Regression Models Course Project

Oleksandr Myronov

12 06 2021

SUMMARY

In this study we explore **mtcars** dataset, extracted from 1974 Motor Trend US magazine. Description for variables can be found **here**. We explore relationship between automated or manual transmission and car mpg. Our linear model lm(mpg~wt*am+qsec*am-am) has significant p-value 8.42*10^-13. Due to this model, answer for the question "Is an automatic or manual transmission better for MPG?": **It depends**. Estimated manual transmission effect can be approximately calculated by formula:

$$MPG_{difference} = 0.61 * qsec - 2.92 * wt$$

- where *MPGdifference* is expected difference in miles per gallon, *qsec* is 1/4 mile time in seconds and *wt* is car weight in 1000 lbs. With positive difference, manual transmission cars have higher mpg. *qsec* coefficient has 95% uncertainty bounds 0.30 .. 0.92 and *wt* coefficient has 95% uncertainty bounds -4.85 .. -1.00.

Due to our linear model, manual transmission is better for light-weighted cars and for heavy-weighted cars model tells that automated transmission is better. In mtcars dataset, sorted by car weight descending, TOP-8 heavy cars have automated transmission, so we have no enough data in this segment. We can consider, that for heavy cars manufacturers prefer automated transmission and this fact supports our hypothesis.

