Regression Models - Course Project

Andrey Budish

April 12, 2017

Instructions

You work for *Motor Trend*, a magazine about the automobile industry. Looking at a data set of a collection of cars, they are interested in exploring the relationship between a set of variables and miles per gallon (MPG) (outcome). They are particularly interested in the following two questions:
1. Is an automatic or manual transmission better for MPG
2. Quantify the MPG difference between automatic and manual transmissions

Executive Summary

1. We used multivariable linear regression analysis and found a signficant difference between the mean MPG for automatic and manual transmission cars.
2. Using manual transmissions results in a higher value of MPG compared to automatic transmission, the increase is approximately 2.1 MPG, if use the best model, which include type of transmission, the weight of a car and horsepower as predictors.

1. Exploratory data analysis

Let's take a quick look on the **mtcars** data set

```
head(mtcars, n = 2)
```

```
##               mpg cyl disp  hp drat    wt  qsec vs am gear carb
## Mazda RX4      21   6  160 110  3.9 2.620 16.46  0  1    4    4
## Mazda RX4 Wag  21   6  160 110  3.9 2.875 17.02  0  1    4    4
```

The data is described here.

Transmission type **am** has numeric class:

```
class(mtcars$am)
```

```
## [1] "numeric"
```

Let's change it to factor and also change 0 values to Automatic and 1 values to Manual for better readability:

```
mtcars$am <- as.factor(mtcars$am)
levels(mtcars$am) <- c("Automatic", "Manual")
```

Check MPG distribution

Since we will be predicting MPG we need to know if data in the sample is normally distributed.
We will use Shapiro-Wilk test:

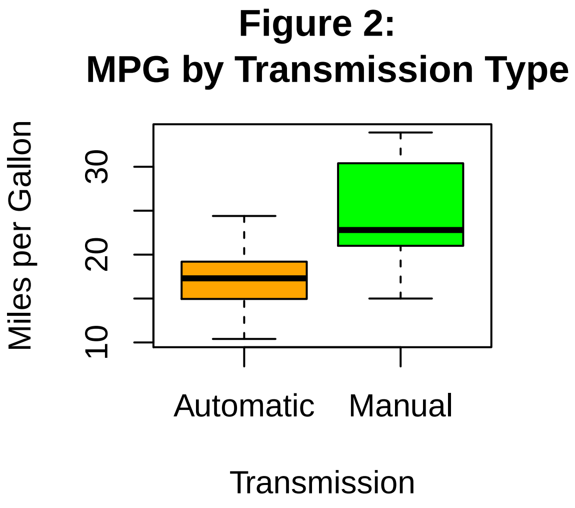
```
shapiro.test(mtcars$mpg)$p.value
```

```
## [1] 0.1228814
```

p-value > 0.05, so we fail to reject the null hypothesis, that data is normally distributed.

Automatic vs manual transmission boxplot

To compare automatic versus manual transmission influence on miles per gallon it would be useful to make a relevant box plot:



Plot shows a certain difference in the MPG by transmission type. We need to test this assumption.

2. Hypothesis testing

Let's look at the means of miles per gallon for both automatic and manual transmission:

```
aggregate(mpg ~ am, data = mtcars, mean)
```

```
##          am      mpg
## 1 Automatic 17.14737
## 2    Manual 24.39231
```

Is this difference, 7.25 MPG, really significant? Let's perform a relevant unpaired t-test with unequal variance, where we assume that using manual transmission results in greater miles per gallon.

```
manual <- subset(mtcars, am == "Manual", select = mpg)
auto <- subset(mtcars, am == "Automatic", select = mpg)
t.test(manual, auto, alternative = "greater")$p.value
```

```
## [1] 0.0006868192
```

p-value is less than 0.05 which confirms our assumption that using manual transmission better than automatic for miles per gallon value. But we didn't consider other variables influence on MPG.

3. Model selection

Simple model - one predictor

Let's begin our model selection with the simplest model by predicting mpg by only one variable - type of transmission

```
basic <- lm(mpg ~ am, mtcars)
```

```
summary(basic)$coef
```

