Regression Models Course Project

Shivangi

9/1/2020

INTRODUCTION

You work for Motor Trend, a magazine about the automobile industry. Looking at a data set of a collection of cars, they 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:

1. "Is an automatic or manual transmission better for MPG"

2. "Quantify the MPG difference between automatic and manual transmissions"

LOADING THE DATASET

Take the {`mtcars`}mtcars data set and write up an analysis to answer their question using regression models and exploratory data analyses.

```
carsdata <- mtcars
head(carsdata)
```

```
##                   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
```

```
summary(carsdata)
```

```
##       mpg             cyl             disp             hp
##  Min.   :10.40   Min.   :4.000   Min.   : 71.1   Min.   : 52.0
##  1st Qu.:15.43   1st Qu.:4.000   1st Qu.:120.8   1st Qu.: 96.5
##  Median :19.20   Median :6.000   Median :196.3   Median :123.0
##  Mean   :20.09   Mean   :6.188   Mean   :230.7   Mean   :146.7
##  3rd Qu.:22.80   3rd Qu.:8.000   3rd Qu.:326.0   3rd Qu.:180.0
##  Max.   :33.90   Max.   :8.000   Max.   :472.0   Max.   :335.0
##       drat             wt             qsec             vs
##  Min.   :2.760   Min.   :1.513   Min.   :14.50   Min.   :0.0000
##  1st Qu.:3.080   1st Qu.:2.581   1st Qu.:16.89   1st Qu.:0.0000
##  Median :3.695   Median :3.325   Median :17.71   Median :0.0000
```

```
##  Mean   :3.597   Mean   :3.217   Mean   :17.85   Mean   :0.4375
##  3rd Qu.:3.920   3rd Qu.:3.610   3rd Qu.:18.90   3rd Qu.:1.0000
##  Max.   :4.930   Max.   :5.424   Max.   :22.90   Max.   :1.0000
##        am              gear            carb
##  Min.   :0.0000   Min.   :3.000   Min.   :1.000
##  1st Qu.:0.0000   1st Qu.:3.000   1st Qu.:2.000
##  Median :0.0000   Median :4.000   Median :2.000
##  Mean   :0.4062   Mean   :3.688   Mean   :2.812
##  3rd Qu.:1.0000   3rd Qu.:4.000   3rd Qu.:4.000
##  Max.   :1.0000   Max.   :5.000   Max.   :8.000
```

CLEANING DATA

```
carsdata[,'am']<-as.factor(carsdata[,'am'])
carsdata[,'cyl']<-as.factor(carsdata[,'cyl'])
carsdata[,'vs']<-as.factor(carsdata[,'vs'])
carsdata[,'gear']<-as.factor(carsdata[,'gear'])
carsdata[,'carb']<-as.factor(carsdata[,'carb'])
```

REGRESSION MODELLING

Considering four models that can play a substantial role in affecting the mpg outcome of the cars.

```
mdl1 <- lm(mpg~am,data=carsdata)

mdl2 <- lm(mpg~.,data=carsdata)

coef(mdl1)
```

```
## (Intercept)         am1
##   17.147368    7.244939
```

```
coef(mdl2)
```

```
## (Intercept)        cyl6        cyl8        disp          hp        drat
## 23.87913244 -2.64869528 -0.33616298  0.03554632 -0.07050683  1.18283018
##          wt        qsec         vs1         am1       gear4       gear5
## -4.52977584  0.36784482  1.93085054  1.21211570  1.11435494  2.52839599
##       carb2       carb3       carb4       carb6       carb8
## -0.97935432  2.99963875  1.09142288  4.47756921  7.25041126
```

```
summary(mdl1)
```

```
##
## Call:
## lm(formula = mpg ~ am, data = carsdata)
##
## 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
```

```
confint(mdl1)
```

```
##                2.5 %   97.5 %
## (Intercept) 14.85062 19.44411
## am1          3.64151 10.84837
```

Let's just begin with a simple linear regression of MPG vs automatic/manual transmissions. The assumptions are that the linear model is appropriate for the mean value of yi, and the error distribution should be normally distributed and independent.

From both the plots in Figure one, the results of our coefficient summary, small p-value, and exclusion of 0 in the confidence internval, we fail to reject the null hypothesis that tranmission affects MPG.

MULTIVARIATE ANALYSIS

While we are exploring MPG for manual vs automatic transmissions, we know that including new variables will increase standard errors of coefficient estimates of other correlated regressors. However, we need to be cautious not to omit variables because that will result in bias in coefficients of regressors which are correlated to the omitted variables.

```
summary(mdl2)
```

```
##
## Call:
## lm(formula = mpg ~ ., data = carsdata)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.5087 -1.3584 -0.0948  0.7745  4.6251
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 23.87913   20.06582   1.190   0.2525
## cyl6        -2.64870    3.04089  -0.871   0.3975
## cyl8        -0.33616    7.15954  -0.047   0.9632
## disp         0.03555    0.03190   1.114   0.2827
## hp          -0.07051    0.03943  -1.788   0.0939 .
## drat         1.18283    2.48348   0.476   0.6407
## wt          -4.52978    2.53875  -1.784   0.0946 .
```

```
## qsec         0.36784    0.93540   0.393   0.6997
## vs1          1.93085    2.87126   0.672   0.5115
## am1          1.21212    3.21355   0.377   0.7113
## gear4        1.11435    3.79952   0.293   0.7733
## gear5        2.52840    3.73636   0.677   0.5089
## carb2       -0.97935    2.31797  -0.423   0.6787
## carb3        2.99964    4.29355   0.699   0.4955
## carb4        1.09142    4.44962   0.245   0.8096
## carb6        4.47757    6.38406   0.701   0.4938
## carb8        7.25041    8.36057   0.867   0.3995
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.833 on 15 degrees of freedom
## Multiple R-squared:  0.8931, Adjusted R-squared:  0.779
## F-statistic:  7.83 on 16 and 15 DF,  p-value: 0.000124
```