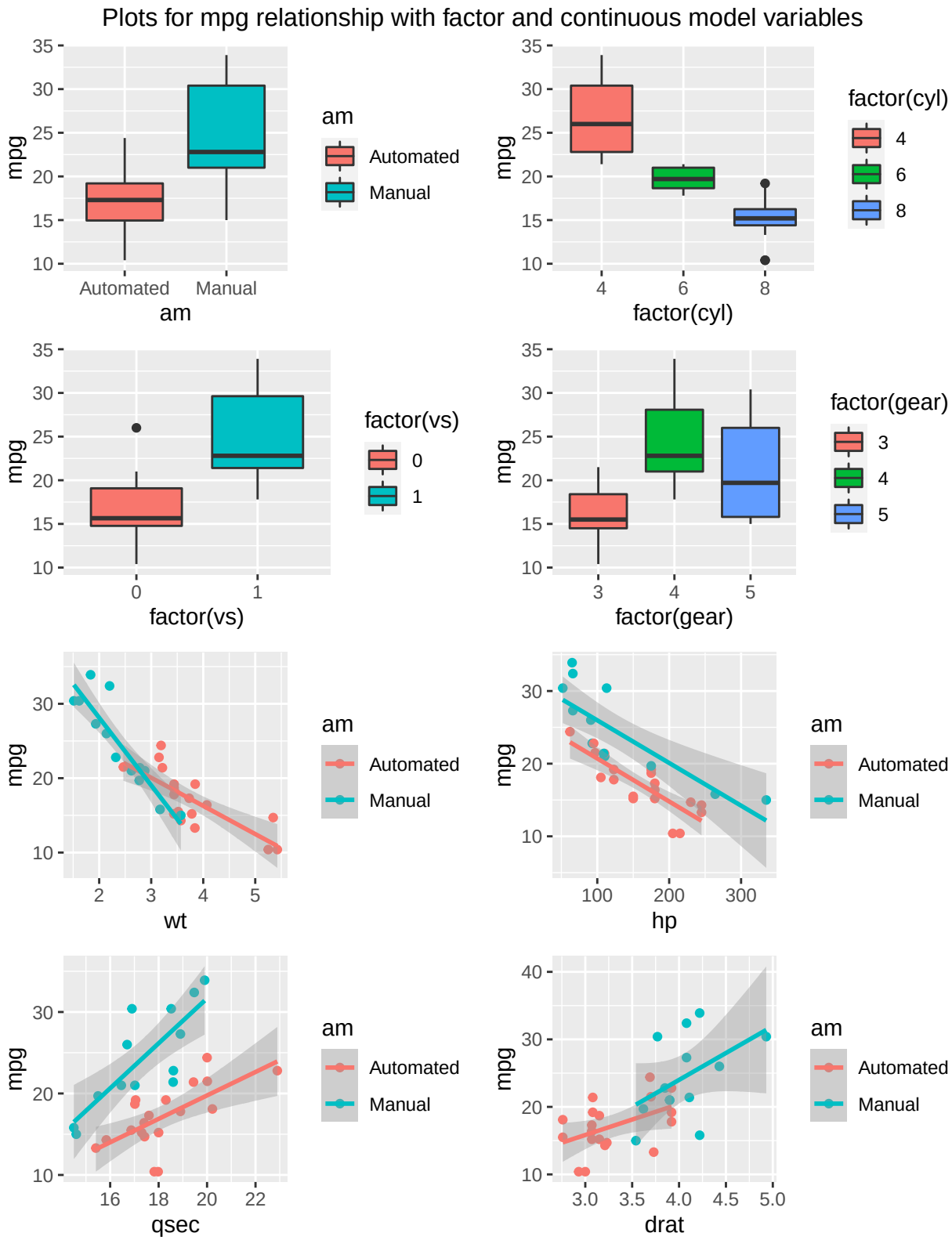
DATA ANALYSIS

Loading and exploring data

Data frame **mtcars** contains 32 car model observations on 11 design aspect variables. We would not use data scale normalization, and SVD for interpretability. In this study we assume significance level alpha=0.05.

First, we perform some exploratory analysis, using boxplots for factor variables and scatterplots for continuous variables.

```
library(ggplot2)   #Loading libraries
library(ggpubr)
data(mtcars)       #Loading data
mtcars$am<-factor(mtcars$am, labels=c("Automated", "Manual")) #Factorizing am
g1<-ggplot(mtcars, aes(am, mpg, fill=am))+geom_boxplot()
g2<-ggplot(mtcars,aes(factor(cyl), mpg, fill=factor(cyl)))+geom_boxplot()
g3<-ggplot(mtcars,aes(factor(vs), mpg, fill=factor(vs)))+geom_boxplot()
g4<-ggplot(mtcars,aes(factor(gear), mpg, fill=factor(gear)))+geom_boxplot()
g5<-ggplot(mtcars, aes(wt, mpg, col=am))+geom_point()+geom_smooth(method=lm)
g6<-ggplot(mtcars, aes(hp, mpg, col=am))+geom_point()+geom_smooth(method=lm)
g7<-ggplot(mtcars, aes(qsec, mpg, col=am))+geom_point()+geom_smooth(method=lm)
g8<-ggplot(mtcars, aes(drat, mpg, col=am))+geom_point()+geom_smooth(method=lm)
annotate_figure(ggarrange(g1,g2,g3,g4,g5,g6,g7,g8, ncol=2, nrow=4),
    top=text_grob("Plots for mpg relationship with factor and continuous model variables"))
```



We can see a lot of variability in data. Manual transmission seemed to be more effective, mean mpg of all cars with manual transmission is higher than mean of automated, but this effect may be caused by other factors. We can see straight influence of weight and hp on mpg, other effects as axle ratio, number of forward gears and 1/4 mile time are more distributed.

Fitting models and model selection

First, we fit "full linear model" with all variables linearly included:

```
{fitall<-lm(mpg~., mtcars)
print(summary(fitall)$coef)           #printing model coefficients
print(summary(fitall)$sigma)}         #printing model residual variation
```

```
##                Estimate  Std. Error    t value   Pr(>|t|)
## (Intercept) 12.30337416 18.71788443  0.6573058 0.51812440
## cyl         -0.11144048  1.04502336 -0.1066392 0.91608738
## disp         0.01333524  0.01785750  0.7467585 0.46348865
## hp          -0.02148212  0.02176858 -0.9868407 0.33495531
## drat         0.78711097  1.63537307  0.4813036 0.63527790
## wt          -3.71530393  1.89441430 -1.9611887 0.06325215
## qsec         0.82104075  0.73084480  1.1234133 0.27394127
## vs           0.31776281  2.10450861  0.1509915 0.88142347
## am           2.52022689  2.05665055  1.2254035 0.23398971
## gear         0.65541302  1.49325996  0.4389142 0.66520643
## carb        -0.19941925  0.82875250 -0.2406258 0.81217871
## [1] 2.650197
```

