Analyzing UK Smoking Data using PCA

Sejal Kankriya

2023-02-18

PCA on UK Smoking Data

In this project we will analyze UK Smoking Data (`smoking.R`):

Description

Survey data on smoking habits from the UK. The data set can be used for analyzing the demographic characteristics of smokers and types of tobacco consumed.

Format

A data frame with 1691 observations on the following 12 variables.

`gender` - Gender with levels Female and Male.

`age` - Age.

`marital_status` - Marital status with levels Divorced, Married, Separated, Single and Widowed.

`highest_qualification` - Highest education level with levels A Levels, Degree, GCSE/CSE, GCSE/O Level, Higher/Sub Degree, No Qualification, ONC/BTEC and Other/Sub Degree

`nationality` - Nationality with levels British, English, Irish, Scottish, Welsh, Other, Refused and Unknown.

`ethnicity` - Ethnicity with levels Asian, Black, Chinese, Mixed, White and Refused Unknown.

`gross_income` - Gross income with levels Under 2,600, 2,600 to 5,200, 5,200 to 10,400, 10,400 to 15,600, 15,600 to 20,800, 20,800 to 28,600, 28,600 to 36,400, Above 36,400, Refused and Unknown.

`region` - Region with levels London, Midlands & East Anglia, Scotland, South East, South West, The North and Wales

`smoke` - Smoking status with levels No and Yes

`amt_weekends` - Number of cigarettes smoked per day on weekends.

`amt_weekdays` - Number of cigarettes smoked per day on weekdays.

`type` - Type of cigarettes smoked with levels Packets, Hand-Rolled, Both/Mainly Packets and Both/Mainly Hand-Rolled

Source National STEM Centre, Large Datasets from stats4schools, https://www.stem.org.uk/resources/elibrary/resource/28452/large-datasets-stats4schools.

Obtained from https://www.openintro.org/data/index.php?data=smoking

Read and Clean the Data

```r
# Load data
source("smoking.R")
```

Take a look into data

```r
# place holder
smoking
```

```
## # A tibble: 1,691 x 12
##    gender   age marital_status highest_qualification nationality ethnicity
##  * <fct>  <int> <fct>          <fct>                  <fct>       <fct>
##  1 Male      38 Divorced       No Qualification       British     White
##  2 Female    42 Single         No Qualification       British     White
##  3 Male      40 Married        Degree                 English     White
##  4 Female    40 Married        Degree                 English     White
##  5 Female    39 Married        GCSE/O Level           British     White
##  6 Female    37 Married        GCSE/O Level           British     White
##  7 Male      53 Married        Degree                 British     White
##  8 Male      44 Single         Degree                 English     White
##  9 Male      40 Single         GCSE/CSE               English     White
## 10 Female    41 Married        No Qualification       English     White
## # i 1,681 more rows
## # i 6 more variables: gross_income <fct>, region <fct>, smoke <fct>,
## #   amt_weekends <int>, amt_weekdays <int>, type <fct>
```

```r
# smoking_data
```

There are many fields there so for this exercise lets only concentrate on smoke, gender, age, marital_status, highest_qualification and gross_income.

Create new data.frame with only these columns.

```r
# place holder
df <- subset(smoking,
  select = c(
    "smoke", "gender", "age", "marital_status",
    "highest_qualification", "gross_income"
  ))
```

