# Regression Models Course Project

MsGret

September 25, 2020

Based on the data set of a collection of cars (**mtcars** data set), we explore the relationship between a set of variables and miles per gallon (MPG) (outcome) and answer two questions:

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

## Exploratory Analyses

Load the data and perform some basic exploratory data analyses

```
library (datasets)
data(mtcars)
dim(mtcars)
```

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

```
str(mtcars)
```

```
## 'data.frame':	32 obs. of  11 variables:
##  $ mpg : num  21 21 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 ...
##  $ cyl : num  6 6 4 6 8 6 8 4 4 6 ...
##  $ disp: num  160 160 108 258 360 ...
##  $ hp  : num  110 110 93 110 175 105 245 62 95 123 ...
##  $ drat: num  3.9 3.9 3.85 3.08 3.15 2.76 3.21 3.69 3.92 3.92 ...
##  $ wt  : num  2.62 2.88 2.32 3.21 3.44 ...
##  $ qsec: num  16.5 17 18.6 19.4 17 ...
##  $ vs  : num  0 0 1 1 0 1 0 1 1 1 ...
##  $ am  : num  1 1 1 0 0 0 0 0 0 0 ...
##  $ gear: num  4 4 4 3 3 3 3 4 4 4 ...
##  $ carb: num  4 4 1 1 2 1 4 2 2 4 ...
```

Data set consists consists of 11 variables and 32 observation for each variable.

Look at relationship between transmission type (`am` as factor variable (0 - automatic, 1 - manual) and miles per gallon (`mpg`) (**Appendix A**).

```
mtcars$am <- factor(mtcars$am, labels = c("automatic", "manual"))
```

Based on boxplot in **Appendix A** we can suppose that there is a significant difference in MPG for different transmission type.

## Statistical Inference

Test our hypothesis: Null hypothesis is "the MPG means for different transmission type is equal" or "true difference in MPG means for different transmission type is equal to 0".

```r
t.test(mpg ~ am, data = mtcars)
```

```
##
##  Welch Two Sample t-test
##
## data:  mpg by 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
```

We can reject null hypothesis that the difference in MPG means for different transmission type is equal to 0 (p-value = 0.001374).

## Regression Analysis

So MPG depends on transmission type, but define how other variables affect on MPG. Build multivariable regression model (results in **Appendix B**):

```r
mtcars$cyl <- factor(mtcars$cyl)
mtcars$vs <- factor(mtcars$vs, labels = c("V", "S"))
mtcars$gear <- factor(mtcars$gear)
mtcars$carb <- factor(mtcars$carb)

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

So none of the coefficients have a p-value less than 0.05 (statistically significant). Find better model (based on removing variables from the model and evaluating the AIC):

