Regression Models Course Project

Motor Trend Magazine - Data Analysis Report

Executive Summary

Looking at the dataset of a collection of cars, the following report explores the relationships between a set of variables and miles per gallon (**MPG**), providing information focusing on two main aspects:

(a) Is an automatic or manual transmission better for MPG

(b) Quantification of the MPG difference between automatic and manual transmissions

All the figures related to the exploratory data analysis are placed in the appendix of this document. First the difference between the automatic and manual transmission in terms of miles per gallon was examined. Results revealed a significant difference between the two groups with manual transmission to be related with an increase number of Miles/(US)gallon (Manual transmission:M=24.39, SD = 6.16; Automatic transmission:M = 17.15, SD = 3.83; p = 0.001374). Next, in order to define the significance of transmission type for MPG and quantify the difference between automatic and manual transmission, we fit and compare a range of regression models.Our analysis showed that cars with manual transmission are better in terms of miles per gallon compared to cars with automatic transmission (see also **figure 1** in Appendix).Based on the results of a linear regression model with transmission type as the only explanatory variable, a change from manual to automatic transmission will increase Miles/(US)gallon by 7.245. However, transmission type explains only 34% of variance.A multivariate linear regression model that includes all the variables in our dataset on the other hand explains 78% of variance. Using the R "step" function, we determined the best variables to include in our model (optimal model by AIC). In addition to transmission variable ("am"), the best subset of cofounder variables explaining our MPG data include:

(a) Number of cylinders ("cyl") **(b)** Gross horsepower ("hp") **(c)** Weight ("wt",lb/1000).

The proportion of variance explained by this model is 87% (see also figure 2). However, transmission type does not appear to be a good explanatory variable in our model, with its contribution not reaching significance, and explaining only about 12% of our mpg data. In contrast, number of cylinders, horsepower and weight appear to be significantly influential for the MPG data, and variables that should be consider over transmission for explaining MPG.Including or excluding the transmission type variable from the best-fit model did not show significant differences (p = 0.21), with the model explaining variance at about 86%.

1. Exploring the Dataset and Variable Transformations

```r
data(mtcars); dim(mtcars); head(mtcars,5)
```

```r
mtcars$cyl <- factor(mtcars$cyl);mtcars$vs <- factor(mtcars$vs); mtcars$gear <- factor(mtcars$gear); mtc
```

2. Statistical Inference

Assuming that transmission data have a normal distribution, we perform a t-test.

```r
t.test(mpg ~ am, data = mtcars)
```

```
##
##  Welch Two Sample t-test
```

```
##
## data:  mpg by am
## t = -3.7671, df = 18.332, p-value = 0.001374
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -11.280194  -3.209684
## sample estimates:
## mean in group 0 mean in group 1
##        17.14737        24.39231
```

```
by(mtcars$mpg, mtcars$am, sd)
```

3. Simple Linear Regression Model

In this part we start exploring the relationships between the variables. We start by fitting a simple linear regression model, with MPG as the outcome, and AM as the predictor. This model will also serve as our base for comparisons with multivariable regression models.

```
fit <- lm(mpg ~ am, data = mtcars); summary(fit)$coef; summary(fit)$r.squared
```

```
##             Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 17.147368   1.124603 15.247492 1.133983e-15
## am1          7.244939   1.764422  4.106127 2.850207e-04

## [1] 0.3597989
```

4. Multivariable Linear Regression Model

We fit a multivariable model including all variables in the dataset and mpg as the outcome. Using stepwise regression, we define the best subset of variables that fit our model.

```
#Stepwise Regression
multifit <- lm(mpg ~ ., data = mtcars)
library(MASS)
#stepwise model selection by exact AIC
bestfit <- stepAIC(multifit, direction = "both") #results are not presented
```

```
summary(bestfit)$coef; summary(bestfit)$r.squared
```

```
##                Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 33.70832390 2.60488618 12.940421 7.733392e-13
## cyl6        -3.03134449 1.40728351 -2.154040 4.068272e-02
## cyl8        -2.16367532 2.28425172 -0.947214 3.522509e-01
## hp          -0.03210943 0.01369257 -2.345025 2.693461e-02
## wt          -2.49682942 0.88558779 -2.819404 9.081408e-03
## am1          1.80921138 1.39630450  1.295714 2.064597e-01

## [1] 0.8658799
```

