Assignment_1

Melvis Onoriode

03/02/2022

```r
library(psych)
library(plotly)
```

```
## Loading required package: ggplot2

##
## Attaching package: 'ggplot2'

## The following objects are masked from 'package:psych':
##
##     %+%, alpha

##
## Attaching package: 'plotly'

## The following object is masked from 'package:ggplot2':
##
##     last_plot

## The following object is masked from 'package:stats':
##
##     filter

## The following object is masked from 'package:graphics':
##
##     layout
```

```r
library(ISLR2)
```

```
## Warning: package 'ISLR2' was built under R version 4.0.5
```

```r
library(leaps)
```

Question 1

Read the data file and report any missing values.

```r
dataFile <-read.csv("HRSexpend.csv", header=TRUE)
dim(dataFile)
```

```
## [1] 9762   17
```

```r
sum(is.na(dataFile))
```

```
## [1] 47792
```

Question 2

```r
dataFile <- na.omit(dataFile)
str(dataFile)
```

```
## 'data.frame':    5488 obs. of  17 variables:
##  $ IDPN    : int  3010 3020 10001010 10004010 10013010 10038010 10050010 10059020 10147010 10206010
##  $ R9OOPMD : num  1660 20060 600 350 3140 ...
##  $ RARACEM : int  1 1 1 1 1 1 3 1 3 2 ...
##  $ R9MSTAT : int  1 1 0 1 0 1 0 1 0 0 ...
##  $ R9AGEY_E: int  72 69 69 68 70 71 66 72 75 71 ...
##  $ H9ITOT  : num  42400 42400 10260 58312 60014376 ...
##  $ R9BMI   : num  24.7 25.7 22.3 29.3 33.5 23.9 25.7 20.3 26.6 31.6 ...
##  $ R9DRINK : int  1 0 1 1 0 1 0 1 0 0 ...
##  $ R9CONDE : int  1 3 0 3 3 3 1 2 3 3 ...
##  $ RAEDYRS : int  12 16 12 16 12 16 17 16 8 12 ...
##  $ R8CESD  : int  0 0 1 1 2 1 1 0 5 1 ...
##  $ R9HOSP  : int  1 0 1 1 1 0 0 0 0 1 ...
##  $ R9NRSHOM: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ R9DOCTOR: int  1 1 1 1 1 1 1 1 1 1 ...
##  $ R9HOMCAR: int  0 0 0 0 0 0 0 0 1 1 ...
##  $ R9HIGOV : int  1 1 1 1 0 1 0 0 1 1 ...
##  $ R1GOVMR : int  0 0 0 0 0 0 0 0 0 0 ...
##  - attr(*, "na.action")= 'omit' Named int [1:4274] 1 2 6 12 13 14 15 16 17 18 ...
##   ..- attr(*, "names")= chr [1:4274] "1" "2" "6" "12" ...
```

```r
dim(dataFile)
```

```
## [1] 5488   17
```

