Cars MPG Analysis

Amade A.

December 25, 2015

Overview

Research objective

The objective of this paper is answering the question whether automatic transmission is better for MPG compared to manual transmission. In addition, the researcher also aims to quantify the difference in MPG between cars with different types of transmission.

Executive summary

This paper analyzes the influence of various car characteristics on fuel economy (measured as miles per galon of fuel).

Exploratory Data Analysis

Dataset

Dataset used for the analysis is **mtcars** dataset, preloaded in R in the *datasets* package. Data originates from the 1974 Motor Trend US magazine, measuring fuel consumption across 10 automobile design aspects for 32 automobiles.

Among the 10 variables, 5 (cyl, vs, am, gear, carb) can be characterized as discrete (factor) variables and remaining 5 (disp, hp, drat, wt, qsec), continuous (numeric) variables. Fuel consumption is represented by the continuous *mpg* variable.

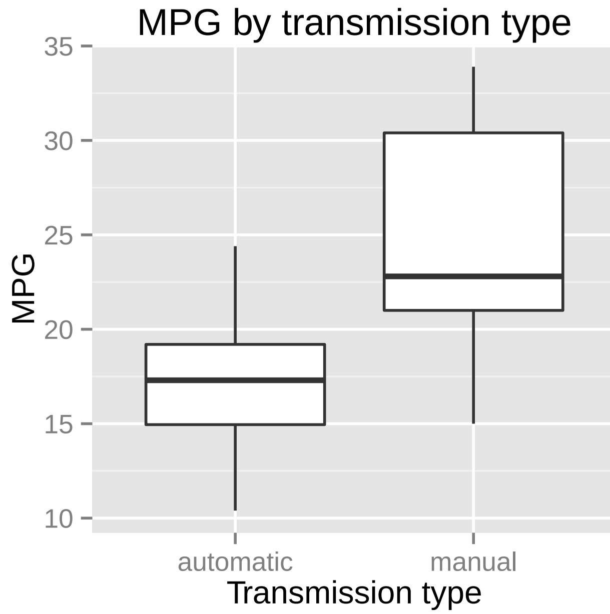
For easier interpretation, variable specifying the transmission type - *am* was recoded, so that the factor levels have straightworward internpretation.

```
library(dplyr, warn.conflicts = FALSE)
mtcars <- mtcars %>%
mutate(am = factor(am, levels = c(0,1), labels = c("automatic", "manual")))
```

MPG by transmission type

In total, there were 19 cars with automatic transmission and 13 with manual. The following graph shows the differences in measured fuel efficiency among cars with different transmission types.

```
library(ggplot2)
p <- ggplot(mtcars, aes(am, mpg))
p <- p + geom_boxplot()
p <- p + ggtitle("MPG by transmission type") + xlab("Transmission type") + ylab("MPG")
p
```



It can be easily seen, that the fuel efficiency is not the same among the cars with different transmission types, which means that *am* variable is a good candidate for independent regression variable. Median MPG for cars with automatic transmission equals **17.3** and for the ones with manual **22.8**. Values seem to be more concentrated around the median in the case of automatic transmission, compared to cars with manual transmission. Standard deviations of MPG value for two groups equaled respectively **3.83** and **6.17**.

Formal modelling

Simple model (automatic/manual only)

As a first step in modeling a simple linear model with one binary variable *am* was fitted. Parameters of the regression are summarized below.

```
fit <- lm(mpg ~ am, data = mtcars)
summary(fit)$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
```

Since the model has only one binary variable *am*, the intercept is interpreted as predicted value of MPG for the base level of *am* variable, which in **mtcars** dataset is “automatic transmission”. Value of 17.15 informs us that cars with automatic transmission are expected to reach 17.15 miles for each galon of fuel. This value is also the average MPG value for cars with automatic transmission.

Change in MPG value for cars with manual transmission is represented by *ammanual* coefficient. Value of 7.24 tells us that the expected (average) MPG for cars with manual transmission is 7.24 higher compared to cars with automatic transmission.

P value for *ammanual* coefficient is very close to 0, informing us that the mean MPG range is significantly different for cars with different transmission type even at 99.9% confidence level.

R^2 of the model equaled 0.36, meaning that 36% of the variation in MPG is explained by the regression model.

Inclusion of additional variable

Since the value of R square coefficient is relatively low (0.36) it is quite likely that there are other variables, which can help in explaining the variation in MPG.

