Regression Models Course Project

Hlib Yudin

10/20/2020

Synopsis

The main goal of this analysis is to find out which kind of transmission is better for MPG (miles per gallon) – automatic or manual. The data is taken from the 'mtcars' dataset, which contains information about fuel consumption and 10 aspects of automobile design and performance for 32 cars (1973-74 models). To answer a given question, an exploratory data analysis is conducted and several regression models are built and analyzed.

Exploratory Data Analysis

First of all, it'd be relevant to briefly look over the dataset:

```
# Download the libraries and the data
library(datasets)
library(car)
```

```
## Loading required package: carData
```

```
data(mtcars)
dim(mtcars)
```

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

```
names(mtcars)
```

```
##  [1] "mpg"  "cyl"  "disp" "hp"   "drat" "wt"   "qsec" "vs"   "am"   "gear"
## [11] "carb"
```

```
head(mtcars)
```

```
##                   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
```

The variable of interest, 'am' (transmission), is numeric. However, it takes only binary values, so we should treat is as a factor. 0 stands for automatic transmission, 1 – manual.

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

```
##  0  1
## 19 13
```

```
summary(mtcars$mpg)
```

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

Lastly, we can plot the data from 'mpg' and 'am' columns (see appendix, plot 1). It shows that cars with manual transmission have bigger MPG. However, this may occur due to some other variables which are correlated with 'am'. We're going to check this.

Fitting regression models

The first thing that comes to mind is to fit a linear model where 'mpg' is the outcome and 'am' is the predictor. However, its coefficients wouldn't be useful (b0 – mean of mpg given am = 0; b1 – mean of mpg (am = 1) minus mean of mpg (am=0)). Let's for a start build a model with all 10 predictors. It turns out that none of its coefficients are significant:

```
fitall <- lm(mpg ~ ., mtcars)
summary(fitall)$coefficients # get p-value
```

```
##                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
## am1          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
```

This may occur because a lot of variables are correlated with each other, which inflates the standard error of the coefficients. For example, the number of cylinders may strongly affect the horsepower, displacement, MPG.

Let's build a model which includes as predictors only the horsepower (the higher it is, the bigger fuel consumption), the weight (the heavier the car, the more fuel it spends), and the transmission. ANOVA test may show if there are spare regressors (the normality of residuals is checked in appendix, plot 2):

```
fit_hp_wt_am <- lm(mpg ~ hp + wt + am, mtcars)
anova(lm(mpg ~ hp, mtcars),
      lm(mpg ~ hp + wt, mtcars),
      lm(mpg ~ hp + wt + am, mtcars))
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ hp
## Model 2: mpg ~ hp + wt
## Model 3: mpg ~ hp + wt + am
##   Res.Df    RSS Df Sum of Sq       F    Pr(>F)
## 1     30 447.67
## 2     29 195.05  1   252.627 39.2340 9.028e-07 ***
## 3     28 180.29  1    14.757  2.2918    0.1413
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Turns out that the information about transmission is insignificant! Why? In the plot 3 (see appendix) it is clear that 'am' strongly depends on weight: almost all cars heavier than 3000 lbs have automatic transmission and almost all cars lighter than 3000 lbs have manual transmission. The similar situation is with horsepower. Let's fit a model omitting the transmission:

```
fit_wt_hp <- lm(mpg ~ I(wt-mean(mtcars$wt)) + I(hp-mean(mtcars$hp)), mtcars)
summary(fit_wt_hp)$coef
```

```
##                            Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)             20.09062500 0.45845476 43.822481 4.690472e-28
## I(wt - mean(mtcars$wt)) -3.87783074 0.63273349 -6.128695 1.119647e-06
## I(hp - mean(mtcars$hp)) -0.03177295 0.00902971 -3.518712 1.451229e-03
```

```
confint(fit_wt_hp)
```

```
##                               2.5 %      97.5 %
## (Intercept)             19.15297973 21.02827027
## I(wt - mean(mtcars$wt)) -5.17191604 -2.58374544
## I(hp - mean(mtcars$hp)) -0.05024078 -0.01330512
```

All coefficients are more than 2 standard errors far from zero, thus they are significant. The intercept coefficient means that cars with average weight and average horsepower will have the expected value of MPG equal 20.09. If horsepower is fixed, then increasing the weight by 1000 lbs will decrease MPG by 3.88 (on average). If the weight is fixed, then increasing the horsepower by one unit will on average decrease the MPG by 0.03.

The residual plot (see appendix, plot 5) shows that the residuals of the model aren't distributed evenly around mean 0. In fact, it looks like the model also should have the quadratic term (e.g. the interaction between hp and wt):

