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 (GGPLOT2 required); Loading the required data (“mtcars”, provided with R); and Switching ‘am’ to a factor variable (for specific processing only).

Note: All pre-processing is hidden.

Exploratory Data Analysis

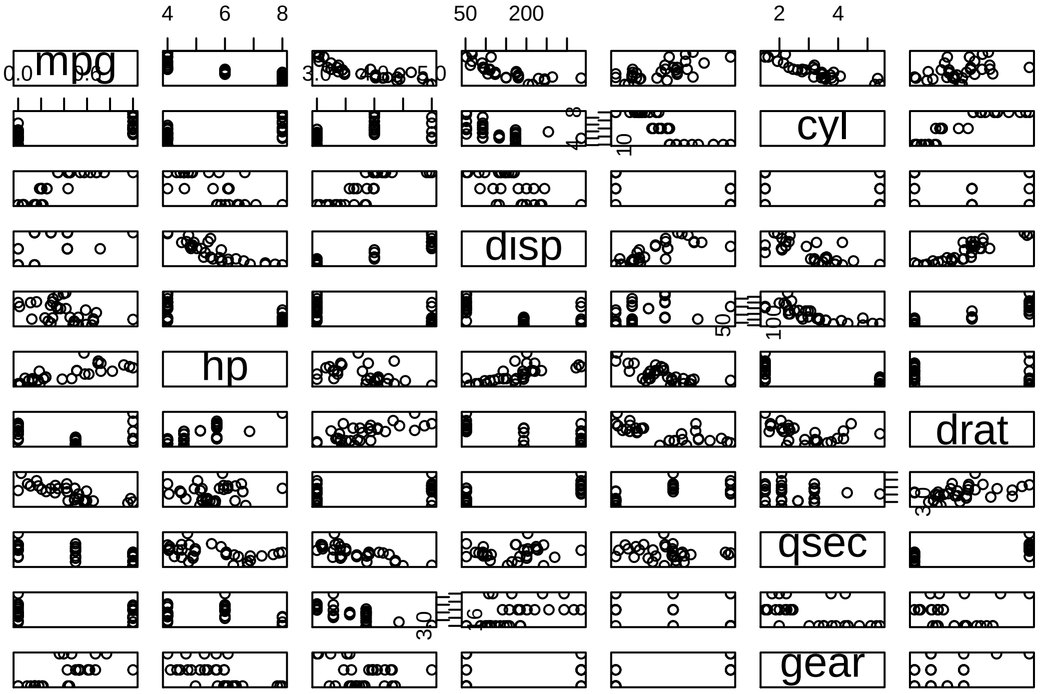
Is transmission (i.e. ‘am’, 0 = automatic and 1 = manual) a good predictor for the ‘mpg’? Using a plot to demonstrate the relation and the linear regression between the variables.

```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’).

