table

Alexander ZHU

```
cancer=read.csv("cancer_registry.csv")
View(cancer)
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.0 --

## v ggplot2 3.3.2     v purrr   0.3.4
## v tibble  3.0.4     v dplyr   1.0.2
## v tidyr   1.1.2     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.0

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

table

```
library(tidyverse)
library(gridExtra)
```

```
##
## Attaching package: 'gridExtra'

## The following object is masked from 'package:dplyr':
##
##     combine
```

```
library(gplots)
```

```
##
## Attaching package: 'gplots'

## The following object is masked from 'package:stats':
##
##     lowess
```

```r
library(glmnet)
```

```
## Loading required package: Matrix
```

```
##
## Attaching package: 'Matrix'
```

```
## The following objects are masked from 'package:tidyr':
##
##     expand, pack, unpack
```

```
## Loaded glmnet 4.0-2
```

```r
library(mice)
```

```
##
## Attaching package: 'mice'
```

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

```
## The following objects are masked from 'package:base':
##
##     cbind, rbind
```

```r
df = read.csv("cancer_registry.csv") %>%
  mutate(PctSomeCol18_24 = 100 - PctNoHS18_24 - PctHS18_24 - PctBachDeg18_24) %>% filter(incidenceRate <
  filter(avgAnnCount < 20000) %>%
  filter(MedianAge < 200) %>%
  filter(AvgHouseholdSize > 1)
vars <- colnames(df)
misc_vars <- c('binnedInc', 'Geography', 'TARGET_deathRate')
vars_1 <- setdiff(vars, misc_vars)
response_vars <- c('TARGET_deathRate')
predict_vars <- paste(vars_1, collapse = ' + ')
df <- df %>% select(- c('binnedInc', 'Geography'))
df <- df %>% mutate('ID' = rownames(df))
df <- data.frame(sapply(df, as.numeric))

column_names <- c()
res <- c()
r_square <- c()
for (i in colnames(df)){
  if(class(df[[i]]) == "numeric") {
     column_names <- c(column_names, i)
     mean <- mean(df[[i]], na.rm = TRUE)
     std <- sd(df[[i]],na.rm = TRUE)
     res <- c(res, mean)
     res <- c(res, std)
```

```
  }
  else {
    print("not numeric")
  }
}

output_table <- matrix(res, ncol=2,byrow=TRUE)
colnames(output_table) <- c("mean", "std")
rownames(output_table) <- column_names
output <- as.table(output_table)
print(output)
```

```
##                                   mean          std
## avgAnnCount                5.898971e+02 1.173098e+03
## avgDeathsPerYear           1.796541e+02 4.085922e+02
## TARGET_deathRate           1.785176e+02 2.749620e+01
## incidenceRate              4.473481e+02 5.146932e+01
## medIncome                  4.710968e+04 1.210860e+04
## popEst2015                 9.848289e+04 2.605595e+05
## povertyPercent             1.686863e+01 6.438273e+00
## studyPerCap                1.578895e+02 5.370895e+02
## MedianAge                  4.082669e+01 5.203901e+00
## MedianAgeMale              3.956826e+01 5.235713e+00
## MedianAgeFemale            4.215003e+01 5.300330e+00
## AvgHouseholdSize           2.529505e+00 2.493296e-01
## PercentMarried             5.181765e+01 6.884004e+00
## PctNoHS18_24               1.820366e+01 8.040158e+00
## PctHS18_24                 3.500264e+01 9.106843e+00
## PctSomeCol18_24            4.062060e+01 1.100089e+01
## PctBachDeg18_24            6.173103e+00 4.552047e+00
## PctHS25_Over               3.478123e+01 7.034002e+00
## PctBachDeg25_Over          1.331206e+01 5.416320e+00
## PctEmployed16_Over         5.417112e+01 8.344874e+00
## PctUnemployed16_Over       7.822866e+00 3.445562e+00
## PctPrivateCoverage         6.438808e+01 1.062576e+01
## PctPrivateCoverageAlone    4.849246e+01 1.006537e+01
## PctEmpPrivCoverage         4.120566e+01 9.445998e+00
## PctPublicCoverage          3.624292e+01 7.869276e+00
## PctPublicCoverageAlone     1.922818e+01 6.135770e+00
## PctWhite                   8.373254e+01 1.634536e+01
## PctBlack                   9.007788e+00 1.451224e+01
## PctAsian                   1.253503e+00 2.624448e+00
## PctOtherRace               1.988069e+00 3.535983e+00
## PctMarriedHouseholds       5.128136e+01 6.563820e+00
## BirthRate                  5.649137e+00 1.992226e+00
## ID                         1.476500e+03 8.523133e+02
```

