Coursera - Regression Models course - Course Project 1

Martin Cote

May 24, 2015

Executive Summary

Using the "mtcars" dataset provided within the R environment, answering the following questions:

- Is an automatic or manual transmission better for MPG
- Quantify the MPG difference between automatic and manual transmissions

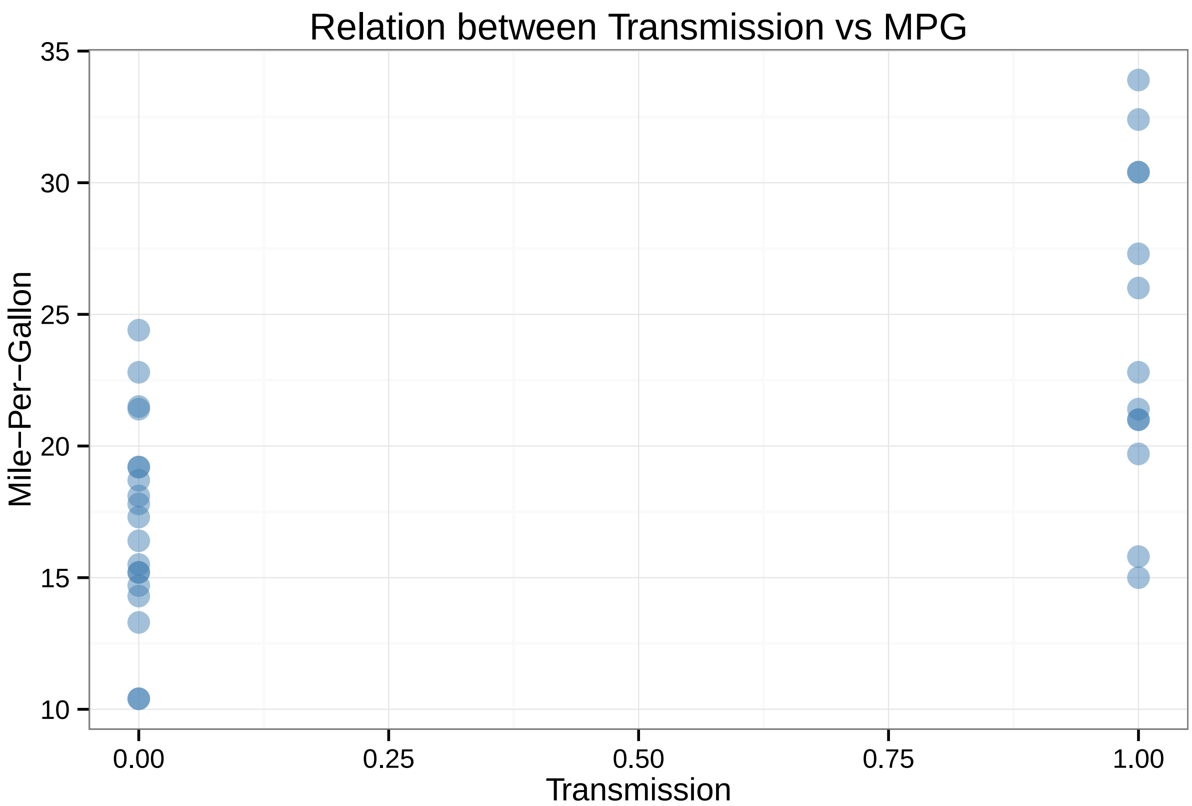
Pre-Processing

Loading libraries and data.

Exploratory Data Analysis

Is transmission (i.e. 'am', 0 = automatic and 1 = manual) a good predictor for the 'mpg'? Using a plot to demonstrate the relation and the linear regression between the variables.

```
g <- ggplot(mtcars, aes(am, mpg)) +
  geom_point(color = "steelblue", size=4, alpha=1/2) +
  #geom_smooth(size=2, method="lm", formula=mtcars$mpg ~ mtcars$am) +
  xlab("Transmission") +
  ylab("Mile-Per-Gallon") +
  labs(title="Relation between Transmission vs MPG") +
  theme_bw()
g
```




```r
t.test(mtcars$am, mtcars$mpg, paired=FALSE, var.equal=FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  mtcars$am and mtcars$mpg
## t = -18.413, df = 31.425, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -21.86356 -17.50519
## sample estimates:
## mean of x mean of y
##   0.40625  20.09062
```

We observe the p-value of the Student's Test Distribution is far below 0.05 which indicates there is a strong relation between Transmission (i.e. 'am') and MPG (i.e. 'mpg').