```r
verInd <- c(1, 2, 3, 4, 5, 7, 10)
pairs(mtcars, verInd = verInd)
```



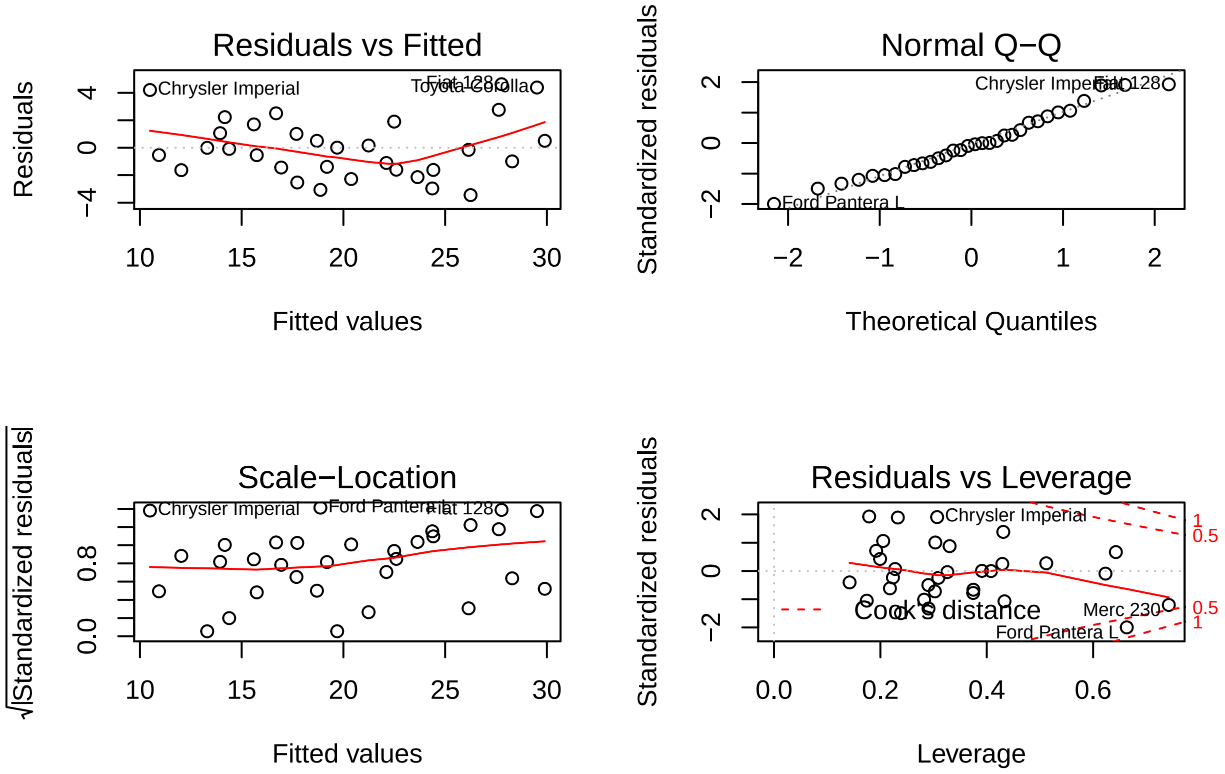
```
summary(lm(mpg ~ ., data = mtcars))
```

```
##
## Call:
## lm(formula = mpg ~ ., data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4506 -1.6044 -0.1196  1.2193  4.6271
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 12.30337   18.71788   0.657   0.5181
## cyl         -0.11144    1.04502  -0.107   0.9161
## disp         0.01334    0.01786   0.747   0.4635
## hp          -0.02148    0.02177  -0.987   0.3350
## drat         0.78711    1.63537   0.481   0.6353
## wt          -3.71530    1.89441  -1.961   0.0633 .
## qsec         0.82104    0.73084   1.123   0.2739
## vs           0.31776    2.10451   0.151   0.8814
## am           2.52023    2.05665   1.225   0.2340
## gear         0.65541    1.49326   0.439   0.6652
## carb        -0.19942    0.82875  -0.241   0.8122
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.65 on 21 degrees of freedom
## Multiple R-squared:  0.869,  Adjusted R-squared:  0.8066
## F-statistic: 13.93 on 10 and 21 DF,  p-value: 3.793e-07
```

```r
summary(lm(mpg ~ . -1, data = mtcars))
```

```
## 
## Call:
## lm(formula = mpg ~ . - 1, data = mtcars)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -3.7721 -1.6249  0.1699  1.1068  4.4666 
## 
## Coefficients:
##      Estimate Std. Error t value Pr(>|t|)  
## cyl   0.35083    0.76292   0.460   0.6501  
## disp  0.01354    0.01762   0.768   0.4504  
## hp   -0.02055    0.02144  -0.958   0.3483  
## drat  1.24158    1.46277   0.849   0.4051  
## wt   -3.82613    1.86238  -2.054   0.0520 .
## qsec  1.19140    0.45942   2.593   0.0166 *
## vs    0.18972    2.06825   0.092   0.9277  
## am    2.83222    1.97513   1.434   0.1656  
## gear  1.05426    1.34669   0.783   0.4421  
## carb -0.26321    0.81236  -0.324   0.7490  
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 2.616 on 22 degrees of freedom
## Multiple R-squared:  0.9893, Adjusted R-squared:  0.9844 
## F-statistic:   203 on 10 and 22 DF,  p-value: < 2.2e-16
```

