Regression models course project

By Arvind Krishna

Executive Summary

We have used 'mtcars' dataset within the 'datasets' library in R to study the impact of transmission type of a car on the miles run per gallon of the fuel burnt by the car as part of the course requirements for the course on 'Regression models' by John Hopkins University on Coursera.

The objective was to determine:

- Is an automatic or manual transmission better for MPG"
- Quantify the MPG difference between automatic and manual transmissions

Our analysis concluded that:

1) Visually, manual transmission gives more miles per gallon than automatic transmission.

2) The mean for MPG obtained in case of manual transmission is significantly higher than that obtained by automatic transmission.

3) Manual transmission gives 1.8 miles more than automatic transmission after adjusting for number of cylinders, weight and horse power.

Data analysis

Data import and processing

Reading the data 'mtcars'

```
library(datasets)
data(mtcars)
```

The variables in data are described in section 1 of the Appendix.

Data visualisation

We visualise pairwise correlation between all variables of the dataset (Section 2 in Appendix). We observe from the pairwise charts that mpg seems to have strong correlation with cyl, disp, hp, drat, wt, vs and am. But we will use linear models to quantify that in the regression analysis section.

Since we are particularly interested in identifying the effects of car transmission type on mpg, we plot boxplots of mpg vs am (Section 3 in the Appendix). This plot clearly depicts an increase in the mpg when the transmission is Manual.

Testing dependence of mpg on transmission

Although through the box plots, it seems like type of transmission does have an impact on mpg, we will perform t-test to verify if the impact is significant.

```
Auto_trans <- subset(mtcars, mtcars$am=="0")
Manual_trans <- subset(mtcars, mtcars$am=="1")
Impact <- t.test(Auto_trans, Manual_trans, alternative="greater", paired=F)
Impact
```

```
##
##  Welch Two Sample t-test
##
## data:  Auto_trans and Manual_trans
## t = 1.8772, df = 348.4, p-value = 0.03066
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  1.918621      Inf
## sample estimates:
## mean of x mean of y
##  46.02645  30.22848
```

We p-value of 0.03 proves that on an average mpg for manual transmission is significantly larger than that for automatic transmission. We will quantify the impact of transmission type on mpg in the next section where we build a regression model.

Regression model

We will begin from an intitial model in which we will try to compute mpg using all the information (variables) that we have in the data. Thereafter we will use 'step' function which runs 'lm' multiple times to build regression models and select the best variables using both forward selection and backward elimination methods. The stepwise algorithm minimises AIC which is an indicator of obtaining the most accurate estimation of the dependent variable using maximum likelihood estimation funtion.

```
mtcars$cyl <- factor(mtcars$cyl)
mtcars$vs <- factor(mtcars$vs)
mtcars$gear <- factor(mtcars$gear)
mtcars$carb <- factor(mtcars$carb)
mtcars$am <- factor(mtcars$am)
initial_model <- lm(mpg ~ ., data = mtcars)
final_model <- step(initial_model, direction = "both",trace=0)
summary(final_model)
```

```
##
## Call:
## lm(formula = mpg ~ cyl + hp + wt + am, data = mtcars)
##
```

```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.9387 -1.2560 -0.4013  1.1253  5.0513
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 33.70832    2.60489  12.940 7.73e-13 ***
## cyl6        -3.03134    1.40728  -2.154  0.04068 *
## cyl8        -2.16368    2.28425  -0.947  0.35225
## hp          -0.03211    0.01369  -2.345  0.02693 *
## wt          -2.49683    0.88559  -2.819  0.00908 **
## am1          1.80921    1.39630   1.296  0.20646
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.41 on 26 degrees of freedom
## Multiple R-squared:  0.8659, Adjusted R-squared:  0.8401
## F-statistic: 33.57 on 5 and 26 DF,  p-value: 1.506e-10
```

The final model obtained from the above computations consists of the variables, cyl, wt and hp as confounders and am as the independent variable.

Model co-effcients

The co-efficients of cyl, hp and wt are negative while that of amManual is positive. This is intuitive as well. The higher the number of cylinders, the larger the weight or the larger the horse power the higher will be the consumption of fuel and hence lower mpg. Co-efficient for manual transmission is positive which is aligned with the box plots and t-test of mpg with transmission type shown in the previous sections.

Explanation of variation in MPG

we observe that the adjusted R square value is 0.84 which is the maximum obtained considering all combinations of variables. Thus we infer that more than 84% varaiance in mpg is explained by the above model.

In the following section, we compare the model with only am as the predictor variable and the final model which we obtained earlier containing confounder variables also.

```
base_model <- lm(mpg ~ am, data = mtcars)
```

Using ANOVA to test if variation in mpg can be explained using only tranmission type or if other variables help explain the variance significantly better.

```
anova(base_model, final_model)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ cyl + hp + wt + am
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     30 720.90
## 2     26 151.03  4    569.87 24.527 1.688e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We observe that the p-value is highly significant. Thus we conclude that number of cylinders, horse power and weight do significantly contribute to the accuracy of the model.

Residuals and diagnostics

We will now observe the residual plots of our regression model and also compute some of the regression diagnostics for our model since it is important to check the residuals for any signs of non-normality and examine the residuals vs. fitted values plot to spot for any signs of heteroskedasticity. Plots are shown in section 4 of appendix.

From the residual plots, we observe:

a) The points in the Residuals vs. Fitted plot are randomly scattered around the fitted line thereby verifying homoskedasticity.

b) The Normal Q-Q plot consists of the points which mostly fall on the line indicating that the residuals are normally distributed.

Conclusion

1) Cars with manual transmission run 1.8 more miles per gallon (mpg) than cars with automatic transmission. (adjusted by hp, cyl, and wt).

2) Miles per gallon decrease by 2.5 (adjusted by hp, cyl, and am) for every 1000 lb increase in wt.

3) Miles per gallon do not decrease significantly with increase in horse power

4) 4 cyliner engine gives 3 more miles per gallon than a 6 cyliner engine and 5 more miles that an 8 cyliner engine (adjusted by hp, wt, and am).

Assumptions

Following are the underlying assumptions behind the conclusions that we have made:

1) The data is representative of the number of miles per gallon obtained with different combinations of horse power, transmission, engine weight etc. The data is unbiased.

2) A significant variable impacting the variance of miles per gallon is not being ommitted. This assumption is justified by the fact that the error terms are homoskedastic.

Appendix

1) Data variables

We have the following variables in the data:

```
colnames(mtcars)
```

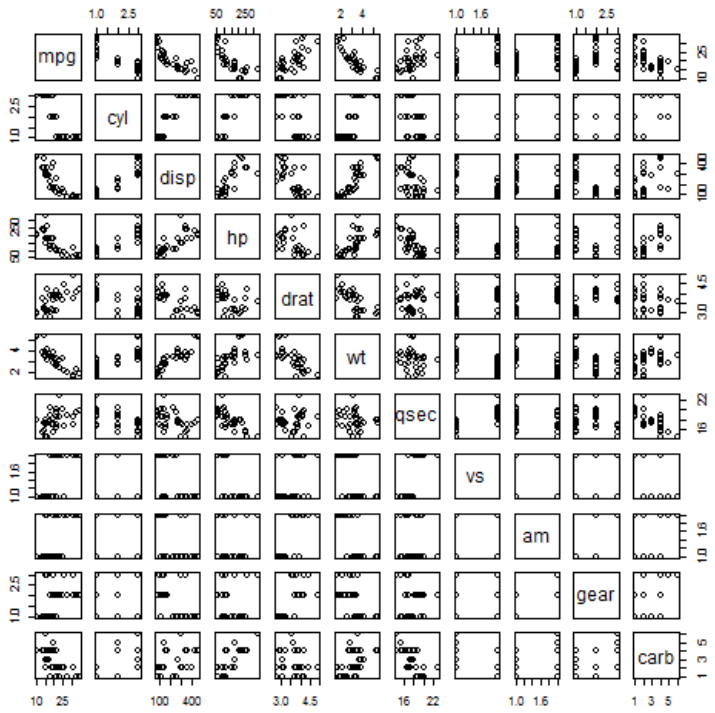
```
##  [1] "mpg"  "cyl"  "disp" "hp"   "drat" "wt"   "qsec" "vs"   "am"   "gear"
## [11] "carb"
```

Below is the description of the variables:

- mpg = Miles/gallon
- cyl = Number of cylinders
- disp = Displacement (cu.in.)
- hp = Gross horsepower drat = Rear axle ratio
- wt = Weight (lb/1000) qsec = ¼ mile time
- vs = V/S
- am = Transmission (0 = automatic, 1 = manual)
- gear = Number of forward gears
- carb = Number of carburetors.

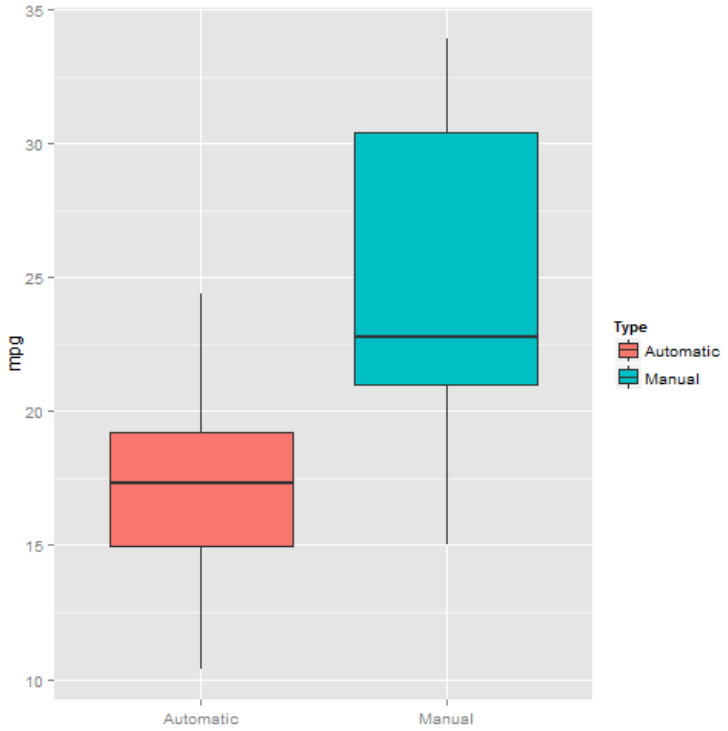
2) Visualising pairwise relation between all variables of the dataset

```
pairs(mtcars)
```



3) Box plot showing variation of mpg with transmission type

```
library(ggplot2)
g = ggplot(mtcars, aes(factor(am), mpg, fill=factor(am)))
g = g + geom_boxplot()
g = g + scale_colour_discrete(name = "Type")
g = g + scale_fill_discrete(name="Type", breaks=c("0", "1"),
            labels=c("Automatic", "Manual"))
g = g + scale_x_discrete(breaks=c("0", "1"), labels=c("Automatic", "Manual"))
g = g + xlab("")
g
```



4) Residual plots of regression model

```
par(mfrow = c(2,2))
plot(final_model)
```