Predictions and Correlation

```r
cor(mtcars$mpg, mtcars$am)
```

```
## [1] 0.5998324
```

```r
var(mtcars$mpg, mtcars$am)
```

```
## [1] 1.803931
```

```r
# 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
```

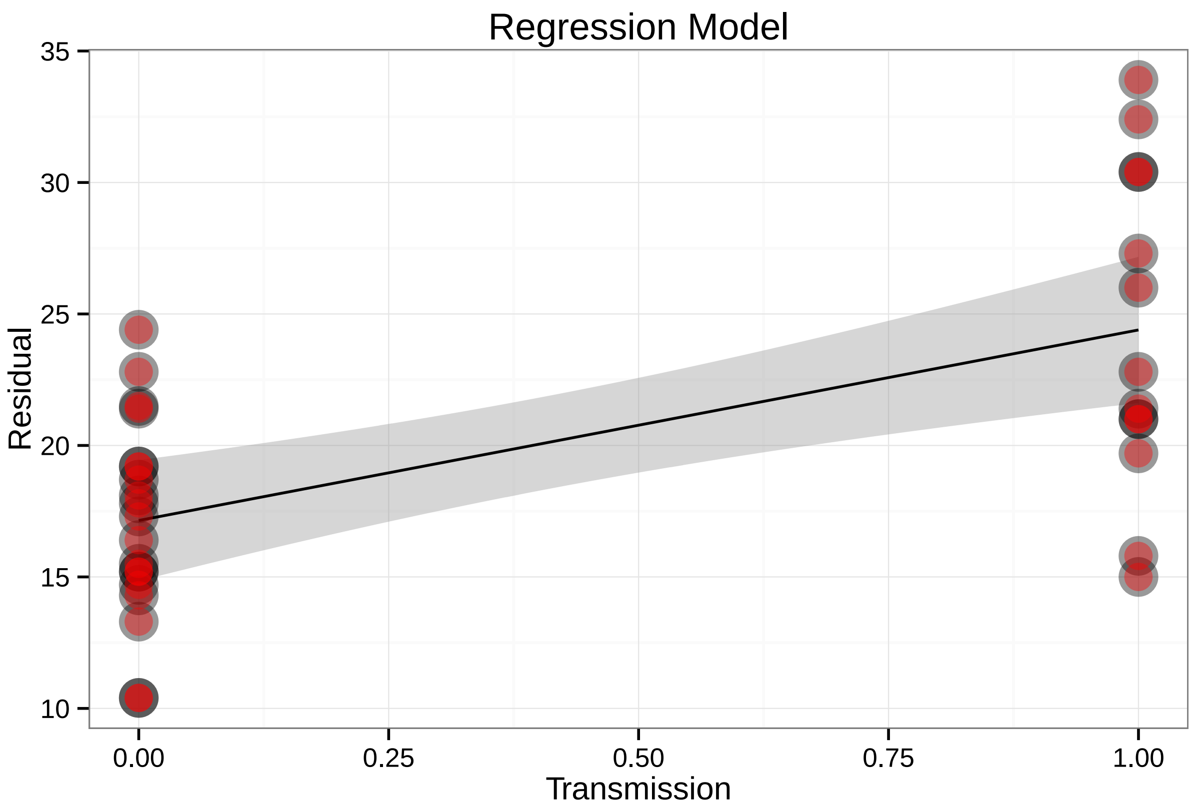
```r
sumCoef <- summary(fit)$coefficients
sumCoef[2,1] + c(-1, 1) * qt(.975, df = fit$df) * sumCoef[2, 2]
```

```
## [1]  3.64151 10.84837
```

```r
#fit <- lm(mpg ~ am -1, data = mtcars)
#summary(fit)
```