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

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

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

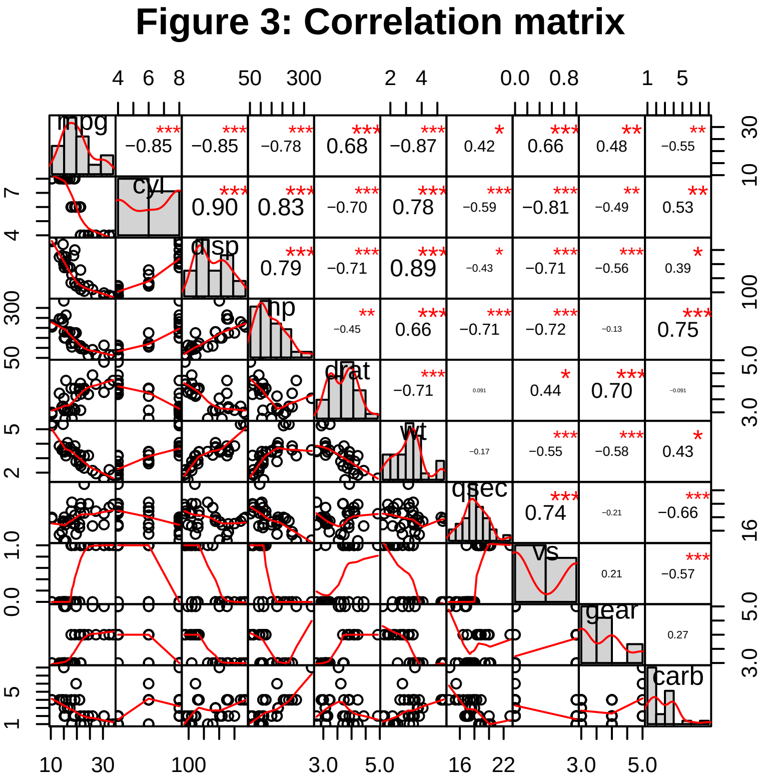
Pr(>|t|) are less than 0.05 for manual and automatic(Intercept) transmission demonstrate that it is right to include transmission type for predicting MPG.

The value of multiple R-squared of 0.36 tells us that only 36% of total variance explained by our model. That means that we should add other regressors to fit our model better.

Choosing multiple predictors

Firstly, to understand what predictors should be included we will investigate correlation matrix, which will show correlation of variables between each other.
We will exclude **am** variable from correlation matrix since we know it should be included in the model anyway.

```
# excluding am variable, 9th column
cars_without_am <- mtcars[, - 9]
library(PerformanceAnalytics)
chart.Correlation(cars_without_am, main = "Figure 3: Correlation matrix")
```



```
# Pick only the correlations of MPG:
cor_mpg <- cor(cars_without_am)[-1, 'mpg']
# Show correllations of MPG in the descending order
sort(round(cor_mpg, digits = 2))
```

```
##    wt   cyl  disp    hp  carb  qsec  gear    vs  drat
## -0.87 -0.85 -0.85 -0.78 -0.55  0.42  0.48  0.66  0.68
```

As we can see **mpg** variable is strongly *(>0.7 and <-0.7)* and moderately*([-0.7,0.5] and [0.5,-0.7])* correlated with almost all variables. Red star chars (***, **, *) in the correlation matrix show how significant those result are. Including all variables in the model would be wrong, since it increases standard error of the beta coefficients.
Let's pick only regressors strongly correlated to MPG and build a relevant model.
That would be: **wt, cyl, disp, hp and am**(should be included anyway)

```
mx <- lm(mpg ~ am + wt + cyl + disp + hp , data = mtcars )
```

Variance Infaltion Factors

Next we will explore variance inflation factors:

```
# library car is loaded to use vif()
library(car)
vif(mx)
```

```
##         am         wt        cyl       disp         hp
##   2.553064   6.079452   7.209456  10.401420   4.501859
```

As we can see both **cyl** and **disp** have high inflation factors, that means that they are highly correlated. We can safely leave out one of them and build a simpler model. Let us omit **disp**.

```
my <- lm(mpg ~ am + wt + cyl + hp, data = mtcars )
vif(my)
```

```
##       am       wt      cyl       hp
## 2.546159 3.988305 5.333685 4.310029
```

Now **cyl** and **hp** are highly correlated. Let us leave out **cyl**

```
mz <- lm(mpg ~ am + wt + hp , data = mtcars )
vif(mz)
```

```
##       am       wt       hp
## 2.271082 3.774838 2.088124
```

In final model **mz** variance inflation factors for all regressors are small (vif < 5). That means we do not need to omit any of the left regressors.