This model is not simply interpretable and has a lot of high p-values. Some of the variables may be highly correlated and can affect linear model. All model coefficients have large p-values, so we have a lot of random noise. We should reduce number of independent variables to achieve reasonable significance level.

Looking at "full linear model" summary, we can see, that number of carbeurators **carb**, engine type **vs**, number of forward gears **gear** and rear axle ratio **drat** have large p-values and also there are no clear logical relationship to mpg. Number of cylinders closely relates to horsepower and displacement and also has large p-values, so we can throw away **cyl** and **disp**. What is left (except transmission type) - weight **wt**, horsepower **hp** and 1/4 mile time **qsec**. Logically, this three variables are also dependent, so we need just two of them. Obviously, car weight influences mpg. Horsepower relates to weight, because heavy cars should have more horsepower. **qsec** depends on both horsepower and weight, but also represents some additional information, such as motor dynamics, car air resistance, inertia of rotating wheels and details, so it should be more informative. Let's look at correlation between **wt** and other variables:

```
cor(mtcars[, c("wt", "hp", "disp", "vs", "qsec", "cyl")], method = "pearson")[1,]
```

```
##         wt         hp       disp         vs       qsec        cyl
##  1.0000000  0.6587479  0.8879799 -0.5549157 -0.1747159  0.7824958
```

qsec is less correlated to weight, so it brings more additional information. Next, we fit "simple model" with three variables: car weight **wt**, 1/4 mile time **qsec** and subject of this study - transmission type **am**.

```
{fit2<-lm(mpg~wt+qsec+am, mtcars)
print(summary(fit2)$coef)               #printing model coefficients
print(summary(fit2)$sigma)}             #printing model residual variation
```

```
##              Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)  9.617781  6.9595930  1.381946 1.779152e-01
## wt          -3.916504  0.7112016 -5.506882 6.952711e-06
## qsec         1.225886  0.2886696  4.246676 2.161737e-04
## am           2.935837  1.4109045  2.080819 4.671551e-02
## [1] 2.458846
```

