Jasen_Missing_Data

Jasen Zhang

11/23/2020

1 Load The Data

```
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()
```

```
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
```

```
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
```

```
library(mice)
```

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

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

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

```
dir = getwd()
data_dir <- paste(substr(dir,1, nchar(dir)-5), "cancer_registry.csv", sep = '')

df <- read.csv(data_dir) %>%
  mutate(PctSomeCol18_24 = 100 - PctNoHS18_24 - PctHS18_24 - PctBachDeg18_24) %>%
  filter(incidenceRate < 1000) %>%
  filter(avgAnnCount < 20000) %>%
  filter(MedianAge < 200) %>%
  filter(AvgHouseholdSize > 1)
```

2 Modifying the Design Matrix

```
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))
```

3. Checking what columns have NA values

```
colSums(is.na(df))
```

```
##             avgAnnCount        avgDeathsPerYear        TARGET_deathRate
##                       0                       0                       0
##           incidenceRate               medIncome             popEst2015
##                       0                       0                       0
##          povertyPercent            studyPerCap              MedianAge
##                       0                       0                       0
##           MedianAgeMale         MedianAgeFemale        AvgHouseholdSize
##                       0                       0                       0
##          PercentMarried           PctNoHS18_24             PctHS18_24
##                       0                       0                       0
##         PctSomeCol18_24         PctBachDeg18_24            PctHS25_Over
##                       0                       0                       0
##       PctBachDeg25_Over      PctEmployed16_Over    PctUnemployed16_Over
##                       0                     144                       0
##      PctPrivateCoverage PctPrivateCoverageAlone      PctEmpPrivCoverage
##                       0                     590                       0
##       PctPublicCoverage  PctPublicCoverageAlone               PctWhite
##                       0                       0                       0
##                PctBlack                PctAsian           PctOtherRace
##                       0                       0                       0
##    PctMarriedHouseholds               BirthRate                      ID
##                       0                       0                       0
```

4. Imputing the two variables with NA

```
imputed_df <- mice(df)
```

```
##
##  iter imp variable
##   1   1  PctEmployed16_Over  PctPrivateCoverageAlone
##   1   2  PctEmployed16_Over  PctPrivateCoverageAlone
##   1   3  PctEmployed16_Over  PctPrivateCoverageAlone
##   1   4  PctEmployed16_Over  PctPrivateCoverageAlone
##   1   5  PctEmployed16_Over  PctPrivateCoverageAlone
##   2   1  PctEmployed16_Over  PctPrivateCoverageAlone
##   2   2  PctEmployed16_Over  PctPrivateCoverageAlone
##   2   3  PctEmployed16_Over  PctPrivateCoverageAlone
##   2   4  PctEmployed16_Over  PctPrivateCoverageAlone
##   2   5  PctEmployed16_Over  PctPrivateCoverageAlone
##   3   1  PctEmployed16_Over  PctPrivateCoverageAlone
##   3   2  PctEmployed16_Over  PctPrivateCoverageAlone
##   3   3  PctEmployed16_Over  PctPrivateCoverageAlone
##   3   4  PctEmployed16_Over  PctPrivateCoverageAlone
##   3   5  PctEmployed16_Over  PctPrivateCoverageAlone
##   4   1  PctEmployed16_Over  PctPrivateCoverageAlone
##   4   2  PctEmployed16_Over  PctPrivateCoverageAlone
##   4   3  PctEmployed16_Over  PctPrivateCoverageAlone
##   4   4  PctEmployed16_Over  PctPrivateCoverageAlone
```

```
##   4   5  PctEmployed16_Over  PctPrivateCoverageAlone
##   5   1  PctEmployed16_Over  PctPrivateCoverageAlone
##   5   2  PctEmployed16_Over  PctPrivateCoverageAlone
##   5   3  PctEmployed16_Over  PctPrivateCoverageAlone
##   5   4  PctEmployed16_Over  PctPrivateCoverageAlone
##   5   5  PctEmployed16_Over  PctPrivateCoverageAlone

## Warning: Number of logged events: 50
```

```
imputed_df2 <- imputed_df$imp

imputed_PctEmployed16_Over <- imputed_df2$PctEmployed16_Over
imputed_PctPrivateCoverageAlone <- imputed_df2$PctPrivateCoverageAlone

avg_imputed_PctEmployed16_Over <- data.frame(apply(imputed_PctEmployed16_Over, 1, mean))
avg_imputed_PctPrivateCoverageAlone <- data.frame(apply(imputed_PctPrivateCoverageAlone, 1, mean))

colnames(avg_imputed_PctEmployed16_Over) <- c('Imputed_PctEmployed16_Over')
colnames(avg_imputed_PctPrivateCoverageAlone) <- c('Imputed_PctPrivateCoverageAlone')

avg_imputed_PctEmployed16_Over <- avg_imputed_PctEmployed16_Over %>% mutate('ID' = as.numeric(rownames(a
avg_imputed_PctPrivateCoverageAlone <- avg_imputed_PctPrivateCoverageAlone %>% mutate('ID' = as.numeric

df_imp <- df %>% left_join(avg_imputed_PctEmployed16_Over, by = 'ID') %>%
  left_join(avg_imputed_PctPrivateCoverageAlone, by = 'ID')

df_imp[is.na(df_imp)] <- 0
```

5. Log transforming skewed variables

```
df_imp2 <- df_imp
log_vars <- c('avgAnnCount', 'avgDeathsPerYear', 'popEst2015', 'studyPerCap', 'PctBachDeg18_24', 'PctAs
log_names <- c()


for(i in log_vars){
  temp <- paste(i, '_log', sep = '')

  log_names <- append(log_names, temp)
  if(i %in% c('studyPerCap')){
    df_imp2 <- df_imp2 %>% mutate(!!as.name(temp) := log(!!as.name(i) + 1))
  } else if(i %in% c('PctAsian', 'PctBachDeg18_24', 'PctOtherRace')){
    df_imp2 <- df_imp2 %>% mutate(!!as.name(temp) := log(!!as.name(i) + exp(-5)))
  } else{
    df_imp2 <- df_imp2 %>% mutate(!!as.name(temp) := log(!!as.name(i)))
  }
}

df_imp3 <- df_imp2 %>% select(- log_vars)
```

```
## Note: Using an external vector in selections is ambiguous.
```

```
## i Use ‘all_of(log_vars)‘ instead of ‘log_vars‘ to silence this message.
## i See <https://tidyselect.r-lib.org/reference/faq-external-vector.html>.
## This message is displayed once per session.
```

6. Save the dataset

```
data_dir <- paste(dir, '/df_imp.RData', sep = '')

save(df_imp3, file = data_dir)
```