Regression Models Course Project

Vijay Ramanujam

February 10, 2018

Executive Summary

Motor Trend, a magazine about the automobile industry, looking at a data set of a collection of cars, and are interested in exploring the relationship between a set of variables and miles per gallon (MPG) (outcome). They are particularly interested in the following two questions:

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

In this analysis, we'll be looking into 2 models. The first model explores the relationship between transmission and MPG as requested by the management. The second one, output of stepwise regression, includes other regressors leading to a better model. Based on our analysis, cars with manual transmissions, on an average, give 1.8 miles more per gallon compared to cars with automatic transmission.

Exploratory Analysis

We begin the analysis by loading the dataset and looking into the structure.

```
library(broom)
data("mtcars")
corrmtcars <- cor(mtcars) #dataset used later in plots
mtcars$am   <- factor(mtcars$am, labels=c("Automatic", "Manual"))
mtcars$cyl  <- factor(mtcars$cyl, labels=c("4 cyl", "6 cyl", "8 cyl"))
mtcars$vs   <- factor(mtcars$vs, labels=c("V engine", "Straight engine"))
mtcars$gear <- factor(mtcars$gear, labels=c("3 gears", "4 gears", "5 gears"))
mtcars$carb <- factor(mtcars$carb)
```

The correlation plot between all variables shows strong relationship between mpg and the variables cyl, disp, wt, hp, vs, drat, and am. Please see Appendix Fig (i) for the plot.

```
library(corrplot)
par(mfrow=c(1, 2))
corrplot(corrmtcars, method = "ellipse", type="upper", title="Fig (i)", mar=c(0,0,1,0))
```

The boxplot clearly shows that manual transmission results in more miles/gal compared to automatic transmission type. Please see Appendix Fig (ii) for the plot.

```
boxplot(mpg ~ am, data=mtcars, xlab="Transmission", ylab="MPG", main="Fig (ii)", col=(c("red","blue")))
```

Regression Analysis

Let's analyse the data further by fitting a simple linear regression model between mpg (dependent) and am (independent). This is the base model.

```
fit <- lm(mpg ~ am, mtcars)
tidy(fit)
glance(fit)[,1:7]
```

```
##          term  estimate std.error statistic      p.value
## 1 (Intercept) 17.147368  1.124603 15.247492 1.133983e-15
## 2    amManual  7.244939  1.764422  4.106127 2.850207e-04
```

```
##   r.squared adj.r.squared    sigma statistic      p.value df    logLik
## 1 0.3597989     0.3384589 4.902029  16.86028 0.0002850207  2 -95.24219
```

The above table shows that the model is statistically significant (p-value 0.000285) and is capable of explaining 33.85% of the variability in mpg using am (p-value 0.000285). It also shows that manual transmission increases the MPG by 7.25 miles compared to automatic trasmission. Since there are other variables in the dataset, stepwise regression technique can be used to find the best fitting model by comparing all the variables. This can be achieved by the following R code.

```
fit2 <- step(lm(mpg ~ ., mtcars), direction = "both")
summary(fit2)$call
summary(fit2)$coefficients
glance(fit)[,1:7]
```

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

```
##                Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 33.70832390 2.60488618 12.940421 7.733392e-13
## cyl6 cyl    -3.03134449 1.40728351 -2.154040 4.068272e-02
## cyl8 cyl    -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
## amManual     1.80921138 1.39630450  1.295714 2.064597e-01
```

```
##   r.squared adj.r.squared   sigma statistic      p.value df    logLik
## 1 0.8658799     0.8400875 2.41012  33.57121 1.505607e-10  6 -70.23345
```

The above table shows that the model is statistically significant (p-value < 0.05) and is capable of explaining 84% of variability in mpg using the variables cyl, hp, wt, and am. Let's call it best model.