```
library(data.table)
```

```
##
## Attaching package: 'data.table'

## The following objects are masked from 'package:dplyr':
```

```
##
##     between, first, last

## The following object is masked from 'package:purrr':
##
##     transpose
```

```
n2=sd(cancer$avgDeathsPerYear)
n3=sd(cancer$TARGET_deathRate)
```

```
# Trying out different set of education vars
model2 = lm(TARGET_deathRate ~ PctHS25_Over + PctBachDeg25_Over +
             PctWhite + PctBlack + PctAsian + PctOtherRace +
             # Interactions
             PctWhite*PctHS25_Over + PctBlack*PctHS25_Over + PctAsian*PctHS25_Over + PctOtherRace*PctHS
             PctWhite*PctBachDeg25_Over + PctBlack *PctBachDeg25_Over + PctAsian *PctBachDeg25_Over + P
             # Confounders
             incidenceRate + medIncome + popEst2015 + povertyPercent + MedianAge,
           data = cancer)
summary(model2)
```

```
##
## Call:
## lm(formula = TARGET_deathRate ~ PctHS25_Over + PctBachDeg25_Over +
##     PctWhite + PctBlack + PctAsian + PctOtherRace + PctWhite *
##     PctHS25_Over + PctBlack * PctHS25_Over + PctAsian * PctHS25_Over +
##     PctOtherRace * PctHS25_Over + PctWhite * PctBachDeg25_Over +
##     PctBlack * PctBachDeg25_Over + PctAsian * PctBachDeg25_Over +
##     PctOtherRace * PctBachDeg25_Over + incidenceRate + medIncome +
##     popEst2015 + povertyPercent + MedianAge, data = cancer)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -110.842  -11.397    0.188   11.385  116.656
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    -3.827e+01  4.561e+01  -0.839  0.40151
## PctHS25_Over                    2.870e+00  9.038e-01   3.175  0.00151 **
## PctBachDeg25_Over               2.066e+00  1.650e+00   1.253  0.21047
## PctWhite                        1.164e+00  4.627e-01   2.515  0.01194 *
## PctBlack                        1.061e+00  5.204e-01   2.039  0.04151 *
## PctAsian                       -5.888e-01  1.984e+00  -0.297  0.76662
## PctOtherRace                   -1.507e+00  1.021e+00  -1.477  0.13978
## incidenceRate                   1.985e-01  6.897e-03  28.784  < 2e-16 ***
## medIncome                       1.238e-04  6.758e-05   1.832  0.06709 .
## popEst2015                     -1.708e-06  1.353e-06  -1.263  0.20684
## povertyPercent                  1.198e+00  1.261e-01   9.504  < 2e-16 ***
## MedianAge                      -2.121e-03  7.921e-03  -0.268  0.78891
## PctHS25_Over:PctWhite          -2.554e-02  9.325e-03  -2.739  0.00620 **
## PctHS25_Over:PctBlack          -3.198e-02  1.065e-02  -3.002  0.00270 **
## PctHS25_Over:PctAsian           1.715e-02  4.162e-02   0.412  0.68025
## PctHS25_Over:PctOtherRace       1.387e-02  2.396e-02   0.579  0.56279
```

```
## PctBachDeg25_Over:PctWhite     -4.068e-02  1.700e-02  -2.393  0.01675 *
## PctBachDeg25_Over:PctBlack     -2.232e-03  1.813e-02  -0.123  0.90204
## PctBachDeg25_Over:PctAsian     -6.033e-03  4.941e-02  -0.122  0.90282
## PctBachDeg25_Over:PctOtherRace  2.918e-02  3.343e-02   0.873  0.38282
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 19.74 on 3027 degrees of freedom
## Multiple R-squared:  0.4973, Adjusted R-squared:  0.4942
## F-statistic: 157.6 on 19 and 3027 DF,  p-value: < 2.2e-16
```

```
# Trying out full interaction
model = lm(TARGET_deathRate ~ PctNoHS18_24 + PctNoHS18_24 + PctSomeCol18_24 +
                              PctBachDeg18_24 + PctWhite + PctBlack + PctAsian + PctOtherRace +
                              PctNoHS18_24*PctWhite + PctNoHS18_24*PctBlack + PctNoHS18_24*PctAsian +
                             PctNoHS18_24*PctOtherRace + PctHS18_24*PctWhite + PctHS18_24*PctBlack + P
                             PctHS18_24*PctOtherRace + PctSomeCol18_24*PctWhite + PctSomeCol18_24*PctB
                              PctSomeCol18_24*PctOtherRace + PctBachDeg18_24*PctWhite + PctBachDeg18_2
                              PctBachDeg18_24*PctAsian + PctBachDeg18_24*PctOtherRace +
                             incidenceRate + medIncome + popEst2015 + povertyPercent + MedianAge,
                            data = cancer)

