Regression Models Project

Marcel M.

4/17/2020

MOTOR TREND REPORT

*Due to setting issues, I had to write as R markdown document and convert to HTML before printing PDF.

Executive Summary

This report looks at a data set of a collection of cars in order to explore the relationship between a set of variables and their impact on miles per gallon. The main goal is to ascertain and quantify potential discrepancies between effects of different transmission systems:

1. Is an automatic or manual transmission better for MPG?
2. Quantify the MPG difference between automatic and manual transmissions.

Exploratory Data Analysis

The data set (mtcars) comprises fuel consumption and 10 aspects of automobile design and performance for 32 automobiles (1973 - 74 models). It is a data frame composed of 11 variables and 32 observations. The variable of interest, Transmission (am), can assume two values: 0, when the transmission system is automatic, and 1, when it is manual.

System	MPG.Min.	MPG.1st.Qu.	MPG.Median	MPG.Mean	MPG.3rd.Qu.	MPG.Max.
Automatic	10.4	14.95	17.3	17.14737	19.2	24.4
Manual	15.0	21.00	22.8	24.39231	30.4	33.9

*Code in Appendix A

The table* above displays the summary statistics for MPG according to transmission systems. It is possible to infer, at first sight, that Manual Transmission allows to cover more miles per gallon. On average, manual systems spawns 24.392 miles/gallon, whereas automatic ones reach only 17.147. Nevertheless, it is important to check whether such difference is in deed significant.

Linear Model: MPG versus Transmission System (Code in Appendix B)

lm(mpg ~ Transmission, mtcars) Coefficients: Automatic System = 17.147 Manual System = 7.245

The linear regression model of MPG on Transmission supports the data previously displayed in the table. Therefore it is expected that, on average, cars with manual transmission system drive 7.245 miles/gallo more in comparison to those equipped with automatic systems.

T-test: Comparing Group Means

```
## [1] 0.00137
```

Given the low p-value (0.00137), the t-test to compare the system means rejects the null hypothesis that the true difference in means is equal to zero. Therefore, the difference in means observed between manual and automatic transmission systems is significant at 5% level of significance (Code in Appendix C).

Model Selection

Detect what variables are more correlated to the outcome “MPG”

	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
mpg	100	85	85	78	68	87	42	66	60	48	55

The analysis in absolute values of the correlation between the outcome and the predictors suggests that it is worth to consider a model including cyl, disp, wt and hp, besides the am variable (Code in Appendix D).

Variance Inflation Factor

	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
VIF	15.373833	21.620241	9.832037	3.374620	15.164887	7.527958	4.965873	4.648487	5.357452	7.908747

The Variance Inflation theory claims that predictors with VIF > 10 must be dropped out of the model in order to avoid multicollinearity. Therefore, by cross-checking predictors-outcome correlation and manageable VIF´s, some feasible variables to include in the model are: hp, drat, gear and vs (Code in Appendix E)

Nested Models

Strategy for model selection which consists in fitting multiple models where the next model comprises de predictors of the previous one. ANOVA is applied to compare the models with each other.

According to Anova analysis, the best model is “fit2” (Code in Appendix F).

fit2 = lm(mpg ~ Transmission + hp, mtcars). Summary in Appendix G.

```
##        (Intercept) TransmissionManual                hp
##         26.5849137          5.2770853         -0.0588878
```

When “mpg” is regressed on “Transmission”(or “am”) and “hp”, the impact of transmission systems appears to increase mpg: an automatic system - on average and holding other variables constant - spawns 26.585 miles per gallon, whereas when horse power is omitted this value was about 17.147 miles. In case of switching the system to manual, there is a gain of 5.2777 miles per gallon in comparison to an automatic transmission, which means 31.862 miles per gallon for automobiles with manual transmission. On the other hand, when gross horsepower (hp) increases by 1 unit, the outcome decreases on average by 0.0589 miles per gallon. Moreover, the new model explains better the variance in “mpg”, since its Adjusted R-squared is of 76.7%, against 35.9% of the univariate model. All values are significant at 0.1% significance level (Appendix E).

Residual Plot

Plots are displayed in Appendix H.

