

Regression Models Course Project

Sarthak

03/11/2020

Executive Summary

In this project, analysis of the `mtcars` data set is done; and the relationship between some variables and the Miles per Gallon (MPG) of cars is explored and analyzed. The main questions that are solved in this project are: *Is automatic or manual transmission better for MPG?*, and *Quantify the MPG difference between automatic and manual transmission*. Various regression models and exploratory data analyses are used to analyze how having automatic or manual transmission in a car affects its MPG rating. To check whether automatic or manual transmission is better for performance, several regression models are fitted on the data set with MPG as outcome, and the one with highest adjusted R-squared value is selected (this high value is obtained by using other variables in the model fit alongside the automatic transmission variable). From the best fit model, considering 1/4 mile time and weight of the car to be constants, the MPG of manual transmission cars is $14.08 + (-4.14) * \text{weight}$ more on average than automatic transmission cars. Performing T-tests between the mpg of cars with manual transmission and automatic transmission suggests that there is a significant difference between the MPG of cars with manual transmission and the MPG of cars with automatic transmission. From the t-test, it is evident that the MPG of manual transmission cars are approximately 7 more than that of automatic transmission cars.

Question 1

1. Exploratory Data Analysis

Some of the variables in the data set that have multiple values of the same type are converted into factor variables

```
# loading the mtcars data set
data("mtcars")

# converting some variables into factor variables
mtcars$am <- as.factor(mtcars$am)
mtcars$vs <- as.factor(mtcars$vs)
mtcars$gear <- as.factor(mtcars$gear)
mtcars$cyl <- as.factor(mtcars$cyl)
mtcars$carb <- as.factor(mtcars$carb)
```

A box-plot is created showing the difference between the `mpg` of the cars with manual transmission and automatic transmission (can be found in *appendix*)

2. Regression Modeling

A linear regression model with `mpg` as outcome and the other variables as the predictors is created, just as a starting point to see how to proceed.

```
# output of code is suppressed
baseModel <- lm(mpg ~ ., data = mtcars)
summary(baseModel)
```

From the model, the adjusted R-squared value is 78%, which means that the model can explain 78% of the variance in the `mpg` variable. Also, all of the coefficients in the model have a significance level (p-value) of more than 0.05, which means that none of the coefficients are significant in the model. The residual standard error for this model is 2.833 with 15 degrees of freedom.

To select the most statistically significant variables, backward selection method is used.

```
# output of code is suppressed
backModel <- step(baseModel, k = log(nrow(mtcars)))
summary(backModel)
```

From the backward selection model, the selected variables are `am`, `qsec` and `wt`, wherein `am` is factor variable with 0 denoting automatic and 1 denoting manual transmission, `qsec` is 1/4 mile time, and `wt` is the car weight in 1000 lbs. The adjusted R-squared is 0.8497, which is more than the previous model. All the variables are significant, with a significance level below 0.05. The residual standard error for this model is 2.46 with 28 degrees of freedom.

In the scatter plot comparing `mpg` vs `wt` with `am` used to color the dots (see in the *appendix*), it is evident that there is some interaction between `wt` and `am`, wherein automatic transmission cars seem to have more weight than manual transmission cars. So, an interaction term can be added to the above model.

```
# output of code is suppressed
backModelInt <- lm(mpg ~ wt + qsec + am + wt:am, data = mtcars)
summary(backModelInt)
```

The model has an adjusted R-squared of 0.8959, which is even more than the best model found from the backward selection method. Also, all of the coefficients have a significance level below 0.05, which means that they are significant coefficients. The residual standard error for this model is 2.084 with 27 degrees of freedom (the lowest out of the three models).

The last model being fitted considers `mpg` as outcome and `am` as the predictor.

```
# output of code is suppressed
simModel <- lm(mpg ~ am, data = mtcars)
summary(simModel)
```

The adjusted R-squared for this model is 0.36, which is much lower than the other three models. The cars with automatic transmission has an average MPG of 17.147, and the average MPG for manual transmission cars is 7.245 more than the automatic. The residual standard error for this model is 4.902 with 30 degrees of freedom. Since the adjusted R-squared value is so low, other variables should be included in the model (as has been done in the previous three models).

The final model selected is the `backModelInt`, with `wt`, `qsec`, `am` and `wt:am` as the predictors.

```
#summary of the coefficients
summary(backModelInt)$coef
```

##	Estimate	Std. Error	t value	Pr(> t)
## (Intercept)	9.723053	5.8990407	1.648243	0.1108925394
## wt	-2.936531	0.6660253	-4.409038	0.0001488947
## qsec	1.016974	0.2520152	4.035366	0.0004030165
## am1	14.079428	3.4352512	4.098515	0.0003408693
## wt:am1	-4.141376	1.1968119	-3.460340	0.0018085763

From the coefficients' summary, it is evident that the manual transmission cars are better than automatic transmission cars for MPG. Considered the `wt` and `qsec` variable constant, the MPG for manual transmission cars is $14.08 + (-4.14) * \text{weight}$ more on average than that of automatic transmission cars. Thus, a manual transmission car that has a weight of 3000 lbs (or 3 in the `wt` variable) will have $14.08 + (-4.14) * 3 = 1.66$ MPG more than the automatic transmission car with the same weight, considered `wt` and `qsec` as constants.