```
fit_inter <- lm(mpg ~ hp + wt + hp*wt, mtcars)
summary(fit_inter)$coef
```

```
##                Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 49.80842343 3.60515580 13.815887 5.005761e-14
## hp          -0.12010209 0.02469835 -4.862758 4.036243e-05
## wt          -8.21662430 1.26970814 -6.471270 5.199287e-07
## hp:wt        0.02784815 0.00741958  3.753332 8.108307e-04
```

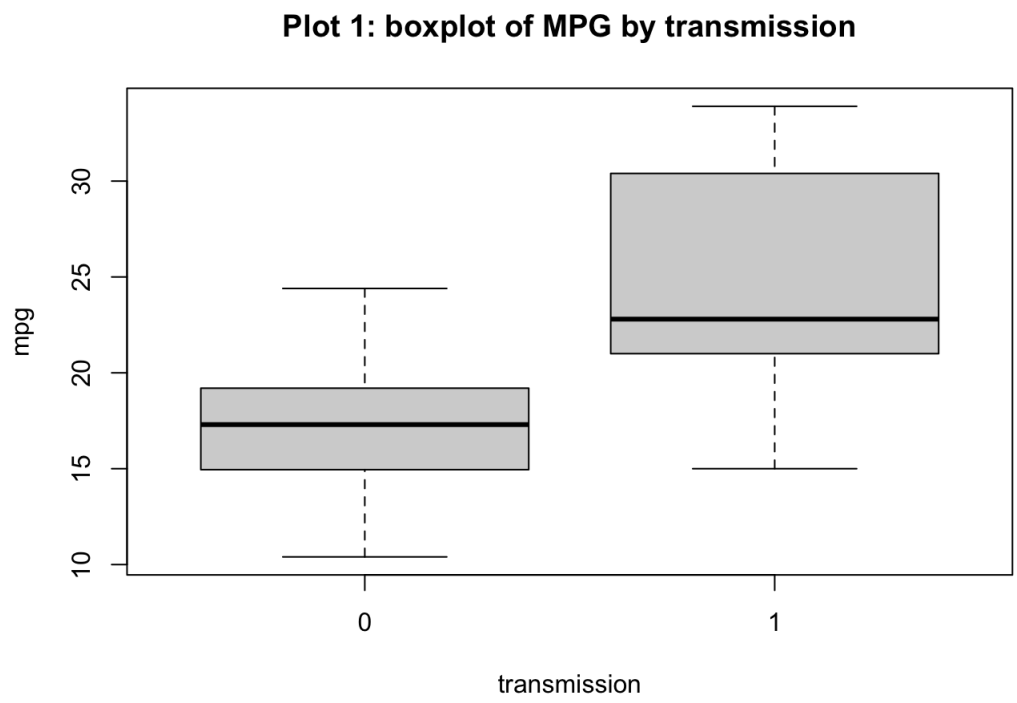
The new residual plot (see appendix, plot 6) looks much better now. As for transmission, it's possible to build a generalized linear model (logistic regression) where 'am' is an outcome, 'wt' and 'hp' are predictors.

Conclusion

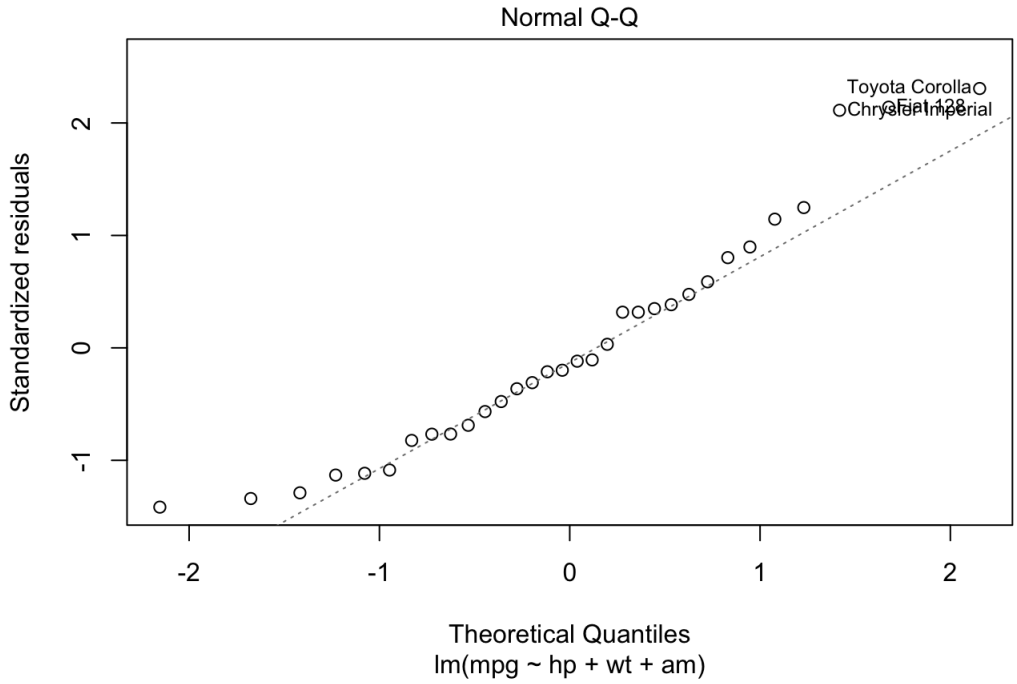
