Regression Models Course Project

Zanin Pavel

March 28, 2016

Link to project on GitHUB
Link to project on RPub

Executive summary

You work for Motor Trend, a magazine about the automobile industry. Looking at a data set of a collection of cars, they are interested in exploring the relationship between a set of variables and miles per gallon (MPG) (outcome). They are particularly interested in the following two questions:
* "Is an automatic or manual transmission better for MPG"
* "Quantify the MPG difference between automatic and manual transmissions"

Analysis

Exploratory analysis

```
library (datasets)
data(mtcars) # Loading data
head(mtcars) # Dataset's head
```

```
##                    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
```

```
dim(mtcars) # Row's numbers and variable's quantity
```

```
## [1] 32 11
```

Let's test hypotesys what automatic and manual transmission are the same on average for MPG?

```
result <- t.test(mtcars$mpg ~ mtcars$am)
result$p.value
```

```
## [1] 0.001373638
```

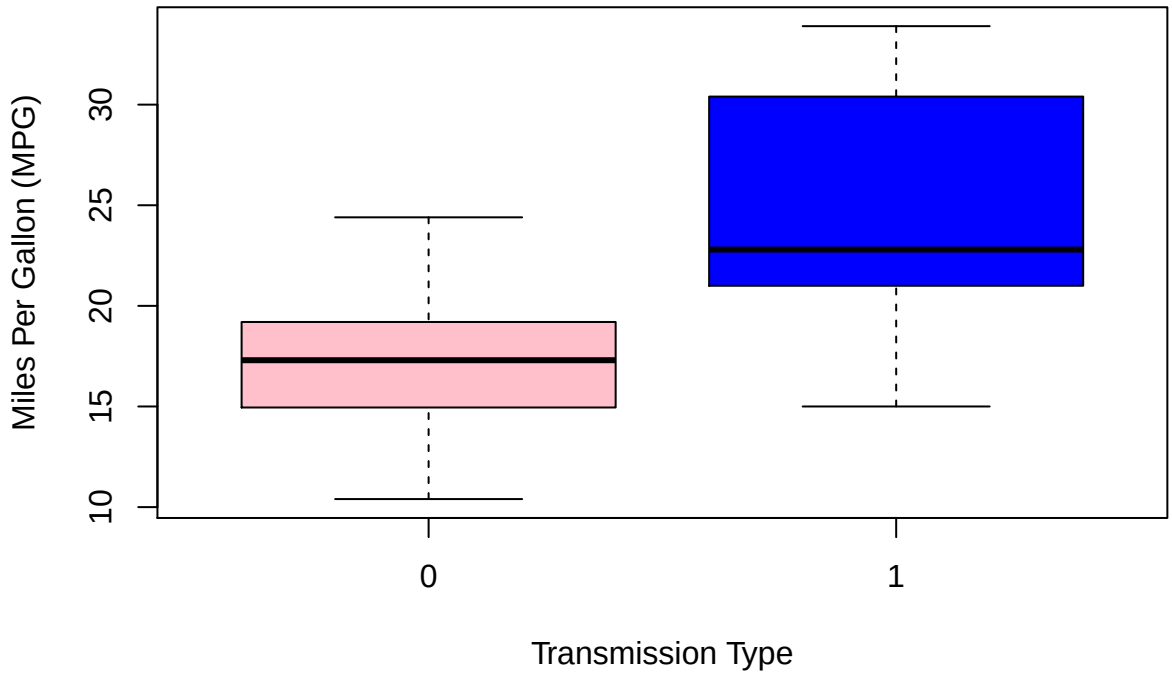
Since the p-value is 0.00137, we reject our null hypothesis. So, the automatic and manual transmissions are from different populations. Let's show difference:

```
result$estimate
```

```
## mean in group 0 mean in group 1
##        17.14737        24.39231
```

```
mtcars$vs <- as.factor(mtcars$vs)
mtcars$am <- as.factor(mtcars$am)

boxplot(mpg ~ am,
        data = mtcars,
        ylab = "Miles Per Gallon (MPG)",
        xlab = "Transmission Type",
        col = (c("pink","blue")))
```



Regression analysis

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

```
##
## Call:
## lm(formula = mpg ~ factor(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 ***
## factor(am)1    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
```

The adjusted R squared value is only 33.8% of the regression variance can be explained by our model. Let's see how will other predictor variables impact.