Testing for Homoscedasticity

Check whether the variance of errors is constant. Code in Appendix I.

```
## [1] 0.0697784
```

A p-value equal to 0.0698 indicates the test fails to reject the null hypothesis that the variance of errors is homocesdastic at a 5% level of significance.

Residual Analysis

The plot Residuals versus Fitted Values indicates there is no systematic pattern in the residuals and the variance of errors is somewhat constant, which is in line with the ncvTest for homoscedasticity. The lack of a systematic pattern in the residuals plot might suggest the model selected is good to predicted mpg values, since the residuals appear not to explain systematic variations of the outcome. Regarding the Normal Q-Q plot, we see points falling about a straight line in the graph, which provides evidence that the residuals follow in deed a normal distribution.

Conclusion and notes on uncertainty

The main goal of this project is to provide evidences to help answering two questions:

1. Is an automatic or manual transmission better for MPG?
2. Quantify the MPG difference between automatic and manual transmissions.

Regarding the question 1, the linear models fitted in this analysis point out that, on average, manual systems spawns 24.392 miles/gallon, whereas automatic ones reach only 17.147. In this way, manual transmission systems appear to be better for MPG. The difference between the two systems is of 7.245 miles/gallon, which answers question 2. When a model considering “horsepower” as an additive predictor, such difference drops to 5.277 on average, holding other variables constant.

Nevertheless, these results must be infered with prudence, since there are some aspects of the regression model that entail a considerable level of uncertainty. Despite that the residual standard error falls to 2.909 in the selected model, it remains high, which means a high deviation of the model from the true regression line. Moreover, many predictors highly correlated to the outcome “mpg” were dropped out of the model due to their high VIF values. Omitting relevant variables from the model is one of the main causes of biased estimators. On the other hand, the “horsepower” inclusion in the model, as suggested by Anova, may increase the variance of the coefficient estimates and make the estimates very sensitive to minor changes in the model, given its VIF is about 10.

APPENDIX

APPENDIX A: Data Analysis

```
# Label "am" and transform it into categorical variable.
mtcars$Transmission <- factor(mtcars$am, labels = c("Automatic", "Manual"))

# Data Analysis of "MPG" according to "Transmission" levels.
library(data.table)
summaryTr <- aggregate(mtcars$mpg, list(mtcars$Transmission), FUN=summary)
names(summaryTr) <- c("System", "MPG")
tableTr <- data.table(summaryTr)
# Table to display results.
library(formattable)
tableStats <- formattable(tableTr, align = c("l","c","c","c","c","c","r"), list("System"= color_tile("aliceblue","cadetblue3"), "MPG.Min."= color_tile("aliceblue","cadetblue3"), "MPG.1st.Qu."= color_tile("aliceblue","cadetblue3"), "MPG.Median"= color_tile("aliceblue","cadetblue3"), "MPG.Mean"= color_tile("aliceblue","cadetblue3"), "MPG.3rd.Qu."= color_tile("aliceblue","cadetblue3"), "MPG.Max."= color_tile("aliceblue","cadetblue3")))
```

APPENDIX B: Linear Regression Model on Transmission Systems.

```
mpgTr <- lm(mpg ~ Transmission, mtcars)
paste("Automatic System =", round(coef(mpgTr)[1],3))
```

```
## [1] "Automatic System = 17.147"
```

```
paste("Manual System =", round(coef(mpgTr)[2],3))
```

```
## [1] "Manual System = 7.245"
```

APPENDIX C: Comparing Group Means.

```
#    Ho : Mean_Automatic = Mean_Manual
#    Ha : Mean_Automatic != Mean_Manual

tTest <- t.test(mpg ~ Transmission, data = mtcars, var.equal = FALSE, paired = FALSE )
round(tTest$p.value,5)
```

```
## [1] 0.00137
```

APPENDIX D: Correlation table

```
corMtcars <- cor(mtcars[,unlist(lapply(mtcars, is.numeric))])
corMPG <- corMtcars[1,]
dfCor <- round(data.frame(t(corMPG)),2) * 100 # display percentage
rownames(dfCor) <- "mpg"

CorrTable <- formattable(abs(dfCor), align = c("l",rep("r", NCOL(abs(dfCor)) - 1)),
list(`Indicator Name` = formatter("span", style = ~ style(color = "grey",font.weight = "bold")), area(col = 1:11) ~ function(x) percent(x / 100, digits = 1), area(col = 1:11) ~ color_tile("#FFCCFF", "#CC00CC")))
```