Omit all incomplete records

```r
# place holder
smoking_data <- na.omit(df)
```

For PCA feature should be numeric. Some of fields are binary (`gender` and `smoke`) and can easily be converted to numeric type (with one and zero). Other fields like `marital_status` has more than two categories, convert them to binary (i.e. is_married, is_devorced). Several features in the data set are ordinal (`gross_income` and `highest_qualification`), convert them to some king of sensible level (note that levels in factors are not in order)

```r
# place holder
smoking_data <- smoking_data %>%
  mutate(
    gender = as.numeric(gender == "Female"),
    smoke = as.numeric(smoke == "Yes"),
    is_married = ifelse(marital_status == "Married", 1, 0),
    is_divorced = ifelse(marital_status == "Divorced", 1, 0),
    is_widowed = ifelse(marital_status == "Widowed", 1, 0),
    is_single = ifelse(marital_status == "Single", 1, 0),
    is_seperated = ifelse(marital_status == "Separated", 1, 0),
    gross_income = as.integer(as.factor(gross_income)),
    highest_qualification = as.integer(as.factor(highest_qualification))
) %>%
select(-marital_status)
```

PCA on all columns except smoking status

```r
# place holder
pca_fit <- prcomp(smoking_data[c(-1)], scale=T)
pca_fit
```

```
## Standard deviations (1, .., p=9):
## [1] 1.430989e+00 1.267082e+00 1.082729e+00 1.029107e+00 1.019001e+00
## [6] 9.444278e-01 8.962502e-01 6.179302e-01 1.329980e-15
##
## Rotation (n x k) = (9 x 9):
##                               PC1         PC2          PC3          PC4
## gender                 0.09236698 -0.21783341  0.184615226 -0.104924240
## age                    0.59613316 -0.08163273 -0.042358742  0.035577318
## highest_qualification  0.36362803 -0.14199494 -0.046606985  0.004463727
## gross_income           0.13042944  0.06092506 -0.327897235  0.088042179
## is_married             0.17421849  0.75139877 -0.005418731  0.003667260
## is_divorced            0.01862082 -0.22901754  0.737960555  0.440077293
## is_widowed             0.42130406 -0.43774488 -0.313354298 -0.011997990
## is_single             -0.52696478 -0.31626529 -0.371168115  0.105887907
## is_seperated          -0.03112234 -0.11539857  0.271776228 -0.880293210
##                               PC5          PC6         PC7         PC8
## gender                 0.70655749  0.460830149 -0.42300321 -0.09168119
## age                   -0.10577464  0.003884101  0.14087361 -0.77707954
## highest_qualification -0.30123718 -0.333799605 -0.77798599  0.19491471
## gross_income           0.62459699 -0.674083959  0.13008252  0.05075660
## is_married             0.05304036  0.131535021 -0.12587759  0.02579714
## is_divorced            0.03926428 -0.258109030  0.11237644  0.03200276
## is_widowed            -0.03286976  0.256008864  0.31058497  0.44777713
## is_single             -0.05561691 -0.057448283 -0.21412666 -0.37748933
## is_seperated          -0.01394184 -0.262767105  0.09084257 -0.04981720
##                                PC9
## gender                 6.704875e-16
## age                   -7.226903e-16
## highest_qualification -3.639077e-17
## gross_income          -1.169731e-16
## is_married             6.069428e-01
## is_divorced            3.565611e-01
```