```r
par(mfrow = c(2, 2))
fit <- lm(mpg ~ ., data = mtcars) # ?influence.measures
plot(fit)
```



```
library(car)
vif(fit)
```

```
##       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
```

No random variable are dropped, hence all of them adds information to the linear model.

Predictions and Correlation

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

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

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

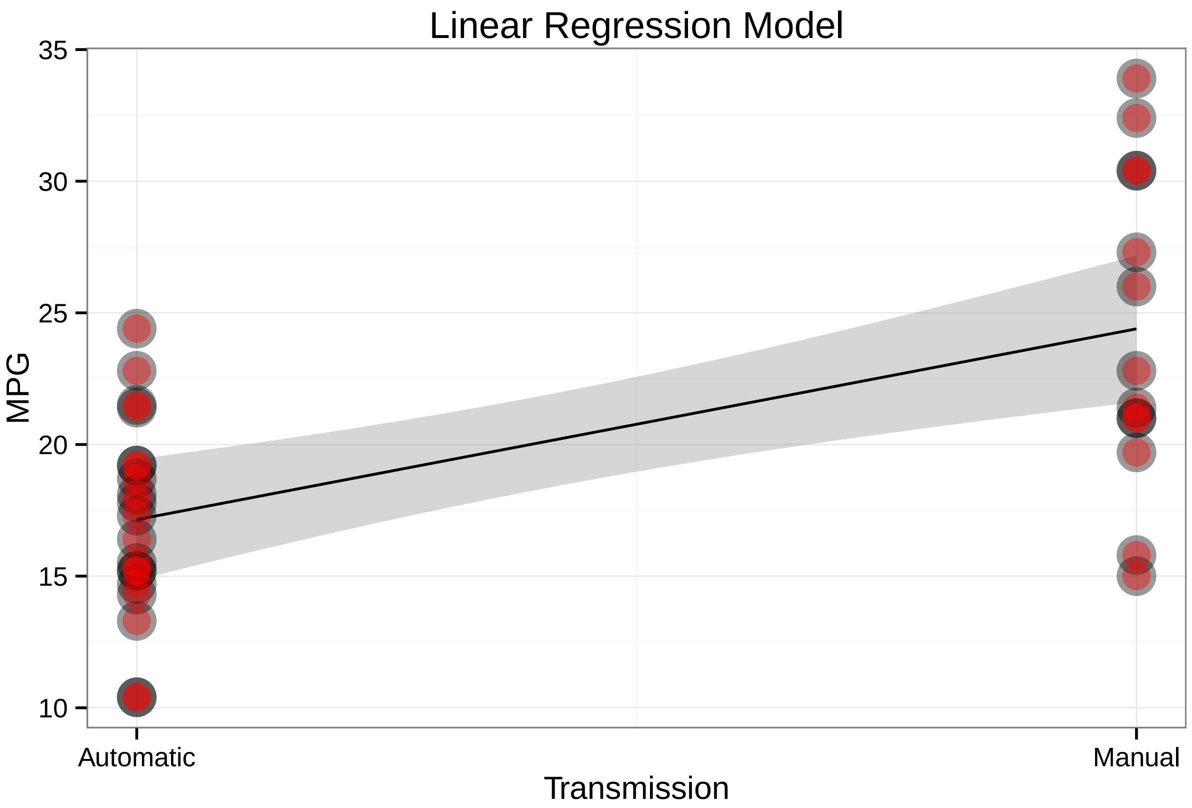
```
## [1] 1.803931
```