Having analyzed the data, we've concluded that there is no difference between automatic and manual transmission for MPG. The reason is that 'am' is highly correlated with weight: there is a very high probability that the transmission will be manual if the car is lighter than 3000 lbs; automatic otherwise. (The similar can be said about horsepower.) Thus the effect of transmission on MPG (see appendix, plot 1) is fully explained by the correlation with the weight and horsepower. ANOVA test showed that adding regressor-transmission into the model is insignificant (p-value = 0.14). Also we've built a model which describes the relationship between MPG and weight with horsepower.

Appendix

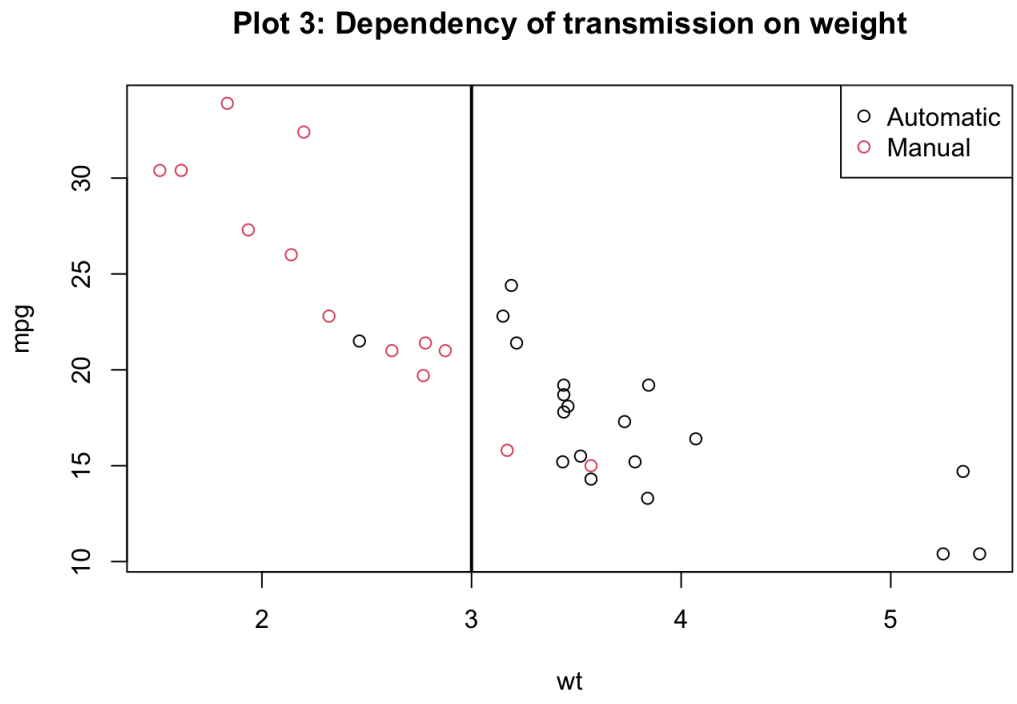
```
# Plot 1: boxplot of MPG by transmission
boxplot(mpg ~ am, mtcars, main = "Plot 1: boxplot of MPG by transmission", xlab = "transmission")
```



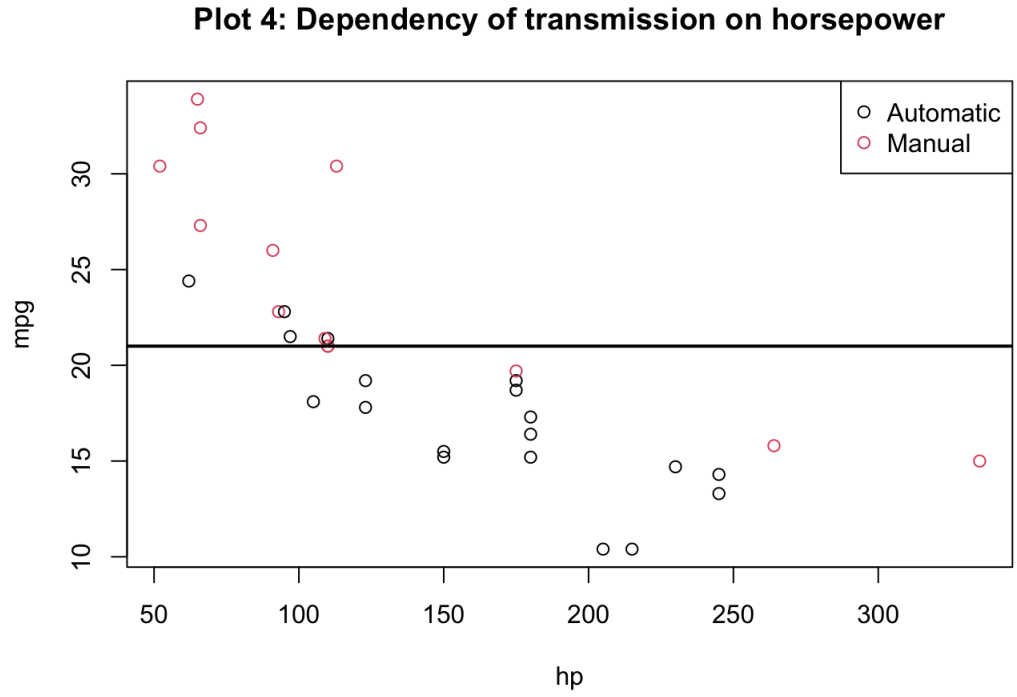