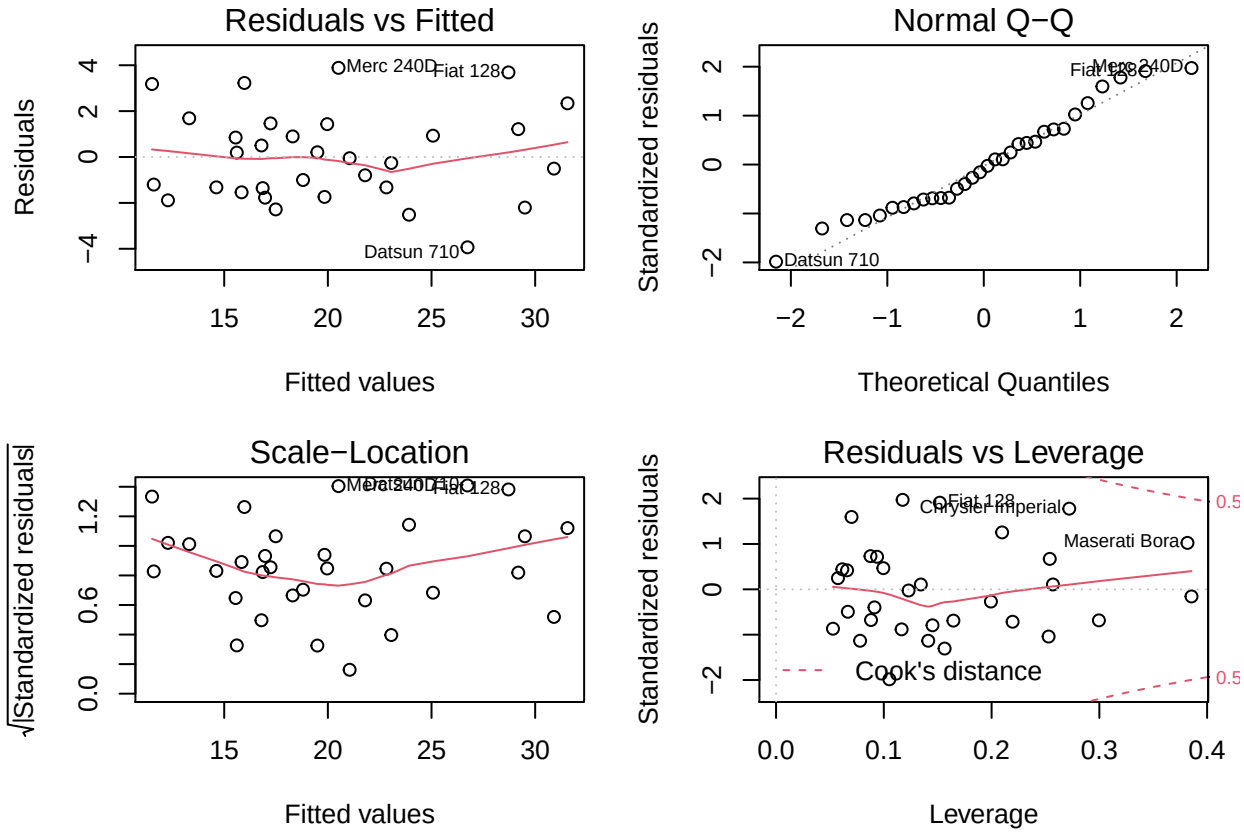
In "simple" model we get most significant p-values, interception is non-significant. Residual error is lower and this model fits data better, than "full linear model". Due to this model, manual transmission increases mpg by 2.9. Assume, that manual transmission doesn't produce mpg by itself and just modifies effect of **wt** and **qsec**. In next "advanced simple model" we include additional **am** product terms and exclude **am**.

```
{fit3<-lm(mpg~wt*am+qsec*am-am, mtcars)
print(summary(fit3)$coef)                #printing model coefficients
print(summary(fit3)$sigma)}              #printing model residual variation
```

```
##               Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 13.9692069  5.7756116  2.418654 2.259367e-02
## wt          -3.1758862  0.6362299 -4.991727 3.114029e-05
## qsec         0.8337859  0.2601709  3.204762 3.458031e-03
## wt:am       -2.9233073  0.9369372 -3.120068 4.271985e-03
## am:qsec      0.6125898  0.1518579  4.033967 4.045127e-04
## [1] 2.096725
```

"Advanced simple model" fits data better and has lower residual variation, all coefficient p-values are significant. This model satisfies our initial significance requirements, and it seems to be more reasonable. Next, we perform basic residual testing for this "advanced simple model".

```
par(mfrow=c(2,2), mar=c(5,4.5,1.5,1))
plot(fit3)
```

