Regression models course project

Georgy Makarov

April 29, 2020

Executive summary

This report explores *Motor Trend* magazine's data set of a collection of cars in order to find out, which is better for fuel consumption - manual transmission or automatic transmission. Cars with manual transmission on average have 7.2 higher *MPG* than cars with automatic. Transmission type explains 36% of variance in *MPG*. Other variables like number of cylinders, power and weight are more important to *MPG* than the transmission. The real difference in transmission types is only 1.8 MPG.

Exploratory data analysis

The data set contains 32 observations of 11 variables. All variables are numeric. There are five variables, which are better to consider as factor variables: *cyl, vs, am, gear, carb*. The distribution of cars by *MPG* is right skewed. The majority of cars has fuel consumption between 15 and 20 MPG. The cars are mostly represented by 8-cylinder cars with automatic transmission. A car transmission usually has 3 gears. Supporting information is in the appendix in the *Figure 1*.

```
data(mtcars)
mtcars$cyl <- as.factor(mtcars$cyl)
mtcars$vs <- as.factor(mtcars$vs)
mtcars$am <- as.factor(mtcars$am)
mtcars$gear <- as.factor(mtcars$gear)
mtcars$carb <- as.factor(mtcars$carb)
summary(mtcars$mpg)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   10.40   15.43   19.20   20.09   22.80   33.90
```

Cars with manual transmission consume less fuel. The median of *MPG* for manual transmission is around 23, while for automatic transmission it is around 17. The reference is in the *Figure 2*.

Regression models

H_0: automatic and manual transimissions *MPG* are the same. We test the H_0 with a t-test. *P-value* of the test is 0.001, confidence interval does not contain zero - there is enough evidence to reject H_0.

```
at <- mtcars[mtcars$am == 0,]
mt <- mtcars[mtcars$am == 1,]
t.test(at$mpg, mt$mpg)$p.value
```

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

```
t.test(at$mpg, mt$mpg)$conf
```

```
## [1] -11.280194  -3.209684
## attr(,"conf.level")
## [1] 0.95
```

Linear regression model with *am* as a predictor quantifies the difference between transmissions. *P-values* are significant - average *MPG* for automatic transmission is 17.1, for manual - 7.2 higher than the automatic. The model explains 36% of the variance in *MPG* as the R^2 is 0.36. We need more variables to explain the variance.

```r
fit1 <- lm(mpg ~ am, data = mtcars)
summary(fit1)$coeff
```

```
##              Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 17.147368   1.124603 15.247492 1.133983e-15
## am1          7.244939   1.764422  4.106127 2.850207e-04
```

```r
summary(fit1)$r.squared
```

```
## [1] 0.3597989
```

One way to choose top predictors is using the *step* function. Top predictors are *cyl, hp, wt, am*.

```r
fit_full <- lm(mpg ~ ., data = mtcars)
step_fit <- step(fit_full, trace = 0)
summary(step_fit)$coeff
```

```
##                Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 33.70832390 2.60488618 12.940421 7.733392e-13
## cyl6        -3.03134449 1.40728351 -2.154040 4.068272e-02
## cyl8        -2.16367532 2.28425172 -0.947214 3.522509e-01
## hp          -0.03210943 0.01369257 -2.345025 2.693461e-02
## wt          -2.49682942 0.88558779 -2.819404 9.081408e-03
## am1          1.80921138 1.39630450  1.295714 2.064597e-01
```

P-values for **multivariate regression** are statistically significant. The R^2 for this model is 0.866. This model shows that manual transmission improves fuel consumption by 1.81 MPG, while increase in number of cylinders from 4 to 6 causes 3.0 MPG loss. Engines with 8 cylinders eat another 2.2 MPG. Every additional *hp* results in decreasing the MPG by 0.03. Every 1000 lbs of weight decrease the MPG by 2.5 US gallons.

```r
fit4 <- lm(mpg ~ am + cyl + hp + wt, data = mtcars)
summary(fit4)$r.squared
```

```
## [1] 0.8658799
```