# Doesn't seem very good, will try indicators for high education and low education
summary(model)
```

```
##
## Call:
## lm(formula = TARGET_deathRate ~ PctNoHS18_24 + PctNoHS18_24 +
##     PctSomeCol18_24 + PctBachDeg18_24 + PctWhite + PctBlack +
##     PctAsian + PctOtherRace + PctNoHS18_24 * PctWhite + PctNoHS18_24 *
##     PctBlack + PctNoHS18_24 * PctAsian + PctNoHS18_24 * PctOtherRace +
##     PctHS18_24 * PctWhite + PctHS18_24 * PctBlack + PctHS18_24 *
##     PctAsian + PctHS18_24 * PctOtherRace + PctSomeCol18_24 *
##     PctWhite + PctSomeCol18_24 * PctBlack + PctSomeCol18_24 *
##     PctAsian + PctSomeCol18_24 * PctOtherRace + PctBachDeg18_24 *
##     PctWhite + PctBachDeg18_24 * PctBlack + PctBachDeg18_24 *
##     PctAsian + PctBachDeg18_24 * PctOtherRace + incidenceRate +
##     medIncome + popEst2015 + povertyPercent + MedianAge, data = cancer)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -113.989  -10.849    0.195   11.260  103.229
##
## Coefficients:
##                                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    2.039e+04  2.135e+04   0.955 0.339900
## PctNoHS18_24                  -2.037e+02  2.131e+02  -0.956 0.339427
## PctSomeCol18_24               -2.015e+02  2.137e+02  -0.943 0.345972
## PctBachDeg18_24               -2.127e+02  2.129e+02  -0.999 0.318205
## PctWhite                      -1.961e+02  2.205e+02  -0.889 0.374184
## PctBlack                      -1.736e+02  2.356e+02  -0.737 0.461307
## PctAsian                      -2.949e+02  8.299e+02  -0.355 0.722385
## PctOtherRace                  -1.866e+01  4.660e+02  -0.040 0.968076
## PctHS18_24                    -2.030e+02  2.135e+02  -0.951 0.342084
```

```
## incidenceRate                 2.108e-01  1.484e-02  14.206  < 2e-16 ***
## medIncome                    -3.960e-04  1.401e-04  -2.827 0.004834 **
## popEst2015                    3.255e-07  2.218e-06   0.147 0.883357
## povertyPercent                9.205e-01  2.760e-01   3.335 0.000896 ***
## MedianAge                    -1.702e-02  1.680e-02  -1.013 0.311433
## PctNoHS18_24:PctWhite         1.968e+00  2.201e+00   0.894 0.371530
## PctNoHS18_24:PctBlack         1.738e+00  2.352e+00   0.739 0.460077
## PctNoHS18_24:PctAsian         2.850e+00  8.282e+00   0.344 0.730877
## PctNoHS18_24:PctOtherRace     2.092e-01  4.661e+00   0.045 0.964216
## PctWhite:PctHS18_24           1.964e+00  2.205e+00   0.890 0.373499
## PctBlack:PctHS18_24           1.740e+00  2.356e+00   0.738 0.460496
## PctAsian:PctHS18_24           3.062e+00  8.307e+00   0.369 0.712511
## PctOtherRace:PctHS18_24       1.555e-01  4.661e+00   0.033 0.973393
## PctSomeCol18_24:PctWhite      1.945e+00  2.207e+00   0.881 0.378492
## PctSomeCol18_24:PctBlack      1.720e+00  2.357e+00   0.730 0.465827
## PctSomeCol18_24:PctAsian      2.875e+00  8.301e+00   0.346 0.729199
## PctSomeCol18_24:PctOtherRace  1.613e-01  4.661e+00   0.035 0.972406
## PctBachDeg18_24:PctWhite      2.047e+00  2.200e+00   0.930 0.352432
## PctBachDeg18_24:PctBlack      1.877e+00  2.352e+00   0.798 0.425133
## PctBachDeg18_24:PctAsian      3.106e+00  8.291e+00   0.375 0.708072
## PctBachDeg18_24:PctOtherRace  3.124e-01  4.642e+00   0.067 0.946359
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 20.9 on 732 degrees of freedom
##   (2285 observations deleted due to missingness)
## Multiple R-squared:  0.4594, Adjusted R-squared:  0.438
## F-statistic: 21.45 on 29 and 732 DF,  p-value: < 2.2e-16
```