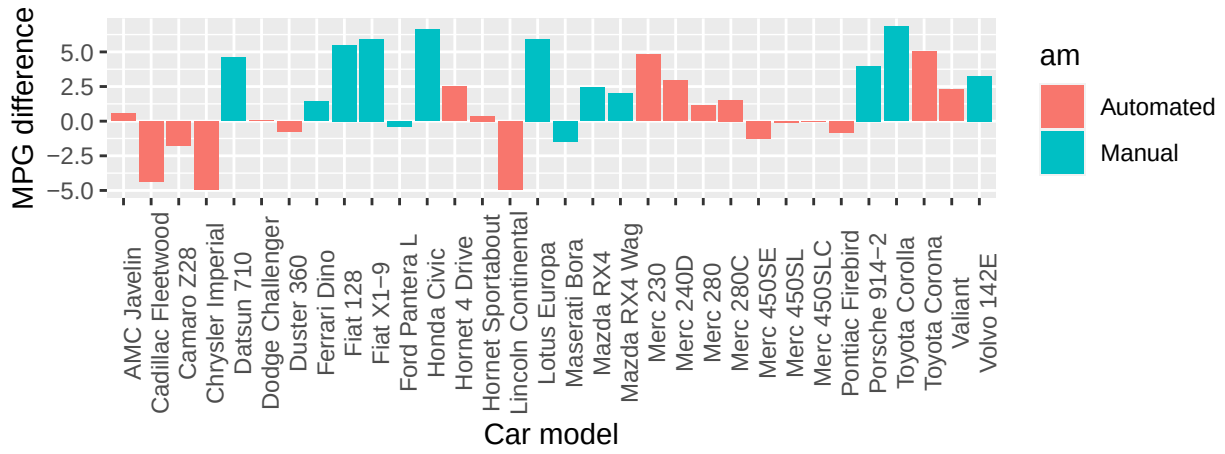


Residual variation of the model tends to normal distribution, and spread of residuals is close to homogeneous. On leverage plot we can see some influential data points, but there are no points out of dashed red line. "Advanced simple model" passes our residuals check.

Model Interpretation

In "advanced simple model", **wt** coefficient have negative slope - with weight increasing **mpg** decreases. **qsec** has positive slope - with 1/4 mile time increasing mpg also increases (less car speed - lower air resistance). **am** coefficients represent **wt** and **qsec** slope change by manual transmission (1). Manual transmission has positive **qsec** influence near 0.6 mpg per each additional second (manual transmission is better). And for **wt** coefficient we have negative slope near -2.9 mpg per 1000 lbs weight (manual transmission is worse). We get non-obvious result - manual transmission is better for light-weighted cars with high acceleration time and automatic transmission is better for heavy-weighted cars with low acceleration time. *Does actually automated transmission save fuel for heavy-weighted cars, or model just overfitted the data?* Trying to answer this question, let's look at predicted difference in mpg for manual transmission with actual transmission type from mtcars dataset:

```
prdiff<-fit3$coefficients[4]*mtcars$wt+fit3$coefficients[5]*mtcars$qsec
ggplot(mtcars, aes(rownames(mtcars), prdiff, fill=am))+labs(x="Car model", y="MPG difference")+
      geom_bar(stat="identity")+theme(axis.text.x = element_text(angle = 90))
```



Mostly, cars with negative predicted difference already have automated transmission (except Mazerati Bora and Ford Panthera L with insignificant difference). We have no enough manual transmission heavy cars data for the answer, but automobile manufacturers prefer automated transmission on heavy cars and this gives us strong support for our hypothesis. Now, let's look at model coefficient confidence intervals and significance level for our model:

```
{print(confint(fit3))                          #printing confidence intervals
stat<-summary(fit3)$fstatistic                 #extracting f-statistics
pf(stat[1], stat[2], stat[3], lower.tail = F)} #printing p-value for whole model
```

```
##                  2.5 %     97.5 %
## (Intercept)  2.1186308 25.8197829
## wt          -4.4813221 -1.8704502
## qsec         0.2999593  1.3676125
## wt:am       -4.8457436 -1.0008711
## am:qsec      0.3010031  0.9241765

##        value
## 8.424102e-13
```

We are interested in differential effect, so interception, **wt** and **qsec** terms would be eliminated by subtraction. Confidence intervals for **am** coefficients are quite large because our dataset is small, we need more data for better estimate. P value for the whole model is significant, we have high probability for true effect.