)", y = "Miles per Gallon")
print(g)
```



As we can see, the engines of manual vehicles run more efficiently based on MPG than their automatic counterparts. To solidify this notion, we will run a two sample T test to see if the means of these two groups are in fact different.

Step 3: Split data by manual and automatic transmission type and find the mean of each groups MPG's.

```
autos <- subset(cars, am == "0")
mans <- subset(cars, am == "1")
mean(autos$mpg)
```

```
## [1] 17.14737
```

```
mean(mans$mpg)
```

```
## [1] 24.39231
```

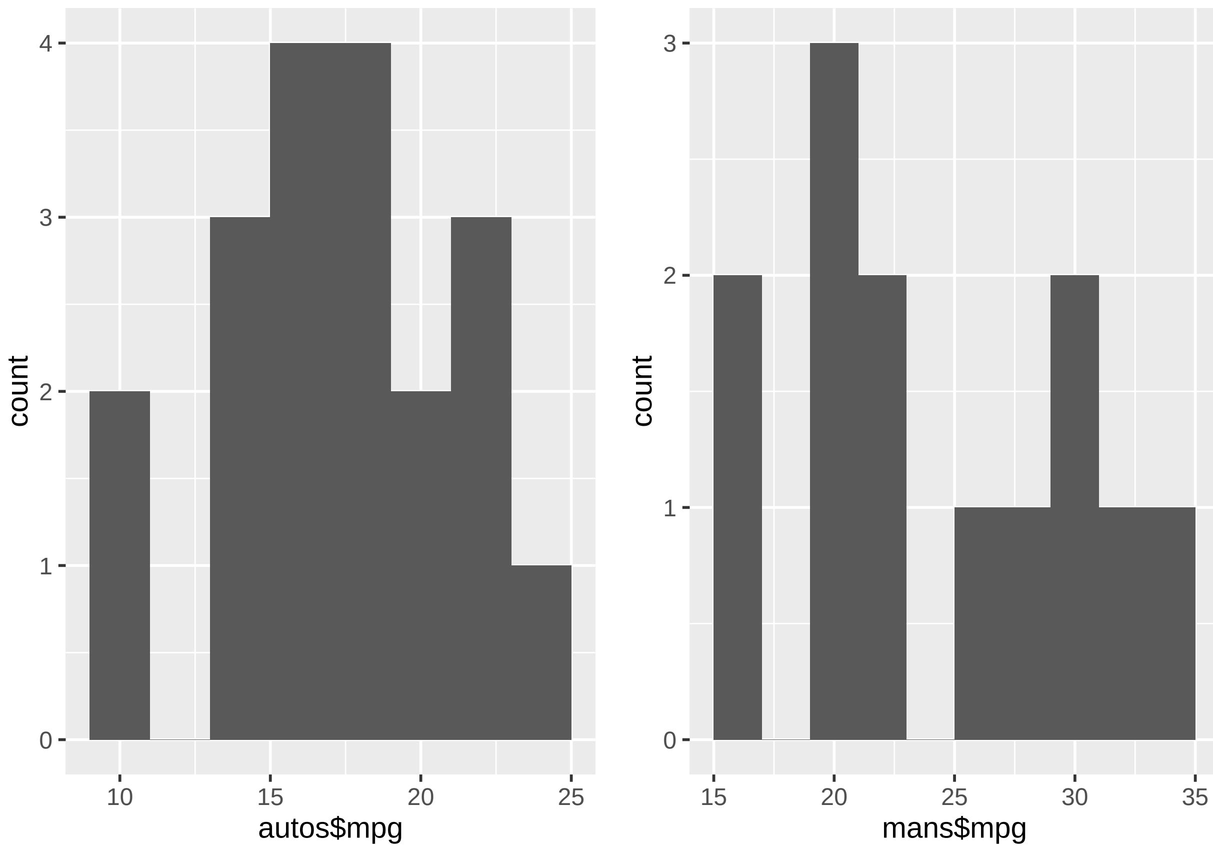
Step 4: Check criteria to see if two-sample t-test is appropriate.

Criteria #1: Samples are Drawn from Normal distributions

```r
require(gridExtra)
```

```
## Loading required package: gridExtra
```

```r
plot1 <- qplot(autos$mpg, binwidth = 2)
plot2 <- qplot(mans$mpg, binwidth = 2)
grid.arrange(plot1, plot2, ncol=2)
```



Criteria #2: Samples must be independent.

Since the probability of being a manual car does not affect the probability of being an automatic vehicle, this condition is met.

Step 5: T-test

```r
t.test(mans$mpg, autos$mpg, alternative = "greater")
```

```
##
##  Welch Two Sample t-test
##
## data:  mans$mpg and autos$mpg
## t = 3.7671, df = 18.332, p-value = 0.0006868
```

```
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  3.913256      Inf
## sample estimates:
## mean of x mean of y
##  24.39231  17.14737
```

With a p-value of 0.0006, we can conclude: there is sufficient evidence that manual transmissions consume less gasoline and are more fuel efficient. There is evidence against the hypothesis that the means of the two groups are the same. Finally, to quantify this difference betweeen the two types of vehicles, we can use our t-test to create a confidence interval. Recalling that our interval was (3.913, + Inf), we are 95% confident the true difference lies within these bounds.

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

```
##
## Call:
## lm(formula = mpg ~ am, data = cars)
##
## 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
```

Now that we see the R^2 valus is around 36% our goal is to find a model where a higher percent of variance can be explained.

