Regression Models Course Project

Week 4

20/02/2018

Overview

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"

Data summary

Basic summary of the data can be viewed in `Appendix 1`.

From exploring the variable types, the variables `am`, `cyl`, `vs`, `gear` and `card` have been convert to factor. Please note that `am` refers to Transmission(0 = automatic, 1 = manual). `mpg` ranges from 10.40 to 33.90 miles/(US)galon. There are cars in this dataset with 4, 6 or 8 cylinders'. Minimum and maximum displacement are 77.1 and 472 cu.in, respectively. Gross horsepower ranges from 52 to 335. There are 3, 4 or 5 forward gears.

Data processing

`am`, `cyl`, `vs`, `gear` and `carb` have been converted to factor using the `factor` function. Please refer to `Appendix 1` to view the code.

Exploratory Analysis

Pair plots including all variables can be found in the `Appendix 1`

Mean of MPG for each transmission type `am`.

```
aggregate(mpg ~ am, data = mtcars, mean)
```

```
##   am      mpg
## 1  0 17.14737
## 2  1 24.39231
```

Manual transmission get more MPG compared to Automatic transmission. Please refer to `Appendix 2` for graphical visualisation.

Based on the t-test results, reject the H_0: mpg distributions for manual and automatic transmissions are the same. See `Appendix 4`.

Regression analysis

Linear regression models will be build in roder to find the best fit.

1. Linear Regression Model regressing all variables against MPG.
2. Use of Linear Regression with STEP function. This will call the MPG variable against all variables available to see which variables affect MPG the most. This will ensure inclusion of variables that are relevant and omission of the ones that aren't.
3. Linear regression model only using `am` as regressor.
4. Compare model in `3` against `2` using ANOVA.

```
fit1 <- lm(mpg ~., data = mtcars);fit2 <- step(fit1, direction = "both",  trace=FALSE)
fit3 <- lm(mpg ~ am, data = mtcars)
```

Summary of the model can be viewed in `Appendix 3`

The adjusted R-squared are 0.7790215, 0.8400875, 0.3384589, for `fit1`, `fit2` and `fit3` repectively. R-suqred for `fit2` is the maximum obtained considering all combinations of variables. It is possible to conclude that 84% of the variability is explained by `fit2`.

```
anova(fit3,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
```

As shown above, the p-value obtained is highly significant and the null hypothesis is rejected. H_0 is the hypotesys that the confounder variables cyl, hp and wt don't contribute to the accuracy of the model.

Conclusion

Based on the analysis performed:

```
## [1] 33.70832390 -3.03134449 -2.16367532 -0.03210943 -2.49682942  1.80921138
```

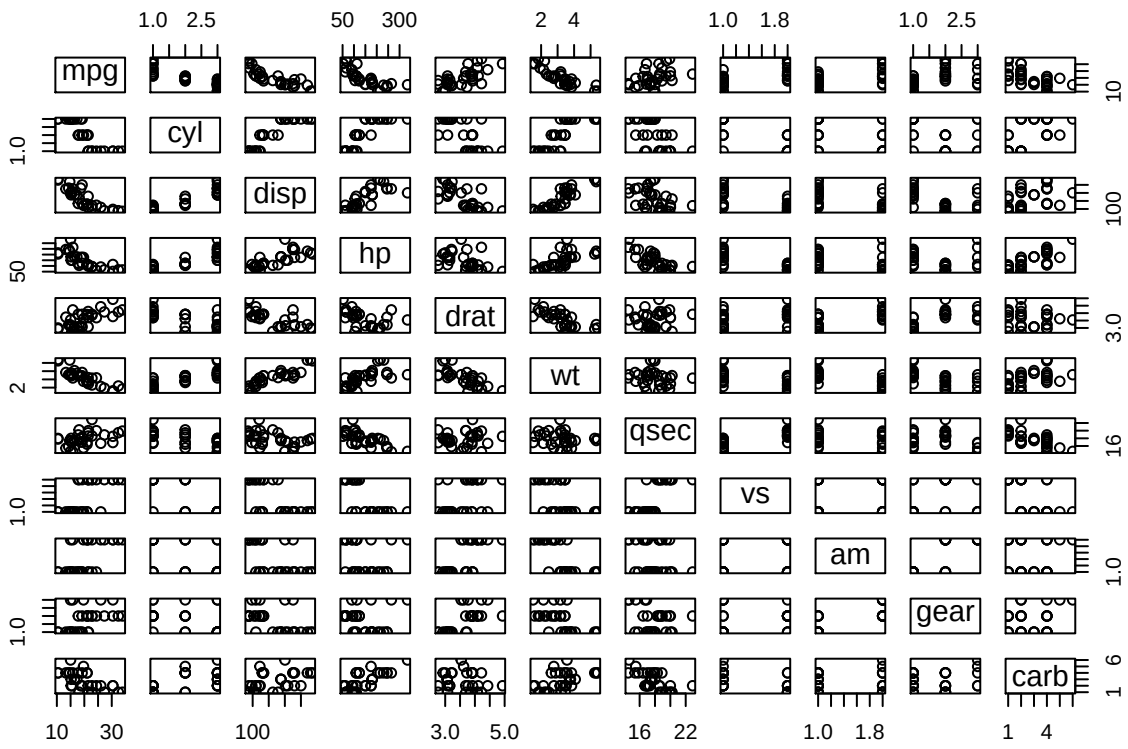
Manual transmission get 1.8 more miles per gallon compared to cars with Automatic transmission. (1.8 adjusted for hp, cyl, and wt). mpg decreases by 0.32 with every increase of 10 in hp.mpg will decrease by 2.5 for every 1000 lb increase in wt. If number of cylinders increase, mpg will decrease(adjusted by hp, wt, and am).