Now, we compare the base model with best model.

```
anova(fit, fit2)
```

```
## 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
```

Here, the p-value of the best model is very significant (i.e. < 0.05, close to 0) compared to the base model. Also, RSS, residual sum of squares, is very low compared to the base model. RSS is used to measure the amount of variance in a data set that is not explained by a regression model.

Residual Plots and Diagnostics

In this section, we plot the residuals and do some analysis for any non-normality. Please see Appendix Fig (iii) for the plot.

```
par(mfrow=c(2, 2))
plot(fit2, main="Fig (iii)")
```

- The Residuals vs. Fitted plot shows that the points are randomly scattered on the plot verifying the independence condition.
- The Normal Q-Q plot shows points which mostly fall on the line indicating that the residuals are normally distributed.
- The Scale-Location plot consists of points scattered in a constant band pattern indicating constant variance.
- The Residuals vs. Leverage plot shows some distinct points (outliers/leverage points) in the top right indicating values of increased leverage of outliers.

Now, we do some regression diagnostics of the model to find out those leverage points. Using hatvalues() function, we can find out the top three points in each case that influence the model coefficients the most resulting from dfbetas() function.

```
leverage <- hatvalues(fit2)
head(sort(leverage, decreasing=TRUE), 3)
```

```
##       Maserati Bora Lincoln Continental       Toyota Corona
##           0.4713671           0.2936819           0.2777872
```

```r
influential <- dfbetas(fit2)
tail(sort(influential[,6]),3)
```

```
## Chrysler Imperial          Fiat 128     Toyota Corona
##         0.3507458         0.4292043         0.7305402
```

The above tables show that our analysis was correct, as the same cars are mentioned in the residual plots.

Inference

Both base and best models showed manual transmission yielding more MPG compared to automatic. To prove, let's perform a t-test on automatic and manual transmission data. Here, we are testing the null hypothesis stating both automatic and manual transmissions data have the same MPG mean.

```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 Automatic    mean in group Manual
##                17.14737                24.39231
```

Based on the results, we reject the null hypothesis that both automatic and manual transmissions data have the same MPG mean.

Conclusions

- Manual transmission cars give better mileage (mpg) compared to automatic transmission.
- On an average, cars with manual transmission give 1.8 miles more per gallon holding all the other variables (cyl, hp, wt) constant.
- MPG will decrease by 2.5 miles for every 1000lb increase in weight of the car holding all the other variables (cyl, hp, am) constant.
- MPG decreases negligibly by 0.03 miles for every single horsepower increase holding all the other variables (cyl, am, wt) constant.
- 4 cylinder cars give 3 MPG more than 6 cylinder cars and 6 cylinder cars give 2.2 MPG more than 8 culinder cars holding all the other variables (am, hp, wt) constant.
- The above conclusions are based on 95% certainty.

Appendix

Fig (i)

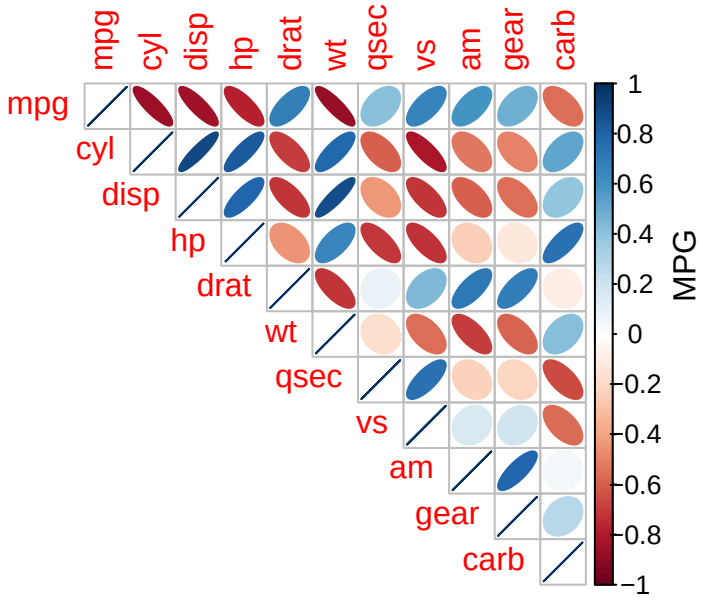

Fig (ii)