Question 3

```r
comment(dataFile$R9AGEY_E) <- c("Age at interview")
comment(dataFile$R9MSTAT) <- c("Current marital status with partnership")
comment(dataFile$H9ITOT) <- c("Total household income for rospondent and Spouse")
comment(dataFile$R9BMI) <- c("BMI")
comment(dataFile$R9DRINK) <- c("Drinking Behaviour")
comment(dataFile$R9CONDE) <- c("Sum of condition ever had")
comment(dataFile$RAEDYRS) <- c("Years of education")
comment(dataFile$R8CESD) <- c("CESD score")
comment(dataFile$R9HOSP) <-  c("Hospital stay in the previous 2 years")
comment(dataFile$R9NRSHOM) <- c("Nursing home stay in the previous 2 year")
comment(dataFile$R9DOCTOR) <- c("Doctors visit in the previous 2 year")
comment(dataFile$R9HOMCAR) <- c("Home health care")
comment(dataFile$R9HIGOV) <- c ("Covered by federal government health insurance program")
comment(dataFile$R1GOVMR) <- c("Has government plan medicare")
str(dataFile)
```

```
## 'data.frame':    5488 obs. of  17 variables:
##  $ IDPN    : int  3010 3020 10001010 10004010 10013010 10038010 10050010 10059020 10147010 10206010
##  $ R9OOPMD : num  1660 20060 600 350 3140 ...
##  $ RARACEM : int  1 1 1 1 1 1 3 1 3 2 ...
##  $ R9MSTAT : int  1 1 0 1 0 1 0 1 0 0 ...
##   ..- attr(*, "comment")= chr "Current marital status with partnership"
##  $ R9AGEY_E: int  72 69 69 68 70 71 66 72 75 71 ...
```

```
##   ..- attr(*, "comment")= chr "Age at interview"
##  $ H9ITOT  : num  42400 42400 10260 58312 60014376 ...
##   ..- attr(*, "comment")= chr "Total household income for rospondent and Spouse"
##  $ R9BMI   : num  24.7 25.7 22.3 29.3 33.5 23.9 25.7 20.3 26.6 31.6 ...
##   ..- attr(*, "comment")= chr "BMI"
##  $ R9DRINK : int  1 0 1 1 0 1 0 1 0 0 ...
##   ..- attr(*, "comment")= chr "Drinking Behaviour"
##  $ R9CONDE : int  1 3 0 3 3 3 1 2 3 3 ...
##   ..- attr(*, "comment")= chr "Sum of condition ever had"
##  $ RAEDYRS : int  12 16 12 16 12 16 17 16 8 12 ...
##   ..- attr(*, "comment")= chr "Years of education"
##  $ R8CESD  : int  0 0 1 1 2 1 1 0 5 1 ...
##   ..- attr(*, "comment")= chr "CESD score"
##  $ R9HOSP  : int  1 0 1 1 1 0 0 0 0 1 ...
##   ..- attr(*, "comment")= chr "Hospital stay in the previous 2 years"
##  $ R9NRSHOM: int  0 0 0 0 0 0 0 0 0 0 ...
##   ..- attr(*, "comment")= chr "Nursing home stay in the previous 2 year"
##  $ R9DOCTOR: int  1 1 1 1 1 1 1 1 1 1 ...
##   ..- attr(*, "comment")= chr "Doctors visit in the previous 2 year"
##  $ R9HOMCAR: int  0 0 0 0 0 0 0 0 1 1 ...
##   ..- attr(*, "comment")= chr "Home health care"
##  $ R9HIGOV : int  1 1 1 1 0 1 0 0 1 1 ...
##   ..- attr(*, "comment")= chr "Covered by federal government health insurance program"
##  $ R1GOVMR : int  0 0 0 0 0 0 0 0 0 0 ...
##   ..- attr(*, "comment")= chr "Has government plan medicare"
##  - attr(*, "na.action")= 'omit' Named int [1:4274] 1 2 6 12 13 14 15 16 17 18 ...
##   ..- attr(*, "names")= chr [1:4274] "1" "2" "6" "12" ...
```

Question 4

```r
dataFile1 <- dataFile[,-c(1)]
set.seed(1)
trainingData <- sample(c(TRUE, FALSE), nrow(dataFile1), replace = TRUE)
sum(trainingData)
```