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

```
##
## 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
cor(mtcars)[1,]
```

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

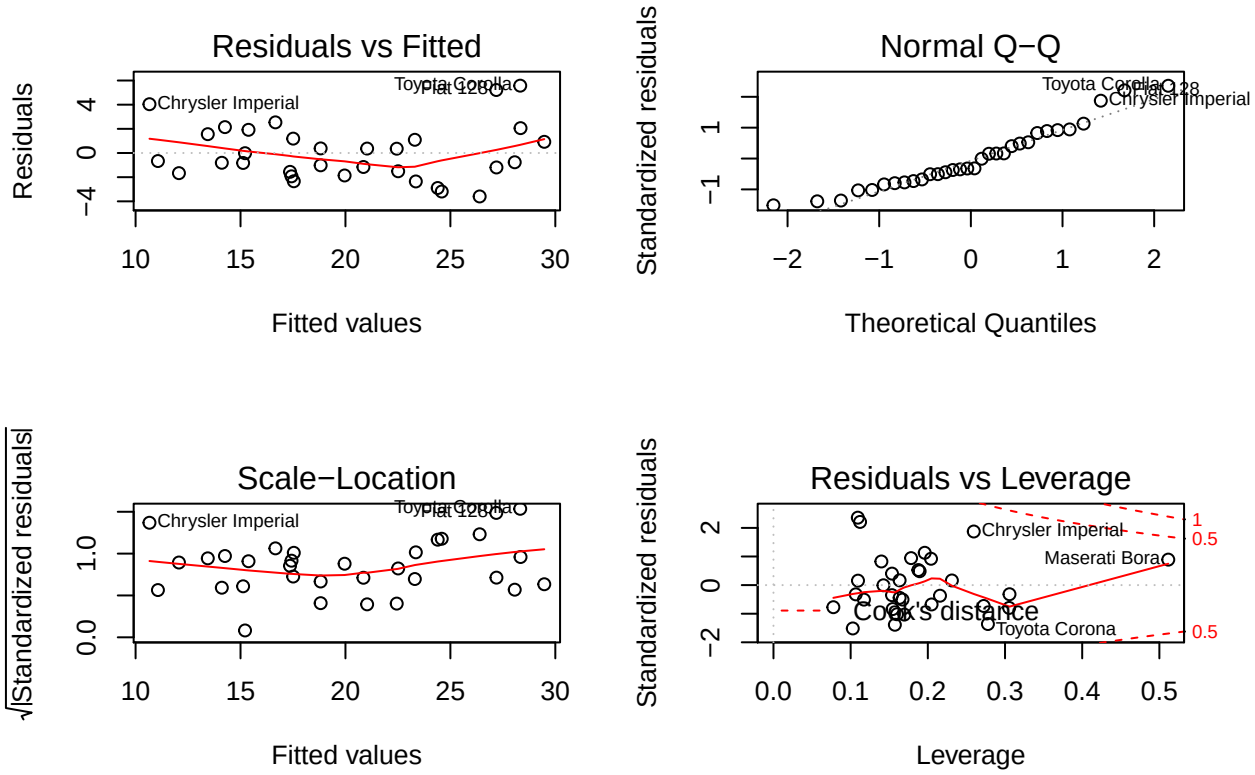
From the output we can see cyl,wt,hp,disp show strong correlations and significance for the model. Hence we choose those variables plus am for a linear model. This gives us the following model below:

```r
fit_MLR_adjusted <- lm(mpg ~ wt+hp+disp+cyl+am, data = mtcars)
summary(fit_MLR_adjusted)
```

```
##
## Call:
## lm(formula = mpg ~ wt + hp + disp + cyl + am, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.5952 -1.5864 -0.7157  1.2821  5.5725
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 38.20280    3.66910  10.412 9.08e-11 ***
## wt          -3.30262    1.13364  -2.913  0.00726 **
## hp          -0.02796    0.01392  -2.008  0.05510 .
## disp         0.01226    0.01171   1.047  0.30472
## cyl         -1.10638    0.67636  -1.636  0.11393
## am           1.55649    1.44054   1.080  0.28984
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.505 on 26 degrees of freedom
## Multiple R-squared:  0.8551, Adjusted R-squared:  0.8273
## F-statistic:  30.7 on 5 and 26 DF,  p-value: 4.029e-10
```

Residual Analysis and Diagnostics

```r
par(mfrow = c(2, 2))
plot(fit_MLR_adjusted)
```



According to the residual plots:
1. The Residuals vs. Fitted plot shows no consistent pattern, supporting the accuracy of the independence assumption.
2. The Normal Q-Q plot indicates that the residuals are normally distributed because the points lie closely to the line.
3. The Scale-Location plot confirms the constant variance assumption, as the points are randomly distributed.
4. The Residuals vs. Leverage argues that no outliers are present, as all values fall well within the 0.5 bands.

Conclusions

Using the final multivariable regression model put together we can see the multiple R squared value is much higher at 0.83, where 83% of the regression variance can be explained by the chosen variables. We can thus conclude that wt, hp, disp and cyl are confounding variables in the relationship between 'am and 'mpg' and that manual transmission cars on average have 1.55 miles per gallon more than automatic cars.