The p-value associated with the transmission type (am) variable is way above our alpha level (p = 0.206), suggesting a non-significant contribution of the transmission type variable in our model.Details related to the calculation of relative importance for each predictor in our model are depicted in **figure 2**

5. Model Comparisons

Comparisons between the base model and the best-fit model, in order to define the best model explaining mpg data.

```
anova(fit, bestfit)
```

Results suggest that the multivariable model is more accurate compared to the simple model.Considering the non-significant contribution of our transmission type, it would be wise to see also differences between our best-fit model and a new model that does not contain the transmission type variable

```
newfit <- lm(mpg ~ cyl + hp + wt, data = mtcars ); summary(newfit)
anova(bestfit, newfit); anova(fit,newfit)
```

6. Residuals and Diagnostics

The related plots can be found in the Appendix (**figure 3**).As shown in the **Residuals vs Fitted plot** data are randomly scattered on the indicating independence of the variable. The curve indicates a slight diversion from normality in the data.

Appendix

figure 1 Boxplot showing the difference in MPG by transmission type (Automatic vs Manual)

```
boxplot(mpg ~ am, data = mtcars, col = (c("green","blue")), ylab = "Miles Per Gallon", xlab = "Transmis
```

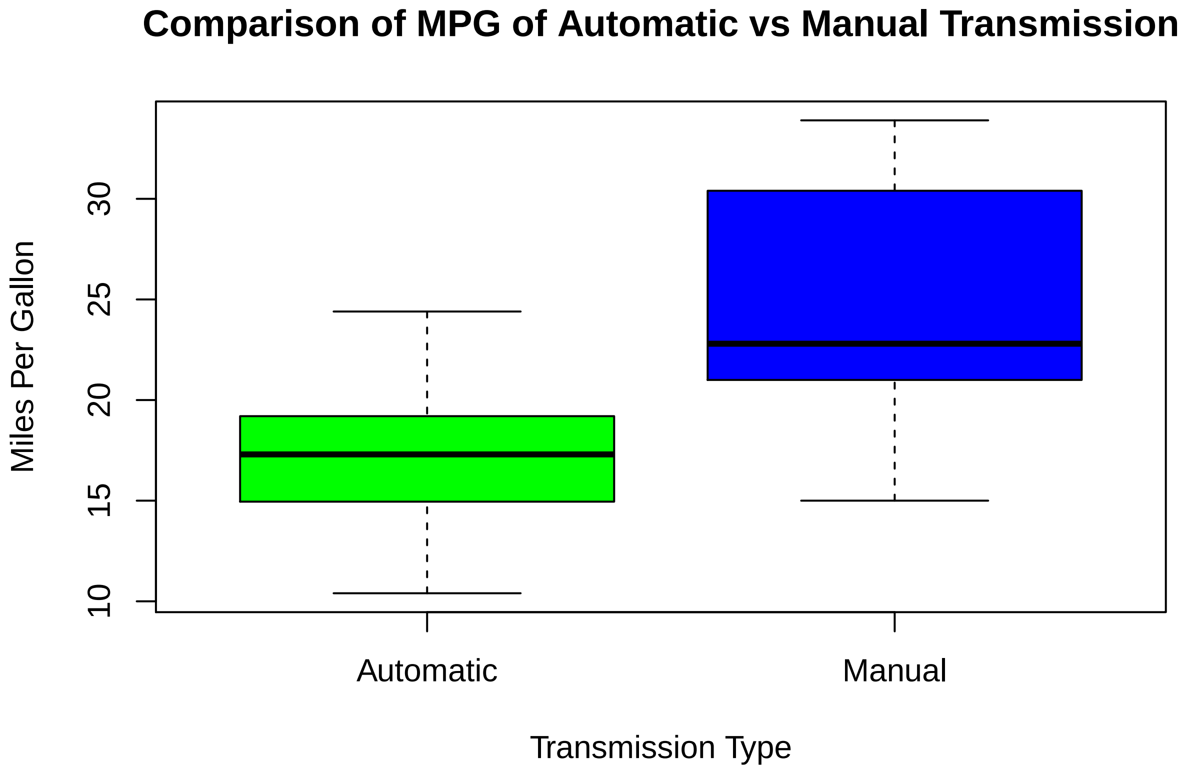


figure 2

Plot showing the accountability of each predictor in our model