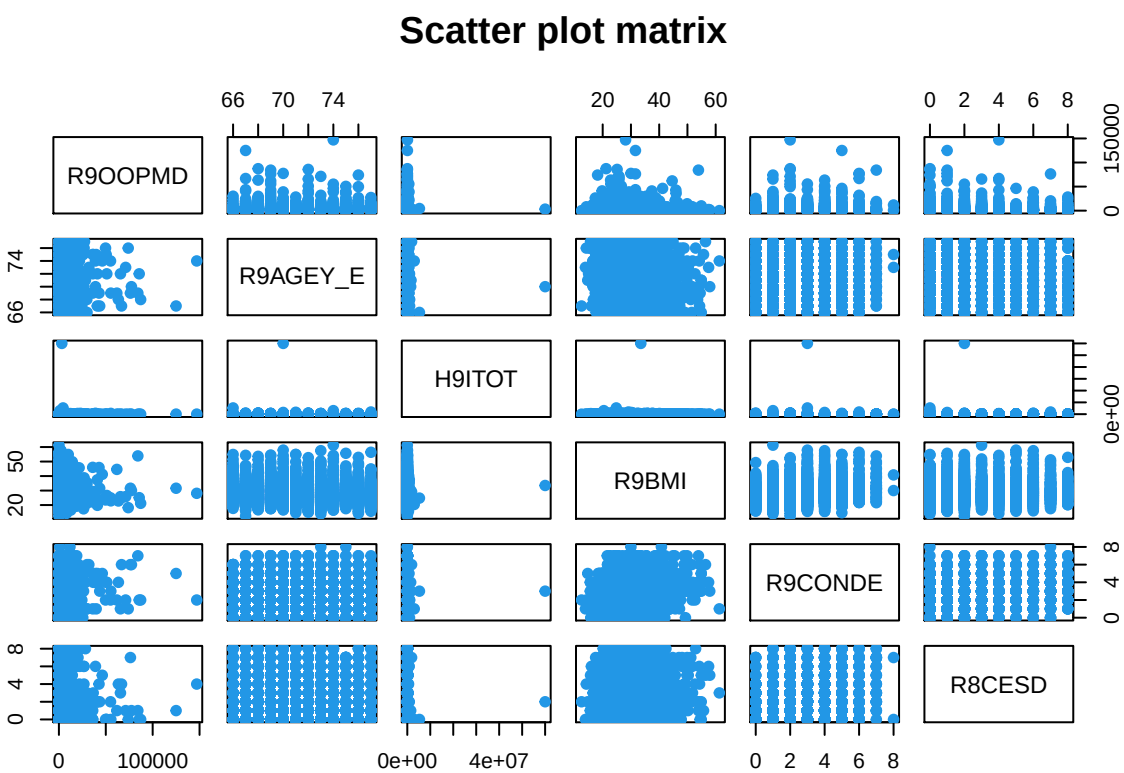
```
## [1] 2704
```

```r
testData <- (!trainingData)
sum(testData)
```

```
## [1] 2784
```