The model with four predictors fits better than the model with transmission type only. The *anova* test shows the significance of *p-value*.

```r
anova(fit1, fit4)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ am + cyl + hp + wt
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     30 720.90
## 2     26 151.03  4    569.87 24.527 1.688e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Residuals and diagnostics

Residuals check and model diagnostics are in the *Figure 3*. *Residual vs Fitted* plot shows that the residuals are independent. *Normal Q-Q* plot shows normal distribution of residuals. Residuals on *Scale-Location* plot are randomly distributed. There are no outliers as the dots on *Residuals vs Leverage* plot are within [-0.5;+0.5] interval.

Appendix

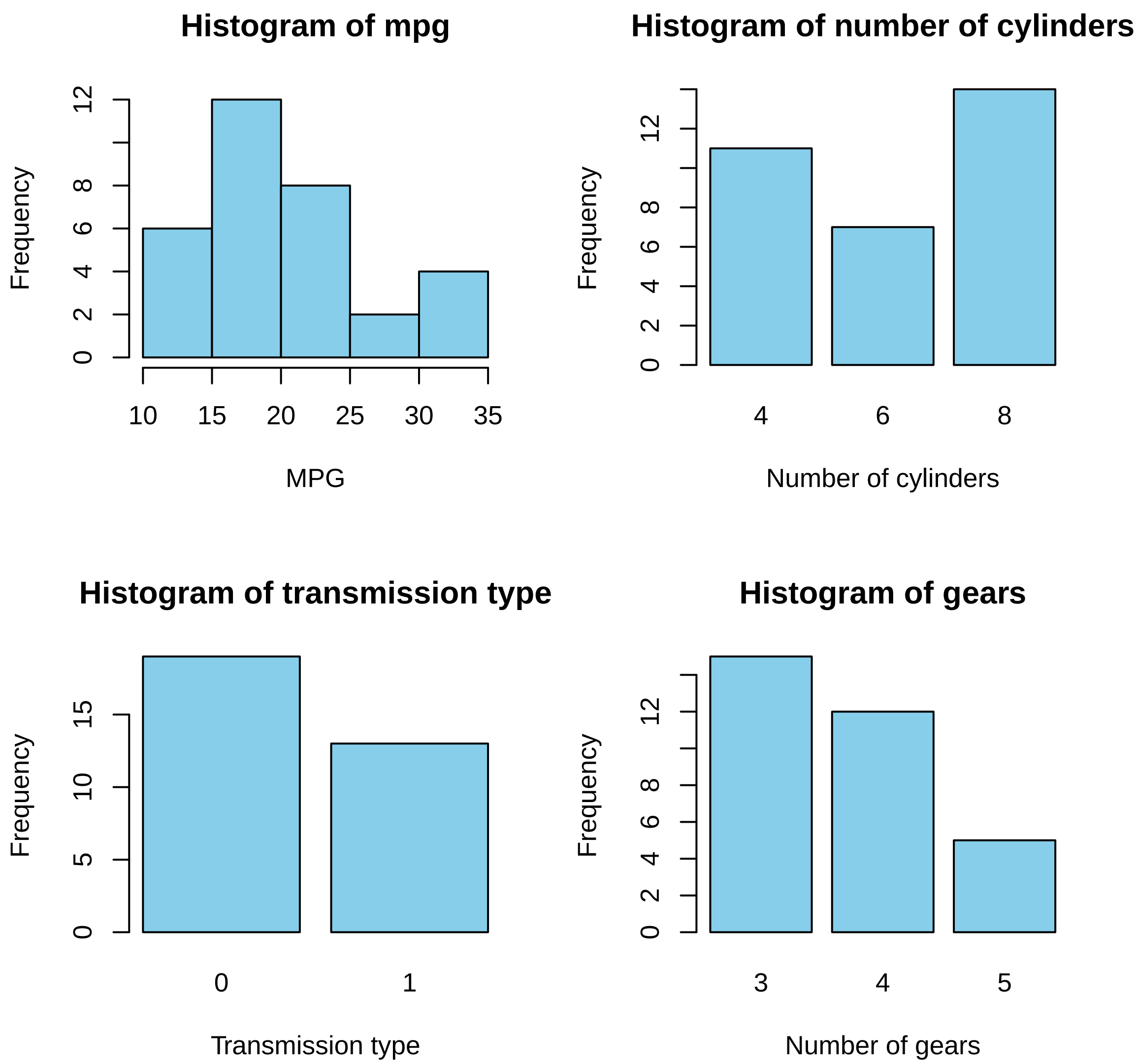

Figure 1: Distribution

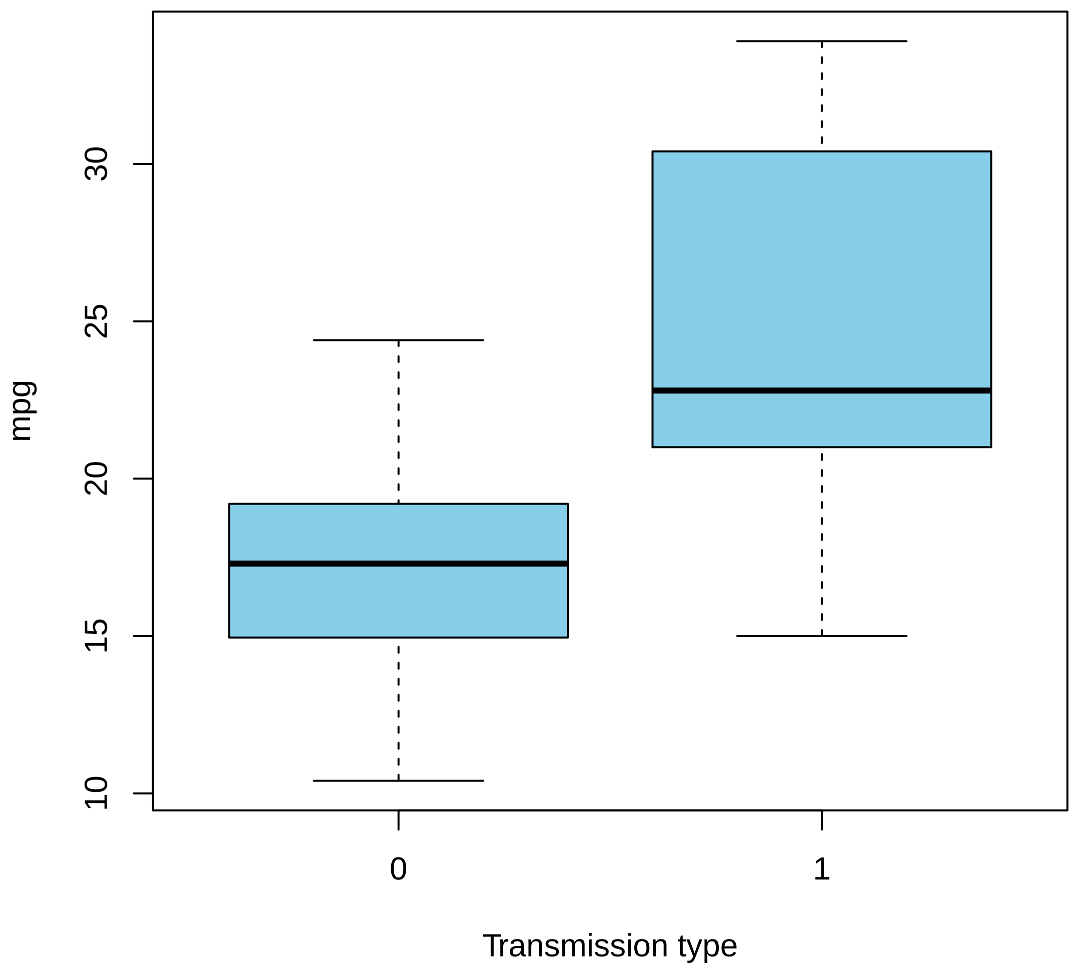

Figure 2: Transmission type

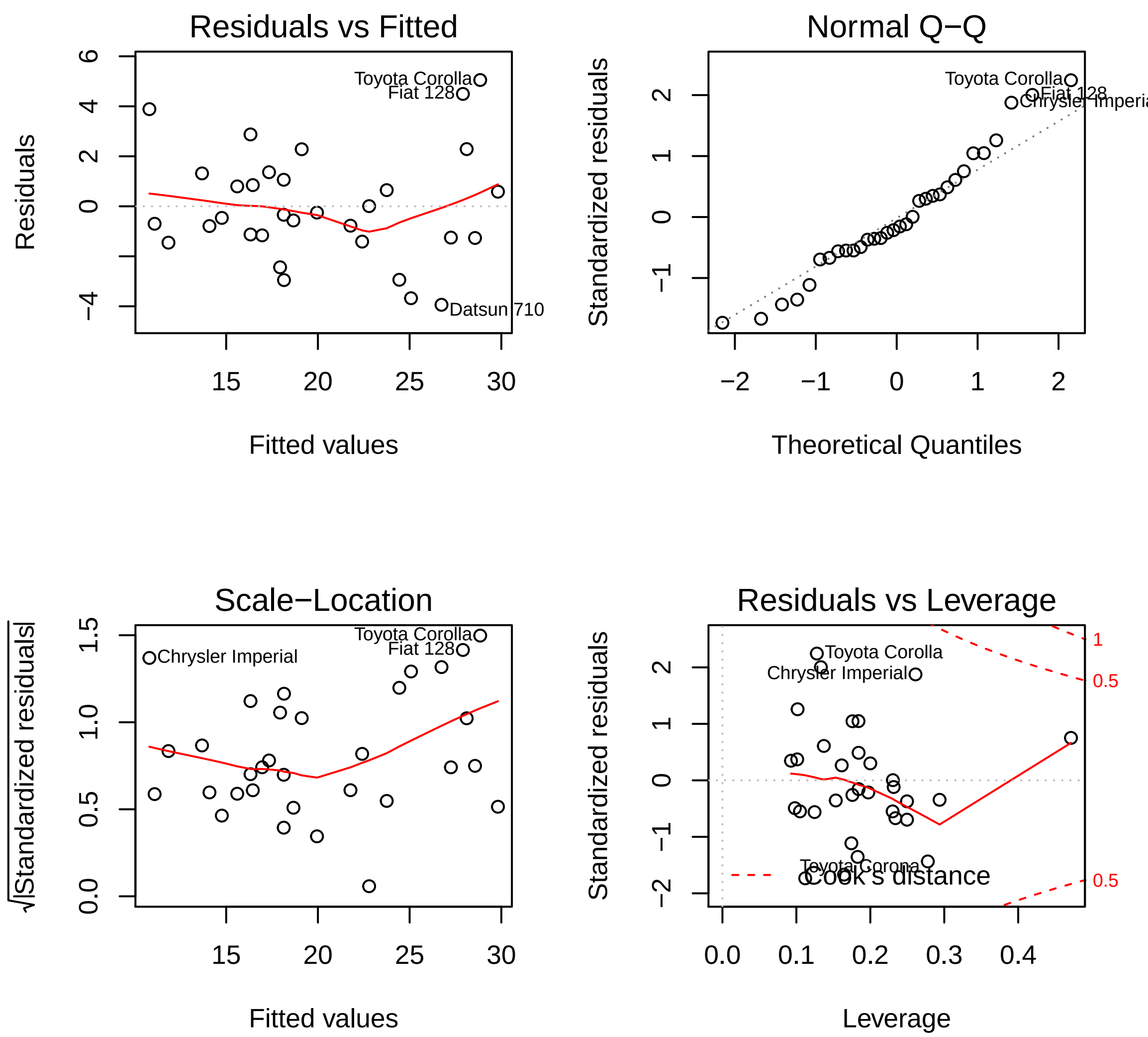

Figure 3: Residuals