Additional variables to the model can be determined by looking at the absolute value of correlation of numeric variables with *mpg* variable.

```
with(mtcars, data.frame(
    absoluteCorrelation = round(c(abs(cor(mpg,disp)), abs(cor(mpg,hp)), abs(cor(mpg,drat)),
        abs(cor(mpg,wt)),abs(cor(mpg,qsec))),2),
    row.names = c("disp", "hp", "drat", "wt", "qsec")
    )
)
```

```
     absoluteCorrelation
disp                0.85
hp                  0.78
drat                0.68
wt                  0.87
qsec                0.42
```

Two most higly correlated variables with *mpg* variable are *wt* and *disp* variables, referring respectively to automobile's weight (in 1000lbs) and engine displacement (in cu.in.).

Model wtih including automobile's weight

Following code fits the linear model using *am* and *wt* variables.

```
fit2 <- lm(mpg ~ am + wt, data = mtcars)
summary(fit2)$coef
```

```
               Estimate Std. Error     t value     Pr(>|t|)
(Intercept) 37.32155131  3.0546385 12.21799285 5.843477e-13
ammanual    -0.02361522  1.5456453 -0.01527855 9.879146e-01
wt          -5.35281145  0.7882438 -6.79080719 1.867415e-07
```

```
summary(fit2)$r.square
```

```
[1] 0.7528348
```

As a result of including the weight variable to the model, *am* variable lost its explanatory power. Its value of -0.02 is very close to 0. Such low value of coefficient would indicate that the mean MPG for cars with manual transmission is lower by -0.02 compared to cars with automatic transmission. The p value of 0.99 doesn't allow us to reject the null hypothesis of no differences in mean MPG between cars with automatic and manual transmission at any conventional conficenve level.

It can be therefore concluded that the regression model with both *am* and *wt* variables is not appropriate for answering the question of impact of transmission type on automobile's fuel efficiency.

Model wtih including engine displacement

The following code fits the model using *am* and *disp* variables. *disp* variable measures the engine displacement.

```
fit3 <- lm(mpg ~ am + disp, data = mtcars)
summary(fit3)$coef
```

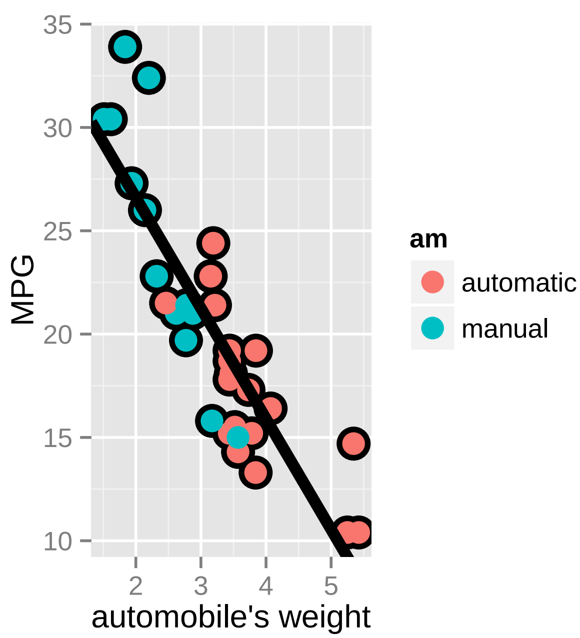
```
               Estimate  Std. Error   t value     Pr(>|t|)
(Intercept) 27.84808111 1.834071377 15.183750 2.452658e-15
ammanual     1.83345825 1.436099585  1.276693 2.118396e-01
disp        -0.03685086 0.005781896 -6.373490 5.747528e-07
```

```
summary(fit3)$r.square
```

```
[1] 0.7333314
```

Also in the case of this model, the p value of 0 for coefficient *ammanual* indicates that it has no significant effect on exaplaining variation in MPG. Therefore, despite relatively large R^2 value of 0.73, this model is also not appropriate for explaining the impact of transmission type on automobile's MPG.

```
g <- ggplot(data = mtcars, aes(x = wt, y = mpg, colour = am))
g <- g + geom_point(size = 6, colour = "black") + geom_point(size = 4)
g <- g + geom_abline(intercept = coef(fit2)[1], slope = coef(fit2)[3], size = 2)
g <- g + geom_abline(intercept = coef(fit2)[1] + coef(fit2)[2], slope = coef(fit2)[3], size = 2)
g <- g + xlab("automobile's weight") + ylab("MPG")
g
```



Best model selection

The most important criterion for model selection is that it must allow us to see the impact of transmission type on MPG. Therefore, such model should include *am* variable. Here perform a likelihood ratio test

Uncertainty measures

Analysis of residuals

Here plot the residuals

Confidence intervals for predictor variables

Specify here the confidence intervals for those variables

Conclusions and interpretation

Interpret the outcome of the model