Question 5

```r
pairs(~ R9OOPMD + R9AGEY_E + H9ITOT + R9BMI + R9CONDE+ R8CESD, data = dataFile1, main = "Scatter plot ma
```




```
dataFile1.cor <- cor(dataFile1)
dataFile1.cor
```

```
##               R9OOPMD      RARACEM      R9MSTAT     R9AGEY_E       H9ITOT
## R9OOPMD   1.000000000 -0.044383899  0.003577841  0.010906384  0.007212820
## RARACEM  -0.044383899  1.000000000 -0.170552571 -0.023469509 -0.017310968
## R9MSTAT   0.003577841 -0.170552571  1.000000000 -0.071841129  0.004020172
## R9AGEY_E  0.010906384 -0.023469509 -0.071841129  1.000000000 -0.014882345
## H9ITOT    0.007212820 -0.017310968  0.004020172 -0.014882345  1.000000000
## R9BMI     0.018513164  0.055304558 -0.003701740 -0.069337227  0.006642208
## R9DRINK  -0.014001901 -0.128879899  0.120684152 -0.055832408  0.006566397
## R9CONDE   0.140266186  0.020687703 -0.082997267  0.073084360 -0.003088146
## RAEDYRS   0.084768682 -0.159057914  0.112965881 -0.051178584  0.029139448
## R8CESD    0.036951650  0.085913267 -0.214949695  0.005235028 -0.011347681
## R9HOSP    0.142399569 -0.007610726 -0.033494873  0.068192453  0.018122964
## R9NRSHOM  0.195185943 -0.019512659 -0.076479195  0.087681607 -0.008571595
## R9DOCTOR  0.054009579 -0.037331401  0.031740534  0.006500768  0.006962854
## R9HOMCAR  0.118705816  0.014802728 -0.087162903  0.070981362 -0.009987106
## R9HIGOV  -0.008804156 -0.034592691 -0.006396966  0.067498534 -0.083689186
## R1GOVMR  -0.016147572  0.044200772 -0.066079998  0.018794886 -0.007974568
##                  R9BMI      R9DRINK      R9CONDE     RAEDYRS       R8CESD
## R9OOPMD   0.018513164 -0.014001901  0.140266186  0.08476868  0.036951650
## RARACEM   0.055304558 -0.128879899  0.020687703 -0.15905791  0.085913267
## R9MSTAT  -0.003701740  0.120684152 -0.082997267  0.11296588 -0.214949695
## R9AGEY_E -0.069337227 -0.055832408  0.073084360 -0.05117858  0.005235028
## H9ITOT    0.006642208  0.006566397 -0.003088146  0.02913945 -0.011347681
## R9BMI     1.000000000 -0.101321209  0.245296680 -0.08639855  0.105006763
## R9DRINK  -0.101321209  1.000000000 -0.176955541  0.23122660 -0.163832818
## R9CONDE   0.245296680 -0.176955541  1.000000000 -0.12306351  0.289354319
## RAEDYRS  -0.086398548  0.231226601 -0.123063514  1.00000000 -0.252470289
```

```
## R8CESD    0.105006763 -0.163832818  0.289354319 -0.25247029  1.000000000
## R9HOSP    0.063885574 -0.067142864  0.311090224 -0.02580025  0.151738686
## R9NRSHOM  0.022350515 -0.066171448  0.159587579 -0.01707023  0.094031631
## R9DOCTOR  0.061958174 -0.012217554  0.151075775  0.06936442  0.023211688
## R9HOMCAR  0.062045984 -0.075411840  0.214024001 -0.06234080  0.136388359
## R9HIGOV   0.044826321 -0.022624802  0.044402225 -0.00138774  0.010634071
## R1GOVMR   0.012140245 -0.047201608  0.098878376 -0.09422330  0.097980483
##                R9HOSP     R9NRSHOM     R9DOCTOR     R9HOMCAR      R9HIGOV
## R9OOPMD   0.142399569  0.195185943  0.054009579  0.118705816 -0.008804156
## RARACEM  -0.007610726 -0.019512659 -0.037331401  0.014802728 -0.034592691
## R9MSTAT  -0.033494873 -0.076479195  0.031740534 -0.087162903 -0.006396966
## R9AGEY_E  0.068192453  0.087681607  0.006500768  0.070981362  0.067498534
## H9ITOT    0.018122964 -0.008571595  0.006962854 -0.009987106 -0.083689186
## R9BMI     0.063885574  0.022350515  0.061958174  0.062045984  0.044826321
## R9DRINK  -0.067142864 -0.066171448 -0.012217554 -0.075411840 -0.022624802
## R9CONDE   0.311090224  0.159587579  0.151075775  0.214024001  0.044402225
## RAEDYRS  -0.025800247 -0.017070231  0.069364419 -0.062340800 -0.001387740
## R8CESD    0.151738686  0.094031631  0.023211688  0.136388359  0.010634071
## R9HOSP    1.000000000  0.256022722  0.080216510  0.343766648  0.028696761
## R9NRSHOM  0.256022722  1.000000000  0.030636656  0.313428826  0.016993375
## R9DOCTOR  0.080216510  0.030636656  1.000000000  0.052505294  0.032543391
## R9HOMCAR  0.343766648  0.313428826  0.052505294  1.000000000  0.020185277
## R9HIGOV   0.028696761  0.016993375  0.032543391  0.020185277  1.000000000
## R1GOVMR   0.066956222  0.028518721 -0.015982652  0.059425661  0.008668345
##               R1GOVMR
## R9OOPMD  -0.016147572
## RARACEM   0.044200772
## R9MSTAT  -0.066079998
## R9AGEY_E  0.018794886
## H9ITOT   -0.007974568
## R9BMI     0.012140245
## R9DRINK  -0.047201608
## R9CONDE   0.098878376
## RAEDYRS  -0.094223297
## R8CESD    0.097980483
## R9HOSP    0.066956222
## R9NRSHOM  0.028518721
## R9DOCTOR -0.015982652
## R9HOMCAR  0.059425661
## R9HIGOV   0.008668345
## R1GOVMR   1.000000000
```