```
library(car)
fit_all <- lm(mpg ~ ., cars)
vif(fit_all)
```

```
##       cyl      disp        hp      drat        wt      qsec        vs
## 15.373833 21.620241  9.832037  3.374620 15.164887  7.527958  4.965873
##        am      gear      carb
##  4.648487  5.357452  7.908747
```

```
sqrt(vif(fit_all))
```

```
##      cyl     disp       hp     drat       wt     qsec       vs       am
```

```
## 3.920948 4.649757 3.135608 1.837014 3.894212 2.743712 2.228424 2.156035
##     gear     carb
## 2.314617 2.812249
```

With the standard deviations of inflation factors stated. We will take the highest and update out model systematically to see which model will best predict MPG of vehicles and have the high R^2 value.

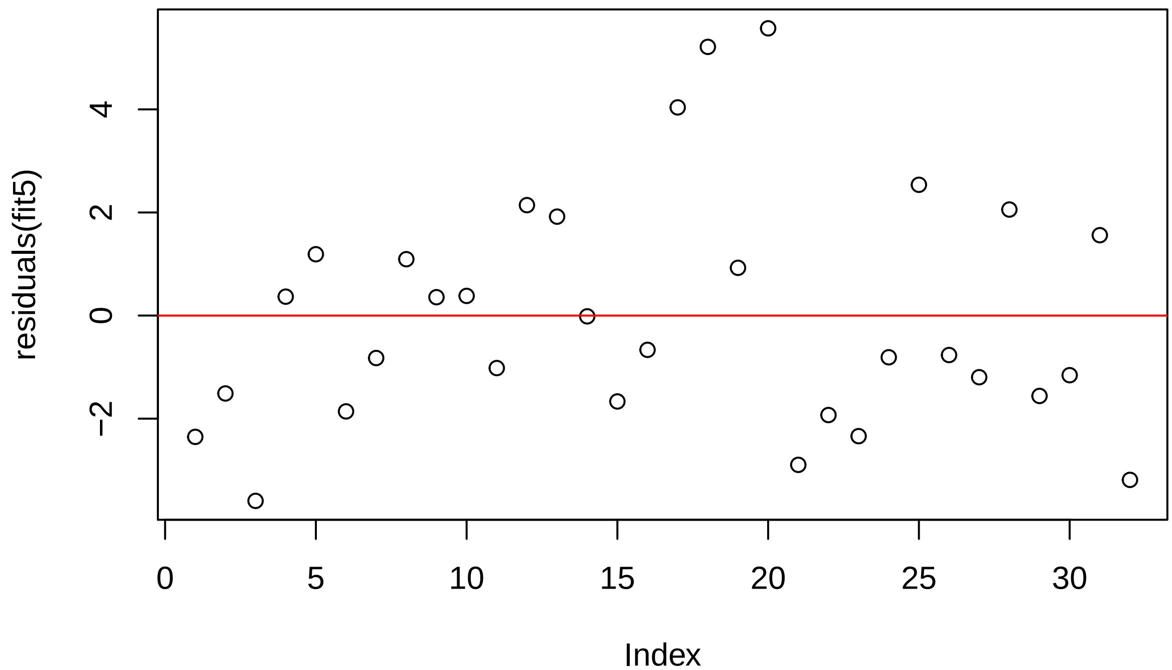
```r
fit1 <- lm(mpg ~ am, data = cars)
fit2 <- update(fit1, mpg ~ am + cyl)
fit3 <- update(fit1, mpg ~ am + cyl + disp)
fit4 <- update(fit1, mpg ~ am + cyl + disp + hp)
fit5 <- update(fit1, mpg ~ am + cyl + disp + hp + wt)

anova(fit1, fit2, fit3, fit4, fit5)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ am + cyl
## Model 3: mpg ~ am + cyl + disp
## Model 4: mpg ~ am + cyl + disp + hp
## Model 5: mpg ~ am + cyl + disp + hp + wt
##   Res.Df    RSS Df Sum of Sq       F    Pr(>F)
## 1     30 720.90
## 2     29 271.36  1    449.53 71.6522 6.037e-09 ***
## 3     28 252.08  1     19.28  3.0732  0.091376 .
## 4     27 216.37  1     35.71  5.6925  0.024609 *
## 5     26 163.12  1     53.25  8.4872  0.007257 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Based off the p-value from the F-statistics, Model 5 and Model 2 end up being the most statistically significant. With that being said, of the two values, a look at the RSS values yields Model5 as the better model in this particular case. Finally, to make sure our model is not flawed, let's look at the residuals to see if there is a pattern.

```r
plot(residuals(fit5))
abline(h = 0, col = "red")
```



Since there is no pattern of consistently over estimating or underestimating the residuals, this model appears ready for action.