Regression Models Course Project

Vamshi Krishna

7/12/2020

Executive Summary

Motor Trend, a magazine about the automobile industry. Below is the data analysis which can answer a question with suitable data. 1. "Is an automatic or manual transmission better for MPG" 2. "Quantify the MPG difference between automatic and manual transmissions"

The below analysis tells us the answers are: 1. The automatic or manual transmission when compared to MPG for the dataset are not statistically significant. 2. The MPG difference is 1.8 for automatic vs manual transmissions.

Loading and preprocessing data

Lets load the data and print the head of the datarame and get an idea of columns and data within the dataframe.

```
data(mtcars)
head(mtcars)
```

```
##                   mpg cyl disp  hp drat    wt  qsec vs am gear carb
## Mazda RX4         21.0   6  160 110 3.90 2.620 16.46  0  1    4    4
## Mazda RX4 Wag     21.0   6  160 110 3.90 2.875 17.02  0  1    4    4
## Datsun 710        22.8   4  108  93 3.85 2.320 18.61  1  1    4    1
## Hornet 4 Drive    21.4   6  258 110 3.08 3.215 19.44  1  0    3    1
## Hornet Sportabout 18.7   8  360 175 3.15 3.440 17.02  0  0    3    2
## Valiant           18.1   6  225 105 2.76 3.460 20.22  1  0    3    1
```

We can see that the rownames are the model names and other variables/features are in columns

Lets see the variables types and size of the data

```
print(str(mtcars))
```

```
## 'data.frame':    32 obs. of  11 variables:
##  $ mpg : num  21 21 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 ...
##  $ cyl : num  6 6 4 6 8 6 8 4 4 6 ...
##  $ disp: num  160 160 108 258 360 ...
##  $ hp  : num  110 110 93 110 175 105 245 62 95 123 ...
##  $ drat: num  3.9 3.9 3.85 3.08 3.15 2.76 3.21 3.69 3.92 3.92 ...
##  $ wt  : num  2.62 2.88 2.32 3.21 3.44 ...
##  $ qsec: num  16.5 17 18.6 19.4 17 ...
```

```
##  $ vs  : num  0 0 1 1 0 1 0 1 1 1 ...
##  $ am  : num  1 1 1 0 0 0 0 0 0 0 ...
##  $ gear: num  4 4 4 3 3 3 3 4 4 4 ...
##  $ carb: num  4 4 1 1 2 1 4 2 2 4 ...
## NULL
```

```
print(dim(mtcars))
```

```
## [1] 32 11
```

It is good that all the variables are in number type with 32 rows and 11 columns

Data cleaning and EDA

By converting some of the variables to factors from numbers, it will be helpful for analysis.

```
columns = c('cyl','vs','am','gear','carb')
mtcars[,columns] <- lapply(mtcars[,columns] , factor)
str(mtcars)
```

```
## 'data.frame':    32 obs. of  11 variables:
##  $ mpg : num  21 21 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 ...
##  $ cyl : Factor w/ 3 levels "4","6","8": 2 2 1 2 3 2 3 1 1 2 ...
##  $ disp: num  160 160 108 258 360 ...
##  $ hp  : num  110 110 93 110 175 105 245 62 95 123 ...
##  $ drat: num  3.9 3.9 3.85 3.08 3.15 2.76 3.21 3.69 3.92 3.92 ...
##  $ wt  : num  2.62 2.88 2.32 3.21 3.44 ...
##  $ qsec: num  16.5 17 18.6 19.4 17 ...
##  $ vs  : Factor w/ 2 levels "0","1": 1 1 2 2 1 2 1 2 2 2 ...
##  $ am  : Factor w/ 2 levels "0","1": 2 2 2 1 1 1 1 1 1 1 ...
##  $ gear: Factor w/ 3 levels "3","4","5": 2 2 2 1 1 1 1 2 2 2 ...
##  $ carb: Factor w/ 6 levels "1","2","3","4",..: 4 4 1 1 2 1 4 2 2 4 ...
```

Lets see the relation of linear regression between automatic/manual transmissions vs MPG. the initial assumptions of the model are appropriate mean value and error should be distributed in normal and independent.

```
fit1 = lm(mpg~am,mtcars)
summary(fit1)
```

```
##
## Call:
## lm(formula = mpg ~ am, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.3923 -3.0923 -0.2974  3.2439  9.5077
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   17.147      1.125  15.247 1.13e-15 ***
```

```
## am1            7.245      1.764   4.106 0.000285 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.902 on 30 degrees of freedom
## Multiple R-squared:  0.3598, Adjusted R-squared:  0.3385
## F-statistic: 16.86 on 1 and 30 DF,  p-value: 0.000285
```

```
#figure1 from appendix
```

lets now look at confidence interval using the formula we know

```
beta_hat = 0.7245
standard_error = 1.764
t_star = qt(1-0.05/2,df = length(mtcars$mpg)-2)
c(beta_hat-t_star*standard_error, beta_hat+t_star*standard_error)
```

```
## [1] -2.878069  4.327069
```

From both the plots in figure1, the results of our coefficient summary, small p-value and presence of 0 in CI, we reject the null hypothesis that transmissions affects MPG.