```
# Plot 2: checking the normality of residuals
plot(fit_hp_wt_am, which = 2)
```



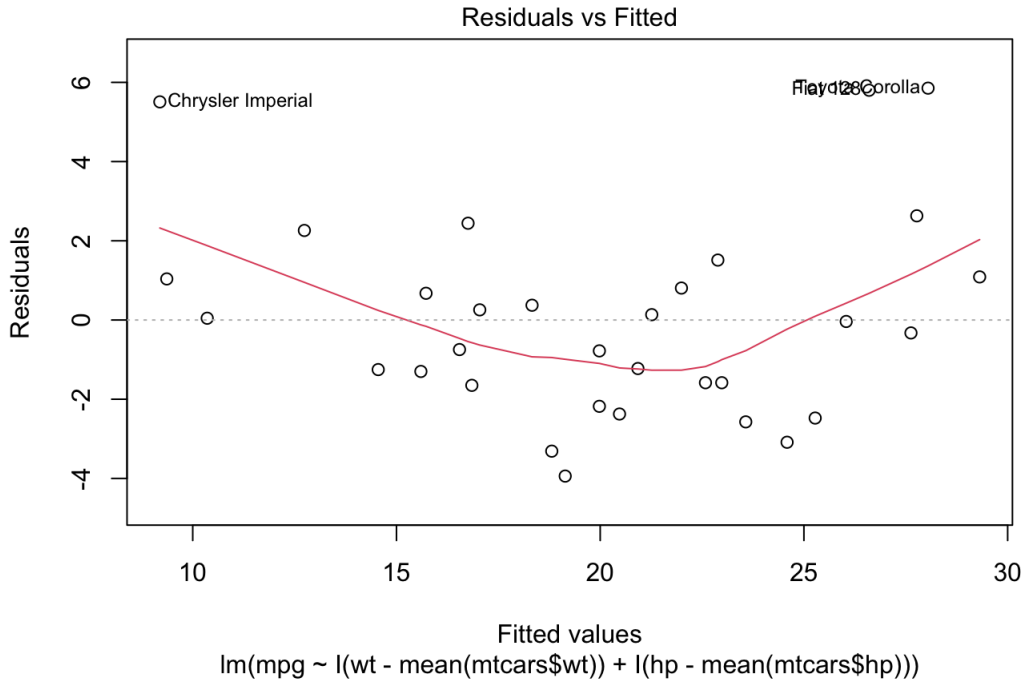