ANOVA

Next we will perform analysis of variance of all **4 nested models: basic, mz, my and mx**.

```
anova(basic, mz, my, mx)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ am + wt + hp
## Model 3: mpg ~ am + wt + cyl + hp
## Model 4: mpg ~ am + wt + cyl + disp + hp
##   Res.Df    RSS Df Sum of Sq       F    Pr(>F)
## 1     30 720.90
## 2     28 180.29  2    540.61 43.0841 5.576e-09 ***
## 3     27 170.00  1     10.29  1.6407    0.2115
## 4     26 163.12  1      6.88  1.0963    0.3047
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

As we can see **adding wt and hp** to the basic model **was significant**, **adding cyl and disp in the next models was not significant**. That means that **Model 2, mz is the best among those 4 models**.

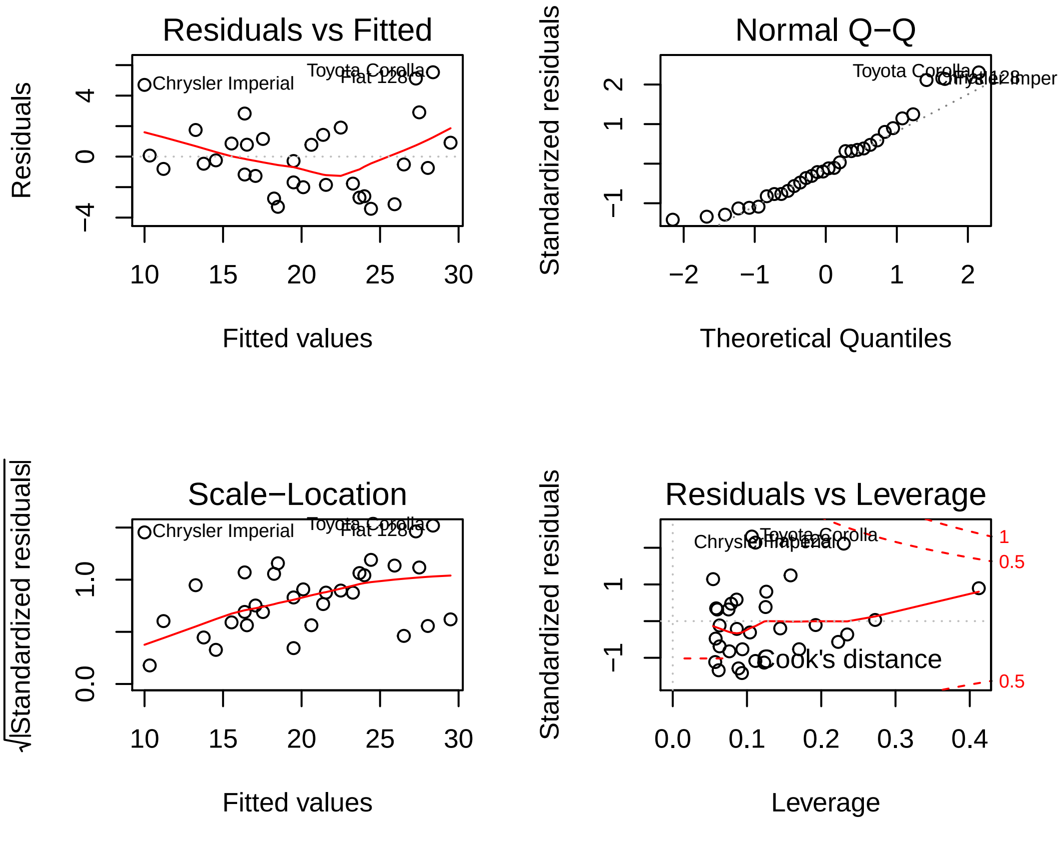
Residuals plots of final model

Let us use residuals plots to check goodness of the model fit:

```
par(mfrow = c(2,2), oma=c(0,0,4,0))
plot(mz, main = NULL)
title("Figure 4: Residuals plots of final model", outer=TRUE)
```

Figure 4: Residuals plots of final model

lm(mpg ~ am + wt + hp)



Residuals plots confirm goodness of fit.

Summary of final model

In the end let us look at the **summary of mz**:

```
sum_mz <- summary(mz)
sum_mz$coefficients
```

```
##                Estimate  Std. Error   t value     Pr(>|t|)
## (Intercept) 34.00287512 2.642659337 12.866916 2.824030e-13
## amManual     2.08371013 1.376420152  1.513862 1.412682e-01
## wt          -2.87857541 0.904970538 -3.180850 3.574031e-03
## hp          -0.03747873 0.009605422 -3.901830 5.464023e-04
```

```
sum_mz$adj.r.squared
```

```
## [1] 0.8227357
```

```
# Difference in MPG between manual and automatic transmissions
sum_mz$coefficients[2, 1]
```

```
## [1] 2.08371
```

4. Conclusion

Our final model mz, which includes am, wt and hp as regressors to predict MPG, explains 82 % of total variance (adjusted r__squared).

The difference between manual and automatic transmissions is 2.1 in MPG, if we hold other variables constant, which still shows us that manual transmission is better for MPG.

At the same time, there are some caveats: we assumed that MPG data comes from normal distribution, which might be wrong; we also used only variables that we were able to get from mtcars data set. There might be other variables that should be included in the model which we don't know about.