APPENDIX E: Variance Inflation Factor

```
library(car)
fitALL <- lm(mpg ~ . - Transmission, mtcars)
inflation <- vif(fitALL)
dfVIF <- data.frame(t(inflation))
rownames(dfVIF) <- "VIF"

vifTable <- formattable(dfVIF, align = c("l",rep("r", NCOL(dfVIF) - 1)), list(
    `Indicator Name` = formatter("span", style = ~ style(color = "grey",font.weight = "bold")),
     area(col = 1:10) ~ color_tile("#FFFF99", "#FF8000")))
```

APPENDIX F: Nested Models and ANOVA

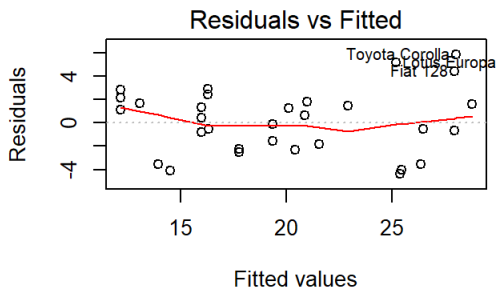
```
fit1 <- lm(mpg ~ Transmission, mtcars)
fit2 <- update(fit1, mpg ~ Transmission + hp)
fit3 <- update(fit2, mpg ~ Transmission + hp + drat)
fit4 <- update(fit3, mpg ~ Transmission + hp + drat + gear)
fit5 <- update(fit4, mpg ~ Transmission + hp + drat + gear + vs)

ANOVA <- anova(fit1,fit2,fit3,fit4,fit5)
```

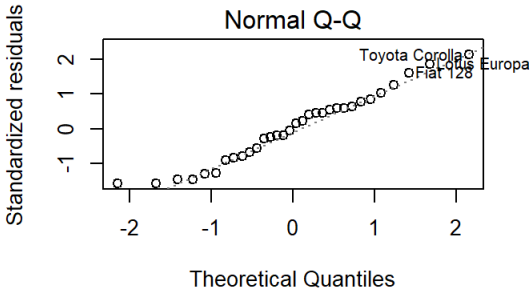
APPENDIX G: Model Fit 2.

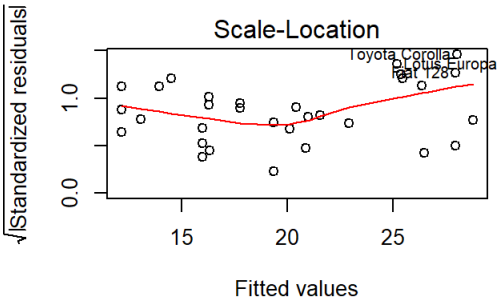
```
summary(fit2)
```

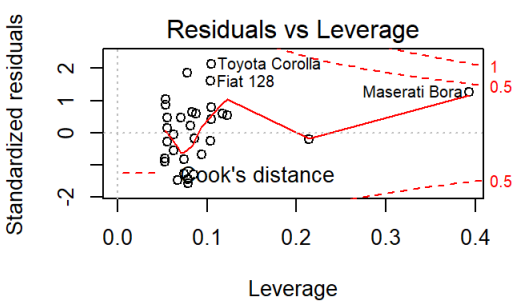
```
##
## Call:
## lm(formula = mpg ~ Transmission + hp, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.3843 -2.2642  0.1366  1.6968  5.8657
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)        26.584914   1.425094  18.655  < 2e-16 ***
## TransmissionManual  5.277085   1.079541   4.888 3.46e-05 ***
## hp                 -0.058888   0.007857  -7.495 2.92e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.909 on 29 degrees of freedom
## Multiple R-squared:  0.782,  Adjusted R-squared:  0.767
## F-statistic: 52.02 on 2 and 29 DF,  p-value: 2.55e-10
```

APPENDIX H: Residual Plot









APPENDIX I: Testing for Homoscedasticity

```
#   Ho : Var_error = Sigma
#   Ha : Var_error = Sigma_ith

hm <- ncvTest(fit2)
hm$p
```

```
## [1] 0.0697784
```