Multivariate Analysis

Lets now include new variables to increase the standard errors of coefficient estimates of other coorelated regressors.

Lets now create a fit for all the other variables with MPG

```
fit2 = lm(mpg~. , mtcars)
```

Now, we determine which variables are important fot the correlation using stepAIC function.

```
library('MASS')
step = stepAIC(fit2, direction="both", trace=FALSE)
summary(step)
```

```
##
## 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 significant variables from the summary in relation to the MPG are cylinders(cyl),horsepower(hp),weight(wt).

Comparing the models:

To conclude that the secoind model using step is better than the first model, we use ANOVA to compare the models.

```
anova(fit1,step)
```

```
## 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
```

The difference in both models is the significant. Lets plot the best fit model to view it clearly.(see figure2 from appendix)

Finding significance of transmission type vs mpg

Lets summarize the coeffients of the best model:

```
coefficients(summary(step))
```

```
##                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
```

Looking at the p-values in summary data, we can see that the p-value from transmissions vs MPG is not significant. We can prove this using confidence interval as we did above.

```r
beta_hat1 = 1.8092
standard_error1 = 1.3963
t_star1 = qt(1-0.05/2, df = length(mtcars$mpg)-2)
c(beta_hat1-t_star1*standard_error1,beta_hat1+t_star1*standard_error1)
```

```
## [1] -1.042425  4.660825
```

Since the CI includes 0 and the p-value is .0.5, the difference between automatic transmission and a manual trnasmission does not affect the MPG significantly. However auto-trnasmission is 1.8 greater than the manual.

Conclusion

By reviewing the above models, the best fit from figure2 shows that the normal Q-Q graph is normally distributed and the scale-location graph has a steady variance which is good when compared to figure1. and finally we can conclude that am does not have a significant impact on mpg.

Appendix

Figure 1:

```r
par(mfrow=c(2,2))
plot(fit1)
abline(fit1)
```

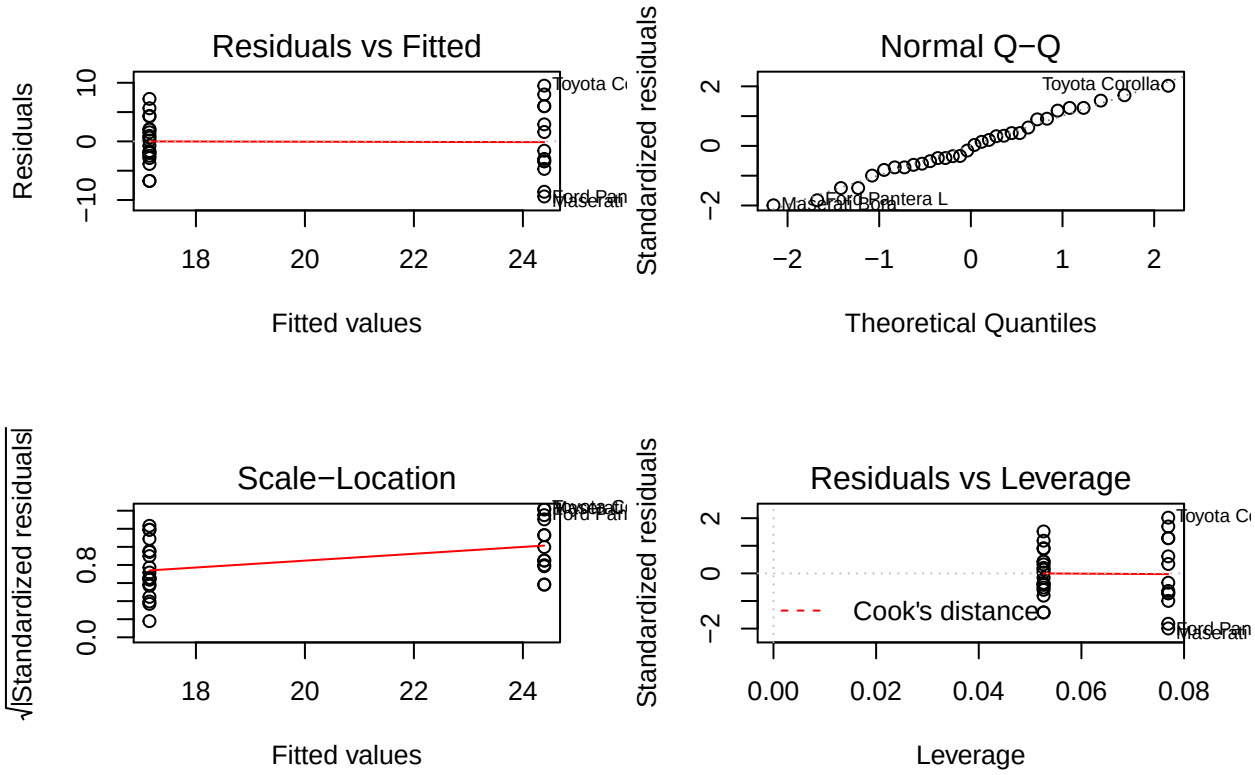
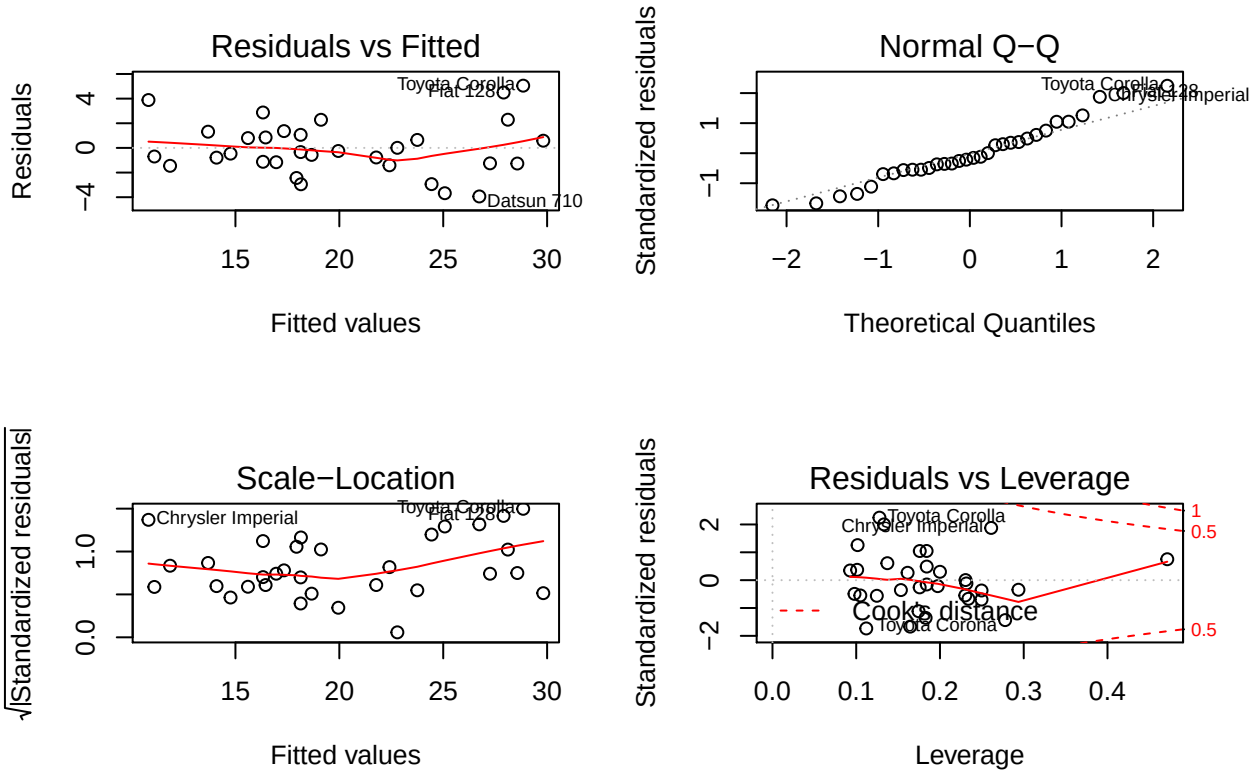


Figure 2:

```
par(mfrow=c(2,2))
plot(step)
abline(step)
```

```
## Warning in abline(step): only using the first two of 6 regression coefficients
```


Residuals vs Fitted
Residuals
Fitted values
Toyota Corolla
Fiat 128
Datsun 710
Normal Q−Q
Standardized residuals
Theoretical Quantiles
Toyota Corolla
Chrysler Imperial
Scale−Location
√|Standardized residuals|
Fitted values
Chrysler Imperial
Toyota Corolla
Fiat 128
Residuals vs Leverage
Standardized residuals
Leverage
Cook's distance
Toyota Corolla
Chrysler Imperial
Toyota Corona