Regression Models Course Project

MsGret

September 28, 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
```

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".

```
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:

```
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)$r.squared
```

```
## [1] 0.8930749
```

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):

```
AICModel <- step(fullModel, direction = "both") # hide results
```

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

```
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
```

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

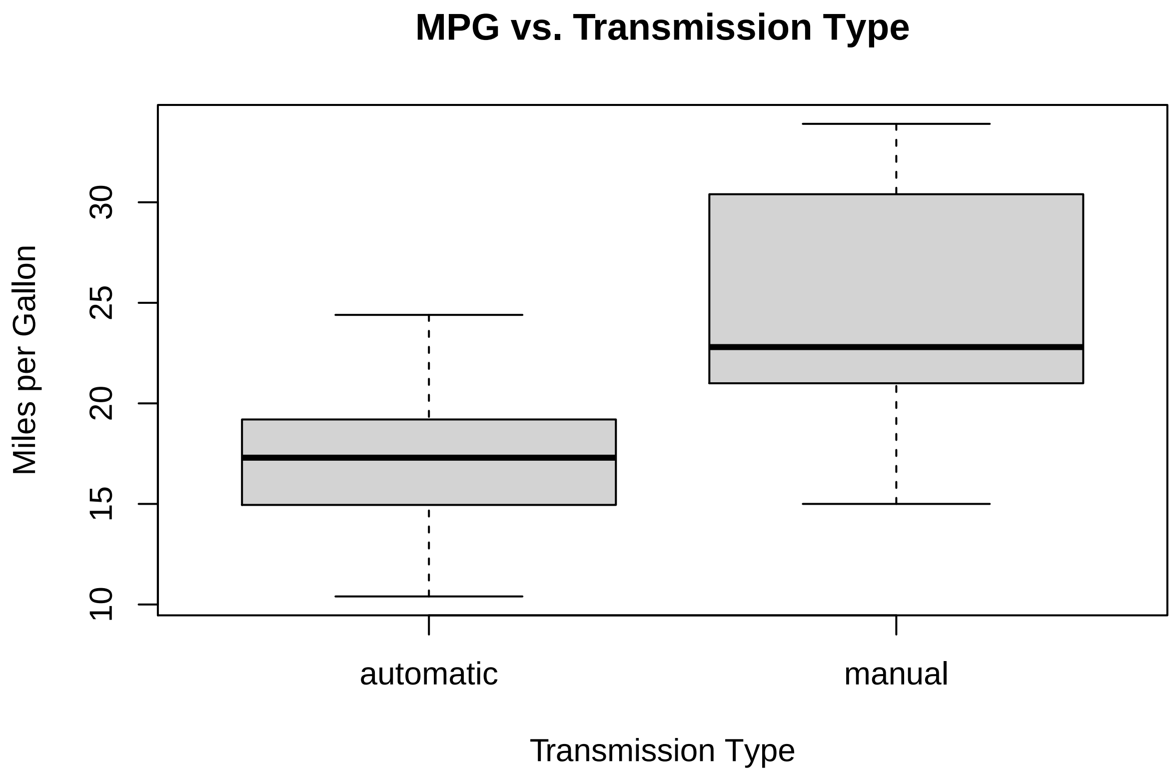
Based on residuals plots (**Appendix B**) we can conclude:

- 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.

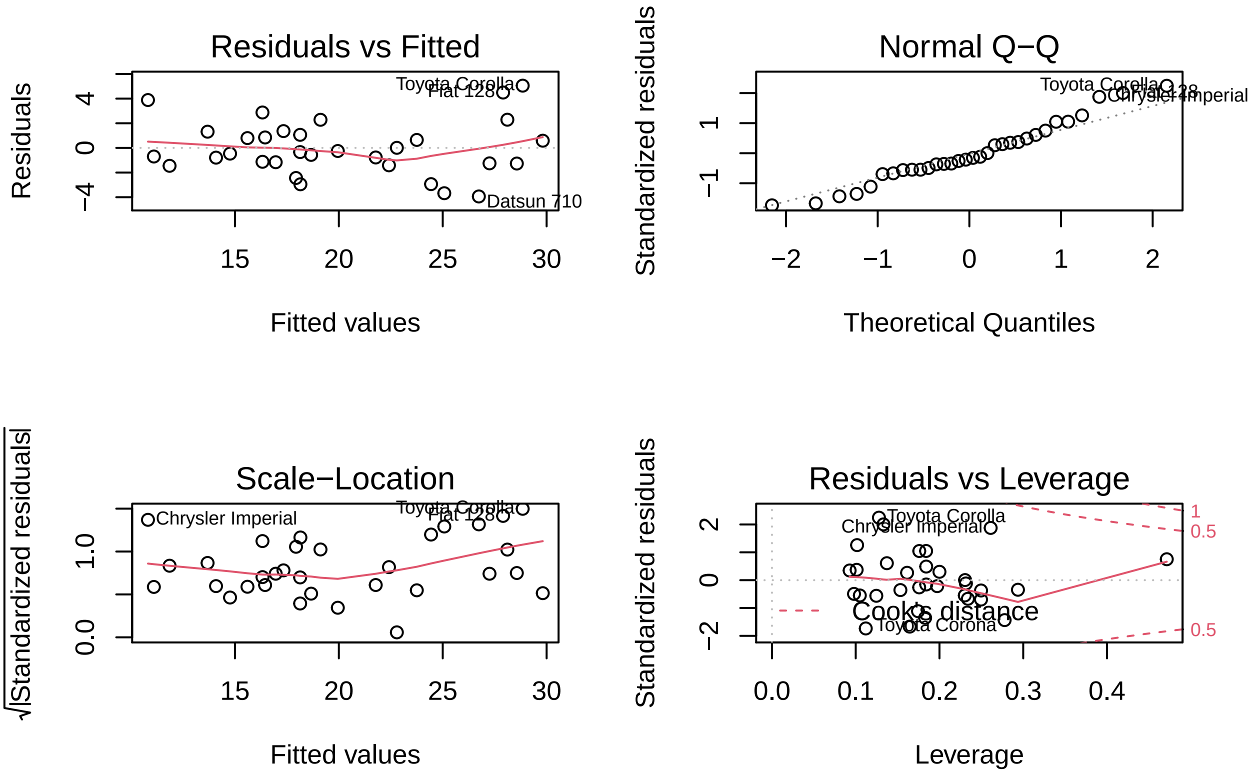
Appendix A



Appendix B

Residuals plots for `AICModel`:

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


Residuals vs Fitted
Residuals
Fitted values
Toyota Corolla
Fiat 128
Datsun 710
Normal Q−Q
Standardized residuals
Theoretical Quantiles
Chrysler Imperial
Scale−Location
√|Standardized residuals|
Residuals vs Leverage
Leverage
Cook's distance
Toyota Corona