```
# linear regression
fit <- lm(mpg ~ am, data = mtcars_f)
summary(fit)
```

```
##
## Call:
## lm(formula = mpg ~ am, data = mtcars_f)
##
## 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 ***
## amManual       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
```

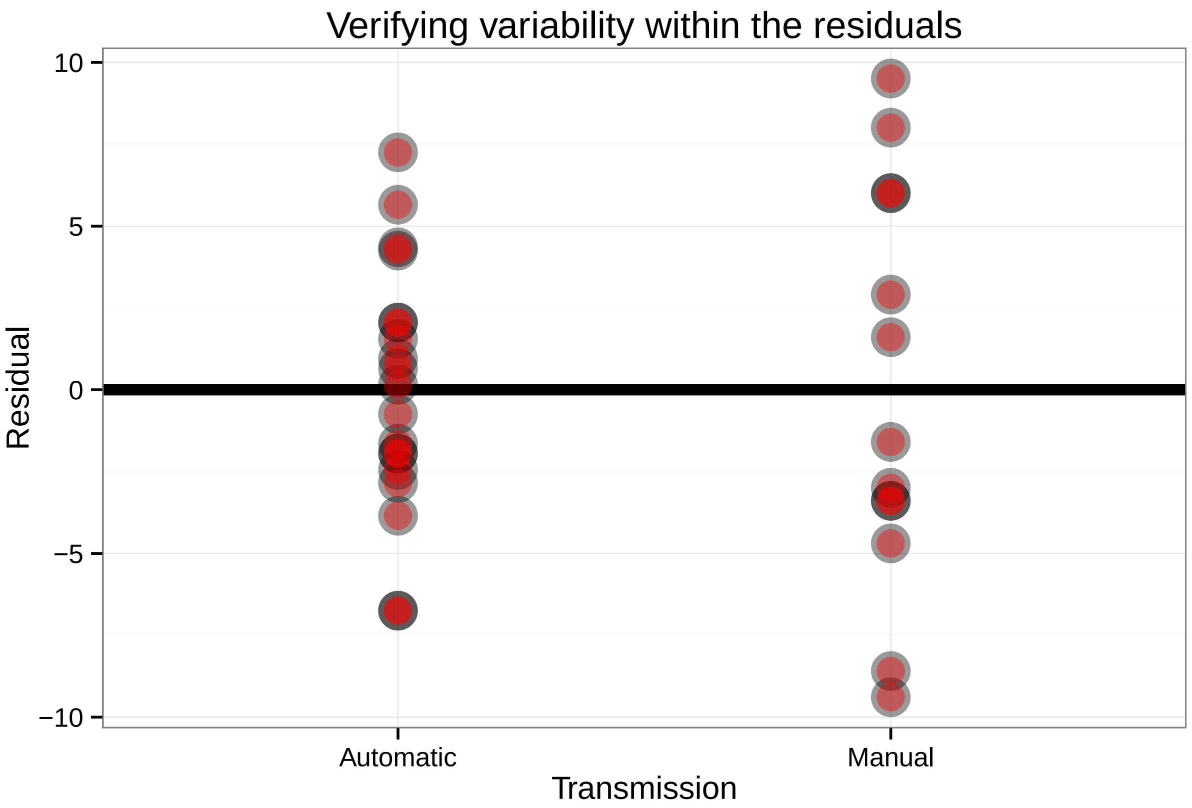
Regression Model…

```r
g = ggplot(data = mtcars, aes(x = am, y = mpg)) +
  geom_smooth(method = "lm", colour = "black") + # linear model regression
  scale_x_continuous(breaks = c(0, 1),
                     name = "Transmission",
                     labels = c("0" = "Automatic", "1" = "Manual")) +
  geom_point(size = 7, colour = "black", alpha = 0.4) + # black contour
  geom_point(size = 5, colour = "red", alpha = 0.4) + # red center
  ylab("MPG") +
  labs(title="Linear Regression Model") +
  theme_bw()
g
```



Studying the residual plot…

```r
g = ggplot(data = mtcars_f, 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
```



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