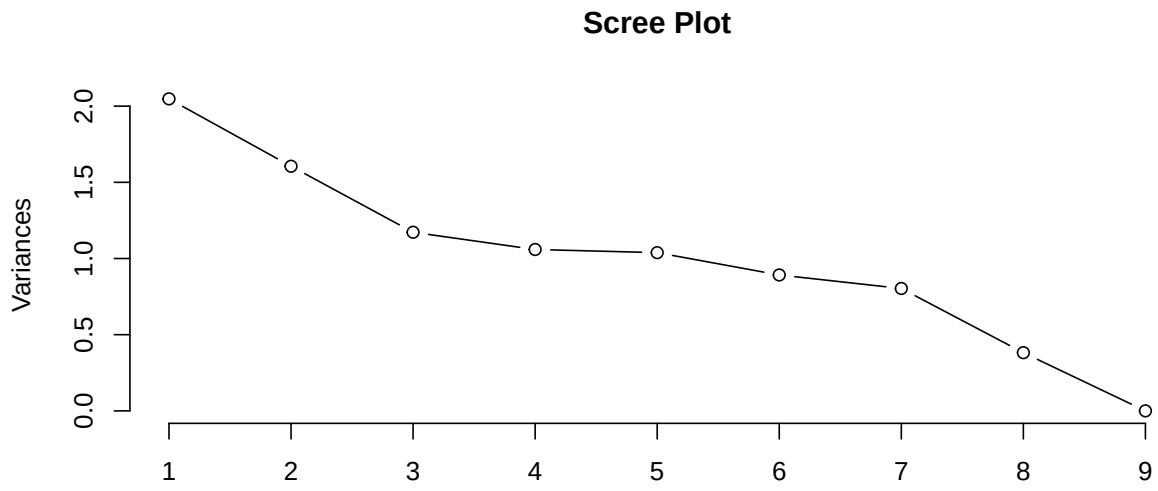
```
## is_widowed              4.110463e-01
## is_single               5.277920e-01
## is_seperated            2.386652e-01
```

```
pca_fit$sdev
```

```
## [1] 1.430989e+00 1.267082e+00 1.082729e+00 1.029107e+00 1.019001e+00
## [6] 9.444278e-01 8.962502e-01 6.179302e-01 1.329980e-15
```

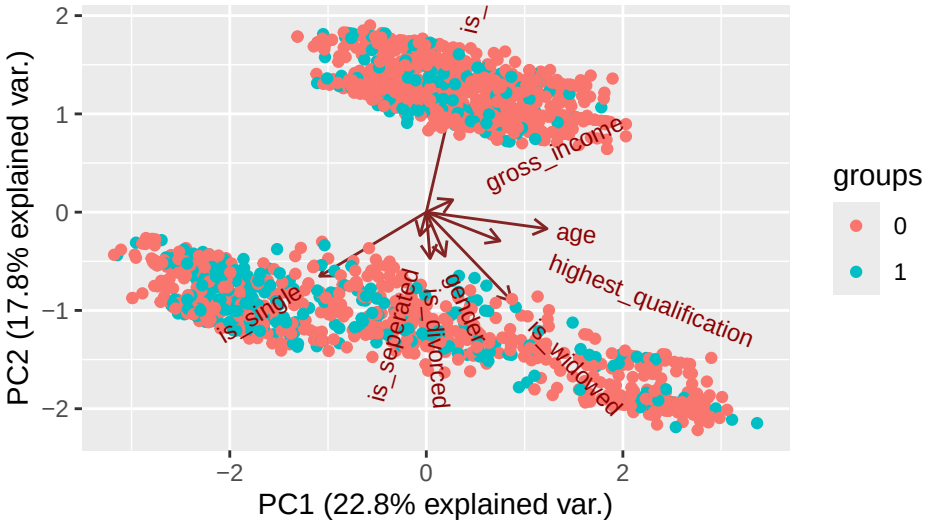
Make a scree plot

```
# place holder
plot(pca_fit, type="line", main="Scree Plot")
```



Biplot color points by smoking field

```
ggbiplot(pca_fit, scale=0, groups = as.factor(smoking_data$smoke))
```



Based on the above biplot, there are two groups - smokers and non-smokers. This can be pointed out by the different colors of the points of two groups. The PC1 seems to be related to unmarried and highest qualification. Whereas, the PC2 appears to be associated with age.

We can use first two to discriminate smoking. However, just by using two PCs wouldn't provide optimal seperation.

Based on the loading vector we can name the PC. Let's say if the PC1 is associated with income and qualification we can name it as social class. and if associated with marital status age, then we can name it as life stage.

For highest_qualification variable, it may be more appropriate to assign numbers for each levels according to their order. for eg. 1 for no qualification, 2 for high school, etc.