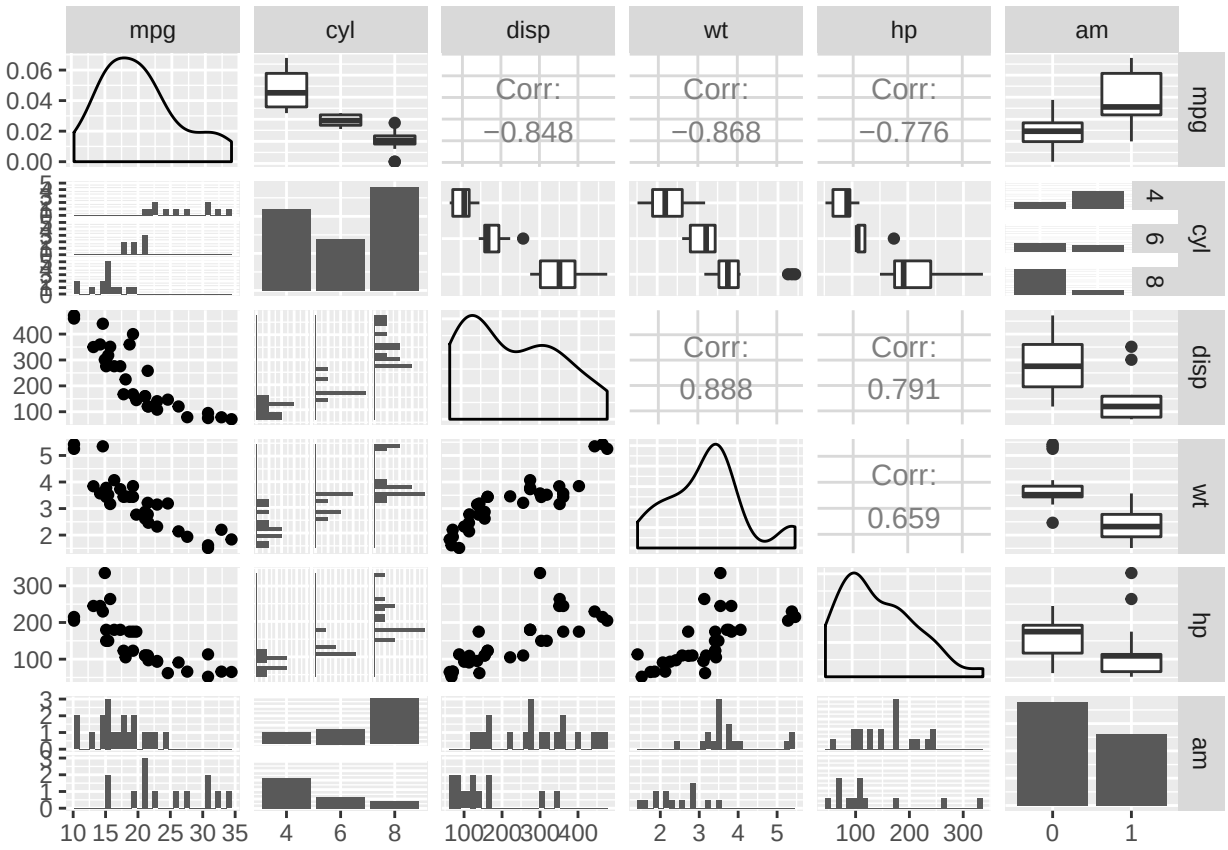
Appendices

Appendix 1

Plot the relationship between all variables in `mtcars`.

```
pairs(mtcars)
```





Data summary

```
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 ...
```

```
summary(mtcars)
```

```
##       mpg        cyl         disp             hp             drat
##  Min.   :10.40   4:11   Min.   : 71.1   Min.   : 52.0   Min.   :2.760
##  1st Qu.:15.43   6: 7   1st Qu.:120.8   1st Qu.: 96.5   1st Qu.:3.080
##  Median :19.20   8:14   Median :196.3   Median :123.0   Median :3.695
##  Mean   :20.09          Mean   :230.7   Mean   :146.7   Mean   :3.597
##  3rd Qu.:22.80          3rd Qu.:326.0   3rd Qu.:180.0   3rd Qu.:3.920
##  Max.   :33.90          Max.   :472.0   Max.   :335.0   Max.   :4.930
##        wt             qsec       vs      am      gear   carb
##  Min.   :1.513   Min.   :14.50   0:18   0:19   3:15   1: 7
##  1st Qu.:2.581   1st Qu.:16.89   1:14   1:13   4:12   2:10
```

```
##  Median :3.325   Median :17.71                5: 5   3: 3
##  Mean   :3.217   Mean   :17.85                       4:10
##  3rd Qu.:3.610   3rd Qu.:18.90                       6: 1
##  Max.   :5.424   Max.   :22.90                       8: 1
```

```
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
```

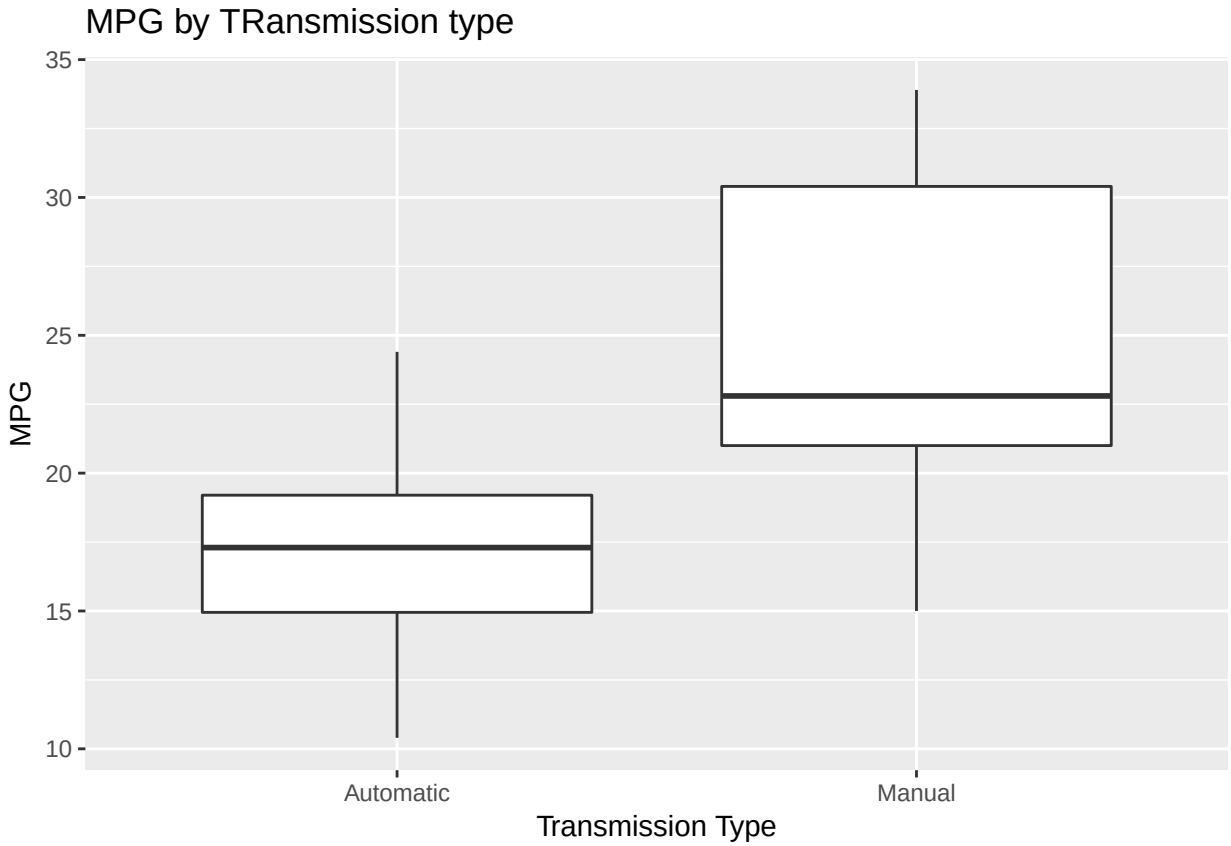
Factors

```
mtcars$am<-as.factor(mtcars$am)
mtcars$cyl  <- factor(mtcars$cyl)
mtcars$vs   <- factor(mtcars$vs)
mtcars$gear <- factor(mtcars$gear)
mtcars$carb <- factor(mtcars$carb)
```