```r
library(relaimpo)
calc.relimp(bestfit)
```

```r
plot(calc.relimp(bestfit))
```

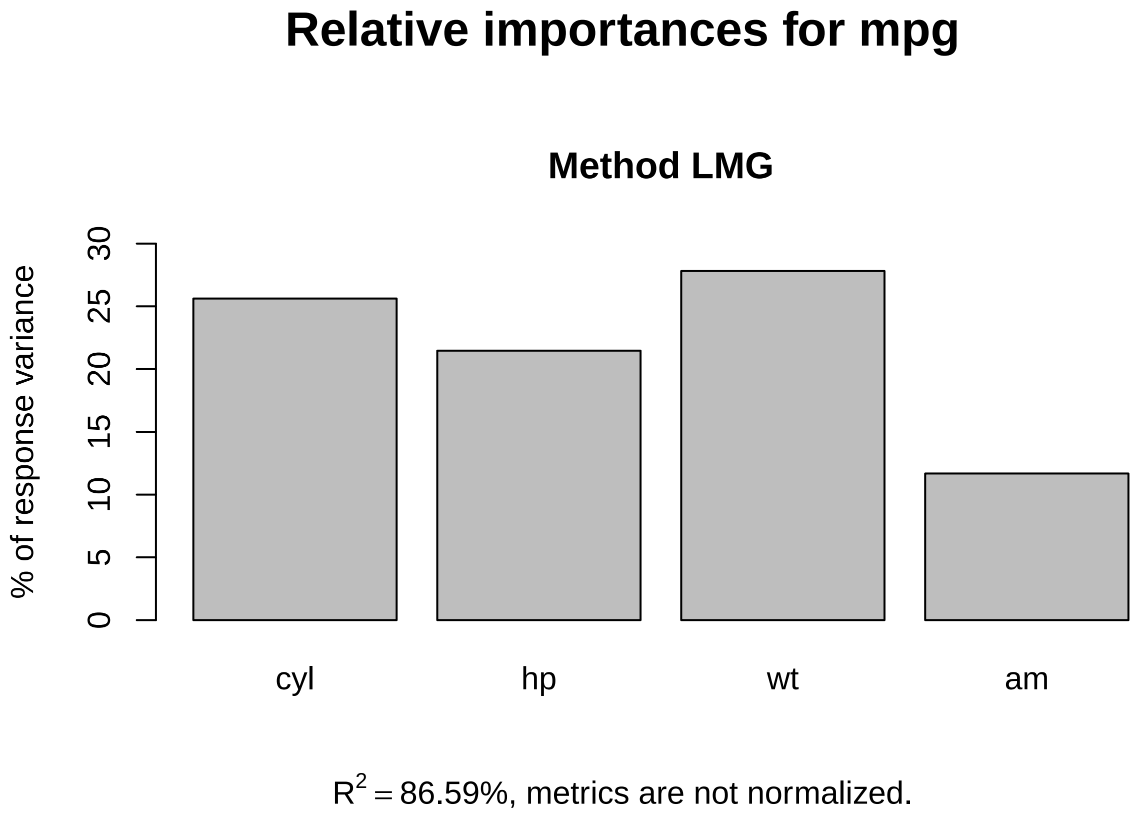
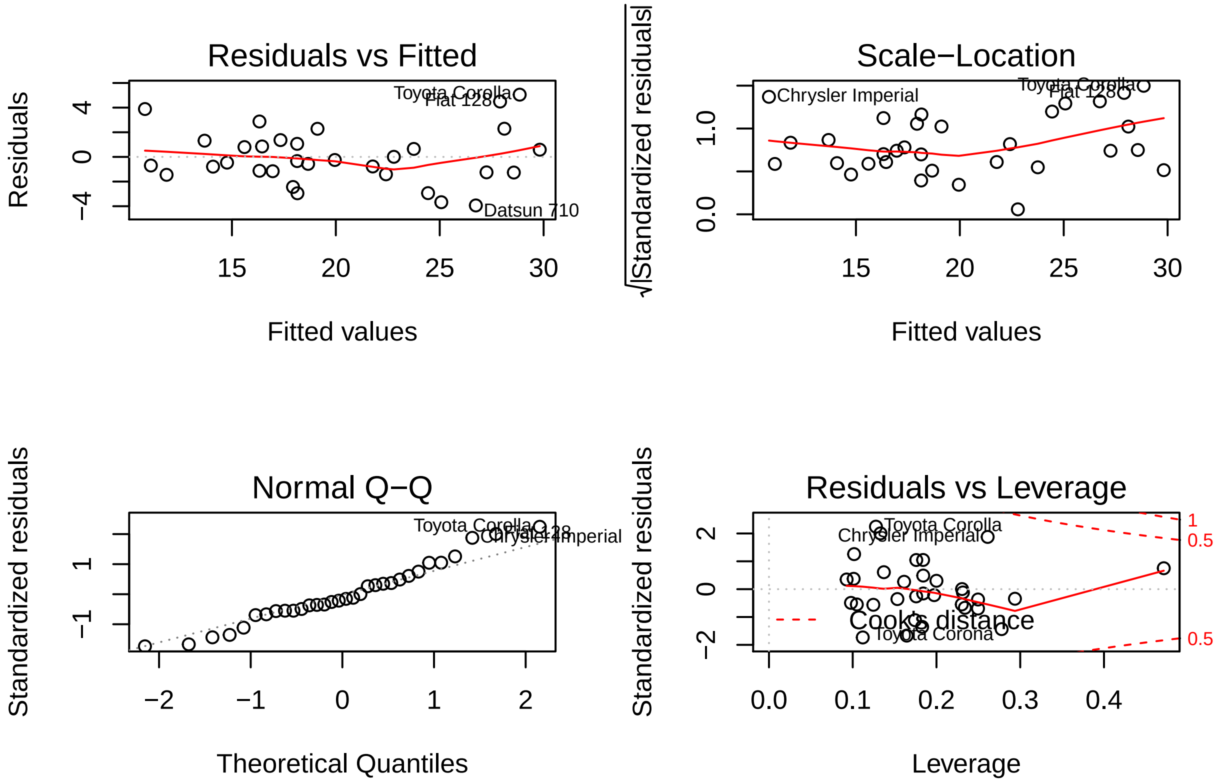


figure 3

```r
# diagnostic plots
layout(matrix(c(1,2,3,4),2,2))
plot(bestfit)
```


Residuals vs Fitted
Residuals
Fitted values
4
0
−4
15
20
25
30
Toyota Corolla
Fiat 128
Datsun 710
Scale−Location
√|Standardized residuals|
Fitted values
0.0
1.0
15
20
25
30
Chrysler Imperial
Toyota Corolla
Fiat 128
Normal Q−Q
Standardized residuals
Theoretical Quantiles
−1
1
−2
−1
0
1
2
Toyota Corolla
Fiat 128
Chrysler Imperial
Residuals vs Leverage
Standardized residuals
Leverage
−2
0
2
0.0
0.1
0.2
0.3
0.4
Toyota Corolla
Chrysler Imperial
Toyota Corona
Cook's distance
1
0.5
0.5