Following the suggestion above and redo PCA and biplot

```
source("smoking.R")
df <- subset(smoking,
  select = c(
    "smoke", "gender", "age", "marital_status",
    "highest_qualification", "gross_income"
  ))
smoking_data <- na.omit(df)
smoking_data_redo <- smoking_data %>%
  mutate(
```

```
    gender = as.numeric(gender == "Female"),
    smoke = as.numeric(smoke == "Yes"),
    is_married = ifelse(marital_status == "Married", 1, 0),
    is_divorced = ifelse(marital_status == "Divorced", 1, 0),
    is_widowed = ifelse(marital_status == "Widowed", 1, 0),
    is_single = ifelse(marital_status == "Single", 1, 0),
    is_separated = ifelse(marital_status == "Separated", 1, 0),
    gross_income = as.integer(as.factor(gross_income)),
    highest_qualification = case_when(
      highest_qualification == "No Qualification" ~ 0,
      highest_qualification == "GCSE/O Level" ~ 1,
      highest_qualification == "Other/Sub Degree" ~ 2,
      highest_qualification == "Higher/Sub Degree" ~ 3,
      highest_qualification == "Degree" ~ 4,
      highest_qualification == "A Levels" ~ 5,
      highest_qualification == "GCSE/CSE" ~ 6,
      highest_qualification == "ONC/BTEC" ~ 7)
  ) %>%
select(-marital_status)
```

```
pca_fit_redo <- prcomp(smoking_data_redo[c(-1)], scale=T)
pca_fit_redo
```

```
## Standard deviations (1, .., p=9):
## [1] 1.446012e+00 1.274239e+00 1.082719e+00 1.029584e+00 1.006816e+00
## [6] 9.404839e-01 8.822542e-01 6.135788e-01 1.066825e-15
##
## Rotation (n x k) = (9 x 9):
##                                 PC1         PC2         PC3         PC4
## gender                   0.12460927  0.21664404 -0.16937364  0.10964394
## age                      0.58996698  0.02749485  0.03820234 -0.02112022
## highest_qualification -0.40138796 -0.16879127 -0.03864477  0.05113583
## gross_income             0.14315974 -0.05981797  0.34306446 -0.09154257
## is_married               0.12677364 -0.75683097  0.01053654 -0.01088975
## is_divorced              0.03590566  0.22583089 -0.73554262 -0.44327781
## is_widowed               0.43623480  0.38878852  0.30829811  0.04197099
## is_single               -0.49623041  0.36437943  0.37040328 -0.11612458
## is_separated            -0.02997164  0.11188881 -0.27800402  0.87445809
##                                 PC5         PC6          PC7         PC8
## gender                  -0.679145685  0.64026980 -0.113186602 -0.11328362
## age                      0.177281576 -0.04411082  0.091376951 -0.77945702
## highest_qualification    0.002341914  0.22681821  0.838625886 -0.22703206
## gross_income            -0.679205493 -0.57294539  0.243396999  0.03201901
## is_married              -0.067413999  0.14868335 -0.121853889  0.03203689
## is_divorced             -0.068123270 -0.24705364  0.129488648  0.02477113
## is_widowed               0.189517553  0.19924342  0.361238994  0.43204843
## is_single               -0.011481654 -0.02757212 -0.232632902 -0.36976342
## is_separated            -0.027796346 -0.29119667  0.008729988 -0.04487765
##                                 PC9
## gender                   6.480448e-16
## age                     -2.712888e-16
## highest_qualification -1.525661e-16
## gross_income             2.369658e-16
```

```
## is_married              6.069428e-01
## is_divorced             3.565611e-01
## is_widowed              4.110463e-01
## is_single               5.277920e-01
## is_separated            2.386652e-01
```

```
pca_fit_redo$sdev
```

```
## [1] 1.446012e+00 1.274239e+00 1.082719e+00 1.029584e+00 1.006816e+00
## [6] 9.404839e-01 8.822542e-01 6.135788e-01 1.066825e-15
```

```
ggbiplot(pca_fit_redo, scale=0, groups = as.factor(smoking_data_redo$smoke))
```