Appendix 2

Exploratory analysis

```
ggplot(data = mtcars, aes(am,mpg)) + geom_boxplot() +
    labs(x= "Transmission Type", y = "MPG", title = "MPG by TRansmission type") +
    scale_x_discrete(breaks=c("0", "1"),
                     labels=c("Automatic", "Manual"))
```



Appendix 2

Model summary

```
summary(fit1)
```

```
##
## Call:
## lm(formula = mpg ~ ., data = mtcars)
##
## 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
```

```
summary(fit2)
```

```
##
## 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
```

```
summary(fit3)
```

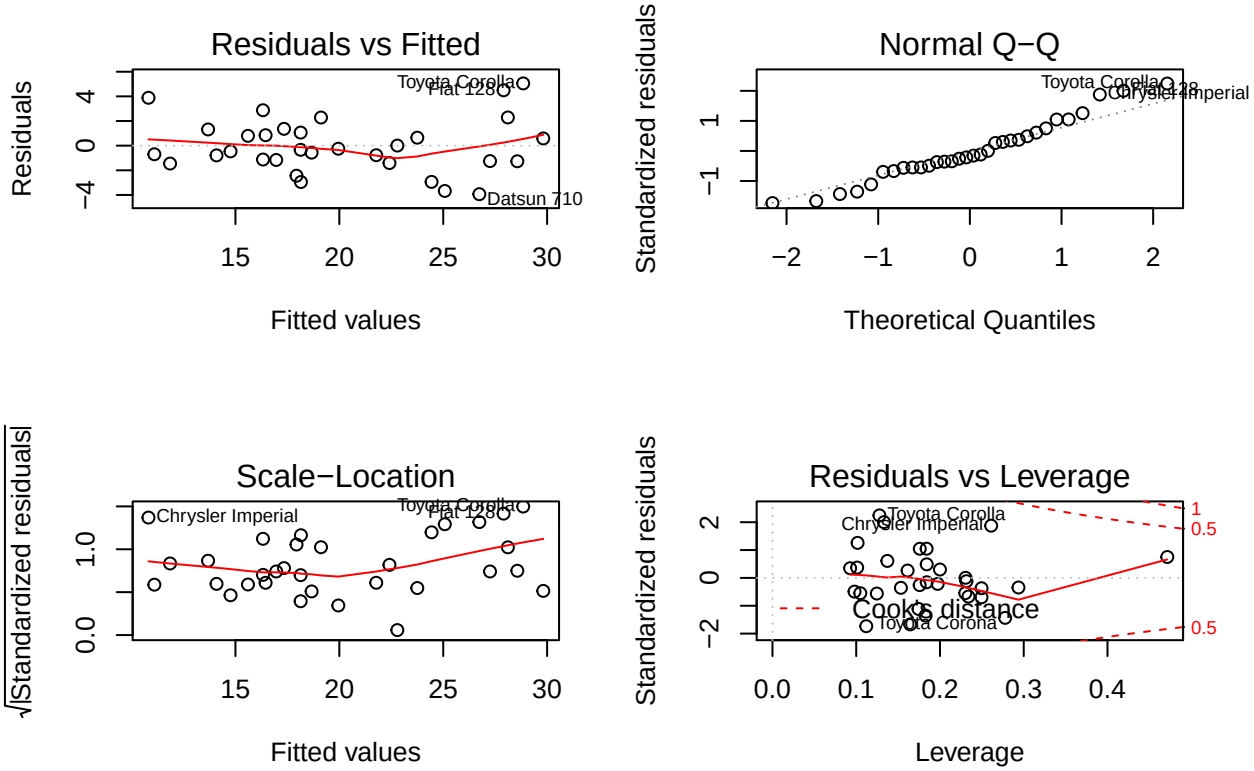
```
##
## Call:
## lm(formula = mpg ~ am, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.3923 -3.0923 -0.2974  3.2439  9.5077
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   17.147      1.125  15.247 1.13e-15 ***
## am1            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
```

Appendix 3

Residual plots of `fit2`

```
par(mfrow = c(2, 2))  # Split the plotting panel into a 2 x 2 grid
plot(fit2)
```



Appendix 4

Statistical inference

```
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
```