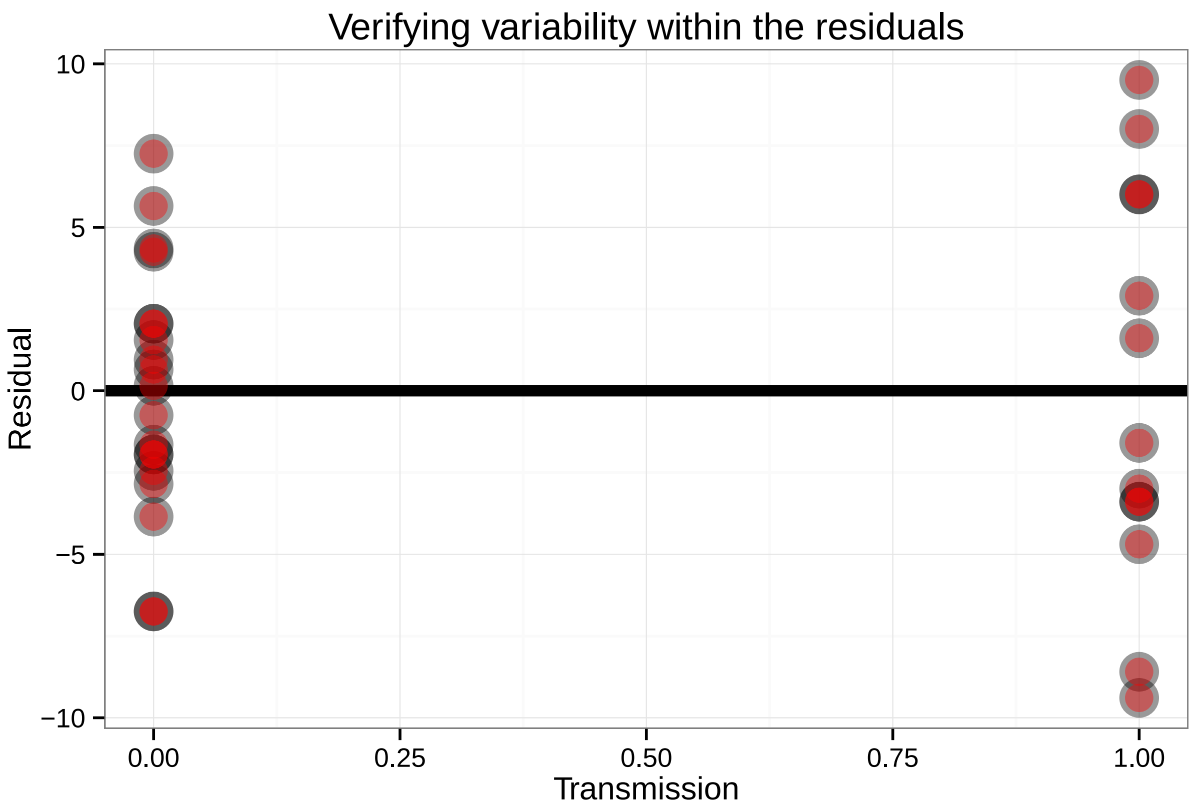
Regression Model…

```r
g = ggplot(data = mtcars, aes(x = am, y = mpg)) +
  geom_smooth(method = "lm", colour = "black") + # linear model regression
  geom_point(size = 7, colour = "black", alpha = 0.4) + # black contour
  geom_point(size = 5, colour = "red", alpha = 0.4) + # red center
  xlab("Transmission") +
  ylab("Residual") +
  labs(title="Regression Model") +
  theme_bw()
g
```



Studying the residual plot. . .

```
g = ggplot(data = mtcars, aes(x = am, y = resid(lm(mpg ~ am))) ) +
  geom_hline(yintercept = 0, size = 2) +
  geom_point(size = 7, colour = "black", alpha = 0.4) + # black contour
  geom_point(size = 5, colour = "red", alpha = 0.4) + # red center
  xlab("Transmission") +
  ylab("Residual") +
  labs(title="Verifying variability within the residuals") +
  theme_bw()
g
```




```
# Validaiting the residuals mean to be 0:
mean(resid(lm(mpg ~ am, data = mtcars)))
```

```
## [1] -6.591949e-17
```

The residual plot seems to indicate a greater variability in MPG for a car *with* transmission (most likely due to the driver's skills and habits? Further investigation required).