According to this summary, the significant variables in relation to the mpg are cyl(cylinders), hp(horsepower) and wt(weight)

COMPARISON OF MDL1 AND MDL2

To verify that the multivariate model from step is the better fit, use ANOVA to compare the two models.

```
anova(mdl1,mdl2)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ cyl + disp + hp + drat + wt + qsec + vs + am + gear + carb
##   Res.Df   RSS Df Sum of Sq      F   Pr(>F)
## 1     30 720.9
## 2     15 120.4 15    600.49 4.9874 0.001759 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Upon comparing the two models, the difference is significant, leading us to rule out the simpler model.

Visually, it may be easier to see if we plot this best fit model. This is done in Figure 2 in the Appendix

```
confint(mdl2)[9,]
```

```
##     2.5 %    97.5 %
## -4.189091  8.050792
```

Since the confidence interval includes 0 and the p-value is greater than .05, the difference between an automatic transmission and a manual transmission does not significantly impact mpg(miles per gallon). It does however show that an automatic transmission is greater than a manual transmission.

CONCLUSION

Upon review of the the models, the best fit in Figure 2, it is shown that the Normal Q-Q graph is normally distributed and the Scale-Location graph has a a steady variance. This is improved from Figure 1 where only am(transmission type) was compared with mpg. Upon further review, it was determined that am did not have a significant impact on mpg.

APPENDIX

FIGURE 1

```
par(mfrow=c(2,2))
plot(mdl1);
abline(mdl1)
```

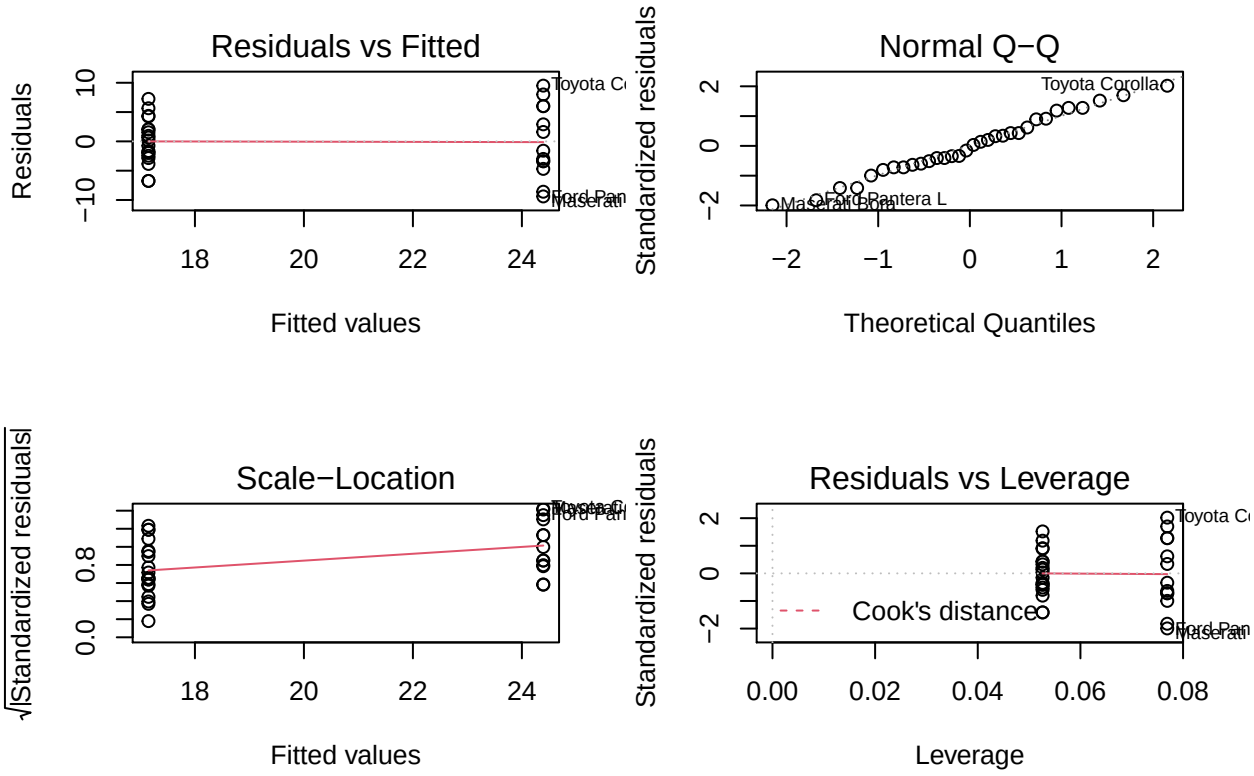

FIGURE 2

```
par(mfrow=c(2,2))
plot(mdl2);
```

```
## Warning: not plotting observations with leverage one:
##   30, 31
```

```r
abline(mdl2$intercept)
```