Module Four Exercise

msurdek

11/3/2021

## load packages

library(readr)  
library(tidyverse)  
library(FactoMineR)

## read and clean claims file

claims <- read\_csv("/home/msurdek/Documents/School/DAT 610 Optimization & Risk Assessment/Assignments/Module Two Exercise/DAT 610 Auto Accident Personal Injury Claims-1.csv")  
claims <- claims[!is.na(claims$Claim\_Number),]

## formula to confirm ridit-ized values

myData=claims[1:502,6:25]  
  
rid=data.frame("RIDIT" = cbind("\_01"= 2\*(0 + 0.5\*table(myData$IND\_01)[1]/502) - 1,"i2"=2\*(table(myData$IND\_01)[1]/502 +   
0.5\*table(myData$IND\_01)[2]/502) - 1,"i5"=2\*(table(myData$IND\_01)[1]/502 +   
table(myData$IND\_01)[2]/502+table(myData$IND\_01)[3]/502+table(myData$IND\_01)[4]/502+0.5\*table(myData$IND\_01)[5]/502) - 1))  
  
rid

## RIDIT.\_01 RIDIT.i2 RIDIT.i5  
## 1 -0.5039841 0.2788845 0.9203187

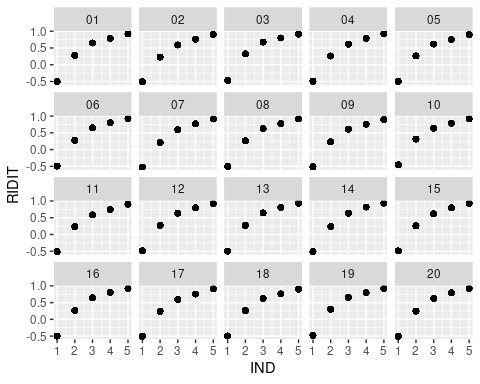
## pivot data so that each claim has 20 rows (1 for each metric) and columns for the metrics IND and RIDIT scores

claims\_direction = claims %>%  
 select(c(1,6:45)) %>%  
 pivot\_longer(2:41, names\_to = c(".value", "variable"), names\_sep = "\_")  
  
claims\_direction

## # A tibble: 10,040 × 4  
## Claim\_Number variable IND RIDIT  
## <dbl> <chr> <dbl> <dbl>  
## 1 5001463 01 1 -0.504  
## 2 5001463 02 1 -0.506  
## 3 5001463 03 1 -0.470  
## 4 5001463 04 4 0.789  
## 5 5001463 05 5 0.896  
## 6 5001463 06 3 0.651  
## 7 5001463 07 3 0.592  
## 8 5001463 08 1 -0.504  
## 9 5001463 09 2 0.231  
## 10 5001463 10 2 0.315  
## # … with 10,030 more rows

## plot IND v RIDIT for each variable

claims\_direction %>%  
 ggplot(mapping = aes(x = IND, y = RIDIT)) +  
 geom\_point() +  
 facet\_wrap(~variable)

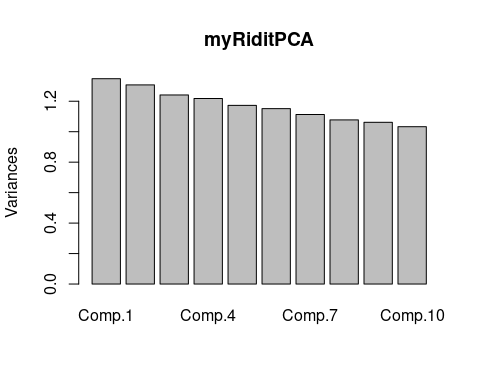
 ## PCA

myRidit = claims[1:502,26:45]  
  
myRiditPCA=princomp(~ ., data = myRidit, cor = TRUE)  
summary(myRiditPCA)

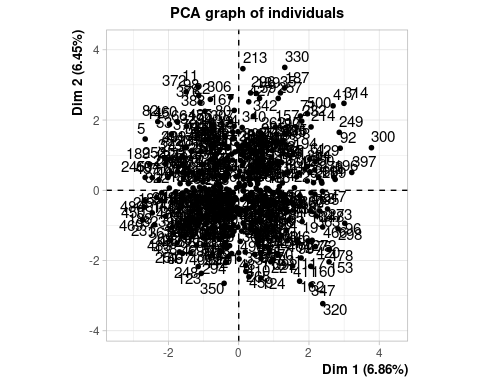
## Importance of components:  
## Comp.1 Comp.2 Comp.3 Comp.4 Comp.5  
## Standard deviation 1.16093641 1.14337267 1.1141535 1.10358813 1.08289782  
## Proportion of Variance 0.06738867 0.06536505 0.0620669 0.06089534 0.05863338  
## Cumulative Proportion 0.06738867 0.13275372 0.1948206 0.25571596 0.31434934  
## Comp.6 Comp.7 Comp.8 Comp.9 Comp.10  
## Standard deviation 1.07303928 1.05489064 1.03805619 1.0303242 1.01625379  
## Proportion of Variance 0.05757067 0.05563971 0.05387803 0.0530784 0.05163859  
## Cumulative Proportion 0.37192001 0.42755972 0.48143775 0.5345161 0.58615474  
## Comp.11 Comp.12 Comp.13 Comp.14 Comp.15  
## Standard deviation 1.00279553 0.96761794 0.9577182 0.94453383 0.9220478  
## Proportion of Variance 0.05027994 0.04681422 0.0458612 0.04460721 0.0425086  
## Cumulative Proportion 0.63643468 0.68324890 0.7291101 0.77371732 0.8162259  
## Comp.16 Comp.17 Comp.18 Comp.19 Comp.20  
## Standard deviation 0.90183907 0.88672695 0.85699268 0.82560990 0.81229003  
## Proportion of Variance 0.04066569 0.03931423 0.03672182 0.03408159 0.03299075  
## Cumulative Proportion 0.85689160 0.89620584 0.93292766 0.96700925 1.00000000

## screeplot

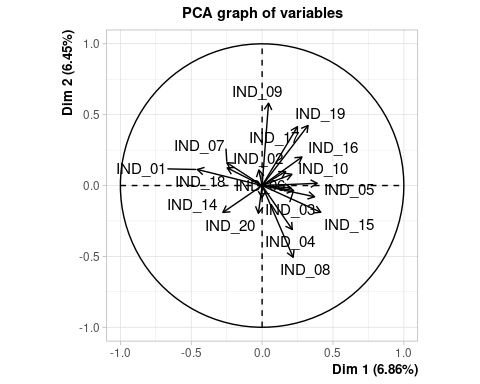
screeplot(myRiditPCA)



res.pca = PCA(claims[,6:25], scale.unit=TRUE, ncp=5, graph=T)



## Warning: ggrepel: 3 unlabeled data points (too many overlaps). Consider  
## increasing max.overlaps



plotellipses(res.pca3,21)