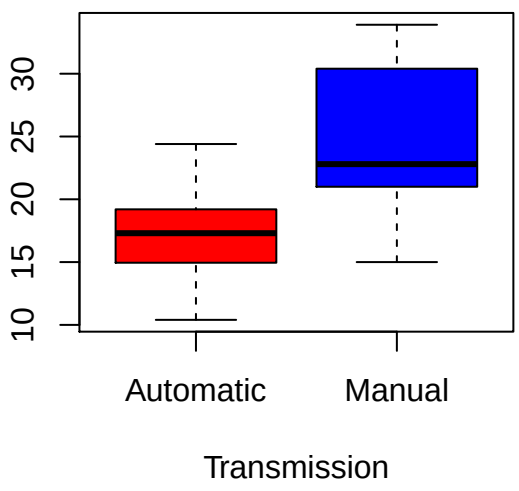

Fig (iii)
Residuals vs Fitted

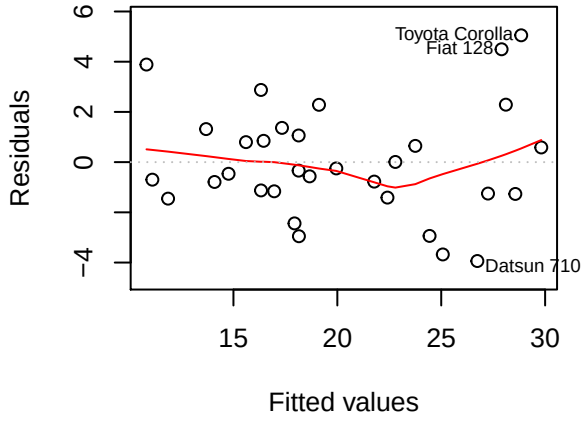

Fig (iii)
Normal Q−Q

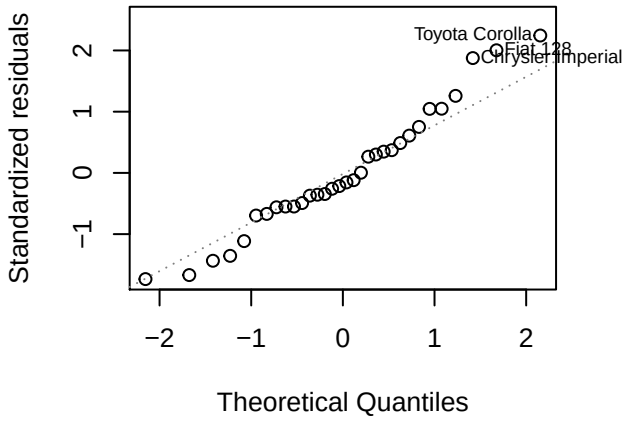

Fig (iii)
Scale−Location

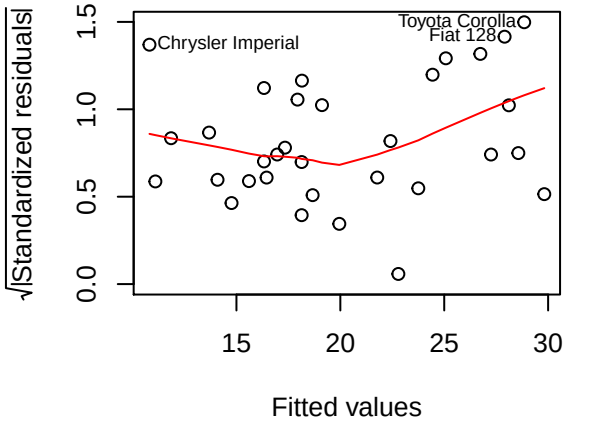

Fig (iii)
Residuals vs Leverage