Question 6

```r
regressionFit <- regsubsets(R9OOPMD~ ., data = dataFile1[trainingData,], nvmax =17)
regressionFitSummary <- summary(regressionFit)
regressionFitSummary
```

```
## Subset selection object
## Call: regsubsets.formula(R9OOPMD ~ ., data = dataFile1[trainingData,
##     ], nvmax = 17)
## 15 Variables  (and intercept)
##          Forced in Forced out
## RARACEM      FALSE      FALSE
```

```
## R9MSTAT      FALSE      FALSE
## R9AGEY_E     FALSE      FALSE
## H9ITOT       FALSE      FALSE
## R9BMI        FALSE      FALSE
## R9DRINK      FALSE      FALSE
## R9CONDE      FALSE      FALSE
## RAEDYRS      FALSE      FALSE
## R8CESD       FALSE      FALSE
## R9HOSP       FALSE      FALSE
## R9NRSHOM     FALSE      FALSE
## R9DOCTOR     FALSE      FALSE
## R9HOMCAR     FALSE      FALSE
## R9HIGOV      FALSE      FALSE
## R1GOVMR      FALSE      FALSE
## 1 subsets of each size up to 15
## Selection Algorithm: exhaustive
##           RARACEM R9MSTAT R9AGEY_E H9ITOT R9BMI R9DRINK R9CONDE RAEDYRS R8CESD
## 1  ( 1 )  " "     " "     " "      " "    " "   " "     " "     " "     " "
## 2  ( 1 )  " "     " "     " "      " "    " "   " "     "*"     " "     " "
## 3  ( 1 )  " "     " "     " "      " "    " "   " "     "*"     "*"     " "
## 4  ( 1 )  " "     " "     " "      "*"    " "   " "     "*"     "*"     " "
## 5  ( 1 )  " "     " "     " "      "*"    " "   " "     "*"     "*"     " "
## 6  ( 1 )  " "     " "     " "      "*"    " "   " "     "*"     "*"     " "
## 7  ( 1 )  " "     " "     " "      "*"    " "   " "     "*"     "*"     " "
## 8  ( 1 )  " "     " "     "*"      "*"    " "   " "     "*"     "*"     " "
## 9  ( 1 )  "*"     " "     "*"      "*"    " "   " "     "*"     "*"     " "
## 10  ( 1 ) "*"     " "     "*"      "*"    " "   " "     "*"     "*"     " "
## 11  ( 1 ) "*"     " "     "*"      "*"    " "   " "     "*"     "*"     " "
## 12  ( 1 ) "*"     " "     "*"      "*"    " "   " "     "*"     "*"     "*"
## 13  ( 1 ) "*"     " "     "*"      "*"    " "   "*"     "*"     "*"     "*"
## 14  ( 1 ) "*"     " "     "*"      "*"    "*"   "*"     "*"     "*"     "*"
## 15  ( 1 ) "*"     "*"     "*"      "*"    "*"   "*"     "*"     "*"     "*"
##           R9HOSP R9NRSHOM R9DOCTOR R9HOMCAR R9HIGOV R1GOVMR
## 1  ( 1 )  " "    "*"      " "      " "      " "     " "
## 2  ( 1 )  " "    "*"      " "      " "      " "     " "
## 3  ( 1 )  " "    "*"      " "      " "      " "     " "
## 4  ( 1 )  " "    "*"      " "      " "      " "     " "
## 5  ( 1 )  " "    "*"      " "      "*"      " "     " "
## 6  ( 1 )  " "    "*"      " "      "*"      "*"     " "
## 7  ( 1 )  "*"    "*"      " "      "*"      "*"     " "
## 8  ( 1 )  "*"    "*"      " "      "*"      "*"     " "
## 9  ( 1 )  "*"    "*"      " "      "*"      "*"     " "
## 10  ( 1 ) "*"    "*"      " "      "*"      "*"     "*"
## 11  ( 1 ) "*"    "*"      "*"      "*"      "*"     "*"
## 12  ( 1 ) "*"    "*"      "*"      "*"      "*"     "*"
## 13  ( 1 ) "*"    "*"      "*"      "*"      "*"     "*"
## 14  ( 1 ) "*"    "*"      "*"      "*"      "*"     "*"
## 15  ( 1 ) "*"    "*"      "*"      "*"      "*"     "*"
```

Question 7

The Best variable are H9ITOT,R9CONDE, RAEDYRS,R9NRSHOM,R9HOMCAR

Question 8

```
test_matrix <- model.matrix(R9OOPMD ~., data = dataFile1[testData,])
```

```
validation_error <- rep(NA, 17)
validation_error
```

```
##  [1] NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA
```

```
for(i in 1:15){
  coe <- coef(regressionFit, id =i)
  coe_1 <- test_matrix[,names(coe)] %% coe
  validation_error[i] <- mean((dataFile1$R9OOPMD[testData]- coe_1)^2)
}
validation_error
```

```
##  [1] 42558361 42554542 43867038 43249982 42951664 43099362 43082601 42890239
##  [9] 43023805 43202077 43213113 43155460 43077076 43097980 43066653       NA
## [17]       NA
```

Question 9

```
which.min(validation_error)
```

```
## [1] 2
```

```
coef(regressionFit,3)
```

```
## (Intercept)     R9CONDE     RAEDYRS    R9NRSHOM
##  -1135.8473    542.0647    193.1061   5126.3339
```