```
# Plot 3-4: dependency of transmission on weight and horsepower
with(mtcars, plot(mpg ~ wt, col = am, main = "Plot 3: Dependency of transmission on weight"))
abline(v = 3, lwd = 2)
legend("topright", pch = 1, col = c(1, 2), legend = c("Automatic", "Manual"))
```



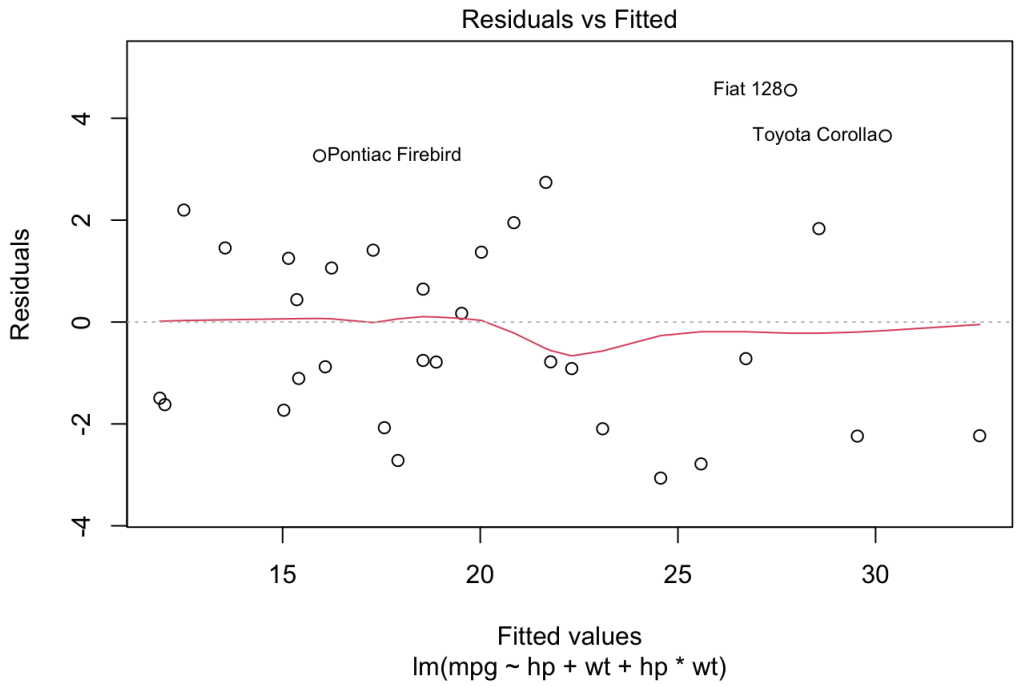

```
with(mtcars, plot(mpg ~ hp, col = am, main = "Plot 4: Dependency of transmission on horsepower"))
abline(h = 21, lwd = 2)
legend("topright", pch = 1, col = c(1, 2), legend = c("Automatic", "Manual"))
```




```
# Plot 5: residual plots, no interaction
plot(fit_wt_hp, which = 1)
```




```
# Plot 6: residual plots, interaction included
plot(fit_inter, which = 1)
```


Residuals vs Fitted
Fiat 128
Toyota Corolla
Pontiac Firebird
Residuals
4
2
0
-2
-4
15
20
25
30
Fitted values
lm(mpg ~ hp + wt + hp * wt)