```r
AICModel <- step(fullModel, direction = "both")
summary(AICModel)
anova(AICModel, fullModel)
```

Comparing the `AICModel` with the `fullModel` we see that removing other predictors has not significantly affected the explanatory ability of the model.

The `AICModel` explains about 87% of the variance in MPG (R-squared is 0.8659). The coefficients conclude that increasing the number of cylinders from 4 to 6 with decrease the MPG by 3.031, but from 4 to 8 with decrease the MPG by 2.164. One additional horsepower is decreases MPG by 0.0321. Weight decreases the MPG by 2.497 for each 1000 lbs increase. A Manual transmission improves the MPG by 1.809.

## Residual Analysis

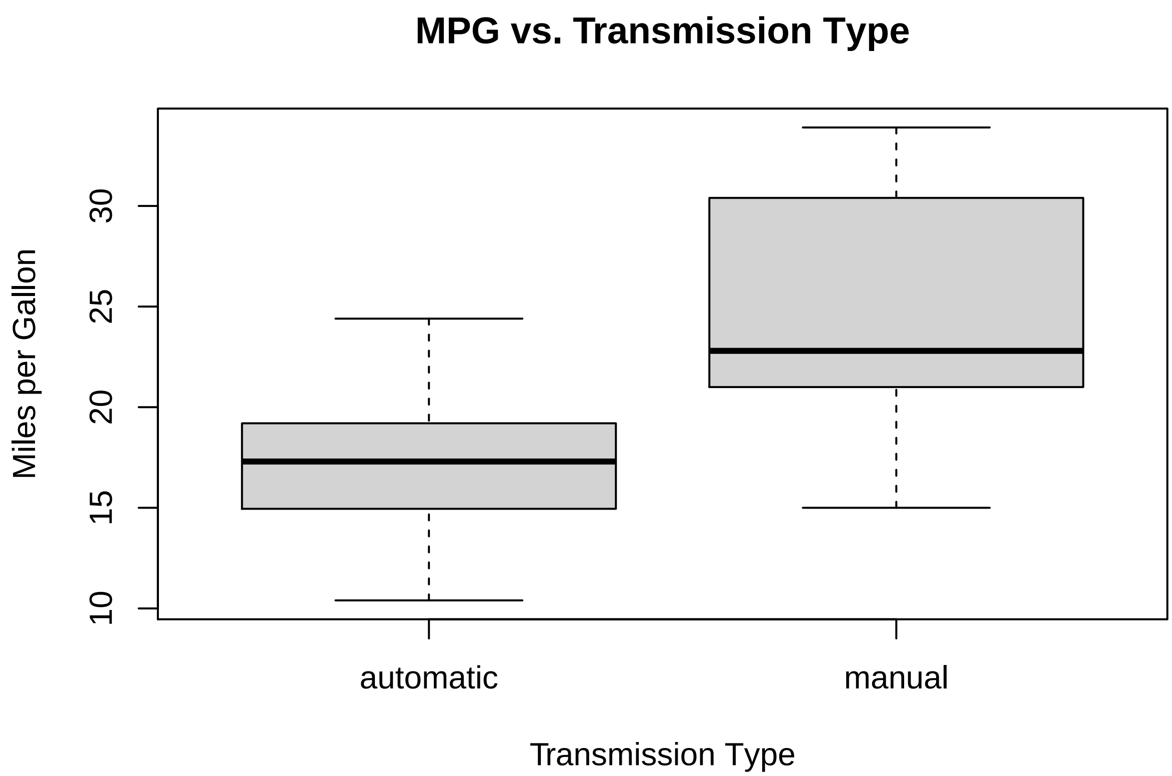
Based on residuals plots (**Appendix C**) we can conclude:

- the `Residuals vs Fitted` plot doesn't show pattern and confirms that residuals are independent;
- the `Normal Q-Q` plot confirms that the residuals are normally distributed (with some deviate from normality at the tails);
- the `Scale-Location` confirms the constant variance assumption;
- the `Residuals vs Leverage` confirms that there are no outliers (all values fall within the 0.5 bands).

### Conclusion

- There is a significant difference in MPG for different transmission type (MPG mean for manual type more automatic type at 7.24).
- Based on `AICModel` we can conclude that number of cylinders, weight and horsepower are more statistically significant then transmission type for determining MPG.

### Appendix A




### Appendix B

```r
summary(fullModel)
```

```
##
## Call:
## lm(formula = mpg ~ ., data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.5087 -1.3584 -0.0948  0.7745  4.6251
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 23.87913   20.06582   1.190   0.2525
## cyl6        -2.64870    3.04089  -0.871   0.3975
## cyl8        -0.33616    7.15954  -0.047   0.9632
```

```
## disp         0.03555    0.03190   1.114   0.2827
## hp          -0.07051    0.03943  -1.788   0.0939 .
## drat         1.18283    2.48348   0.476   0.6407
## wt          -4.52978    2.53875  -1.784   0.0946 .
## qsec         0.36784    0.93540   0.393   0.6997
## vsS          1.93085    2.87126   0.672   0.5115
## ammanual     1.21212    3.21355   0.377   0.7113
## gear4        1.11435    3.79952   0.293   0.7733
## gear5        2.52840    3.73636   0.677   0.5089
## carb2       -0.97935    2.31797  -0.423   0.6787
## carb3        2.99964    4.29355   0.699   0.4955
## carb4        1.09142    4.44962   0.245   0.8096
## carb6        4.47757    6.38406   0.701   0.4938
## carb8        7.25041    8.36057   0.867   0.3995
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.833 on 15 degrees of freedom
## Multiple R-squared:  0.8931, Adjusted R-squared:  0.779
## F-statistic:  7.83 on 16 and 15 DF,  p-value: 0.000124
```

```
summary(AICModel)
```

```
##
## 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
```

```
anova(AICModel, fullModel)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ cyl + hp + wt + am
## Model 2: mpg ~ cyl + disp + hp + drat + wt + qsec + vs + am + gear + carb
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     26 151.03
## 2     15 120.40 11    30.623 0.3468 0.9588
```

## Appendix C

Residuals plots for `AICModel`:

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