Project 1: PCA Analysis on a Birth Weight Dataset

Trevor Carpenter, Christina De Cesaris, Michelle Tran
STA 135: Multivariate Data Analysis
Professor Maxime Pouokam
University of California, Davis
May 4, 2021

```
## Loading required package: ggplot2
## Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve3WBa
```

I. Introduction

Although low birth weighted infants can be healthy, under improper care or other circumstances, it can cause serious long-term health problems. A low birth weight is defined as being below 2.5kg. It can be caused by a multitude of factors —poor socioeconomic situation, premature birth, a mother's pre-existing health conditions, and more. In this project, we will highlight how smoking and other factors can cause low birth weights in infants in comparison to non-smoking parents.

We want to find out what variables related to smoking and birth data are related to eachother and how strong those relationship is. In order to answer this, we will be using principal component analysis to analyze the data. PCA is a statistical technique that reduces a dimensionality of a dataset to make it interpretable while also preserving the variability of the data. [more explanation?] We can use it to see what variables group together, identify possible outliers, recognize correlations between variables, and more.

II. Summary of Data

Before conducting PCA onto a dataset, we must first examine the data and decide on what variables we want to specifically look at by looking at different plots and summaries.

The plot above shows....

When conducting the analysis, we also want to know if we want to use the covariance matrix or the correlation matrix. [Discuss the positive and negative aspects of using the covariance matrix for a PCA rather than the correlation matrix.]

III. Analysis

1. [If the overall goal is to compare babies according with their mother's smoking status, how would you proceed ?

My suggestion: Run PCs on these two groups (use the smoking status to split the data and run PCs on them).Then comment on their similarities and differences. You can compare their PCs, scree plots, etc ...]

***Using the dataset for answering our question of interest.

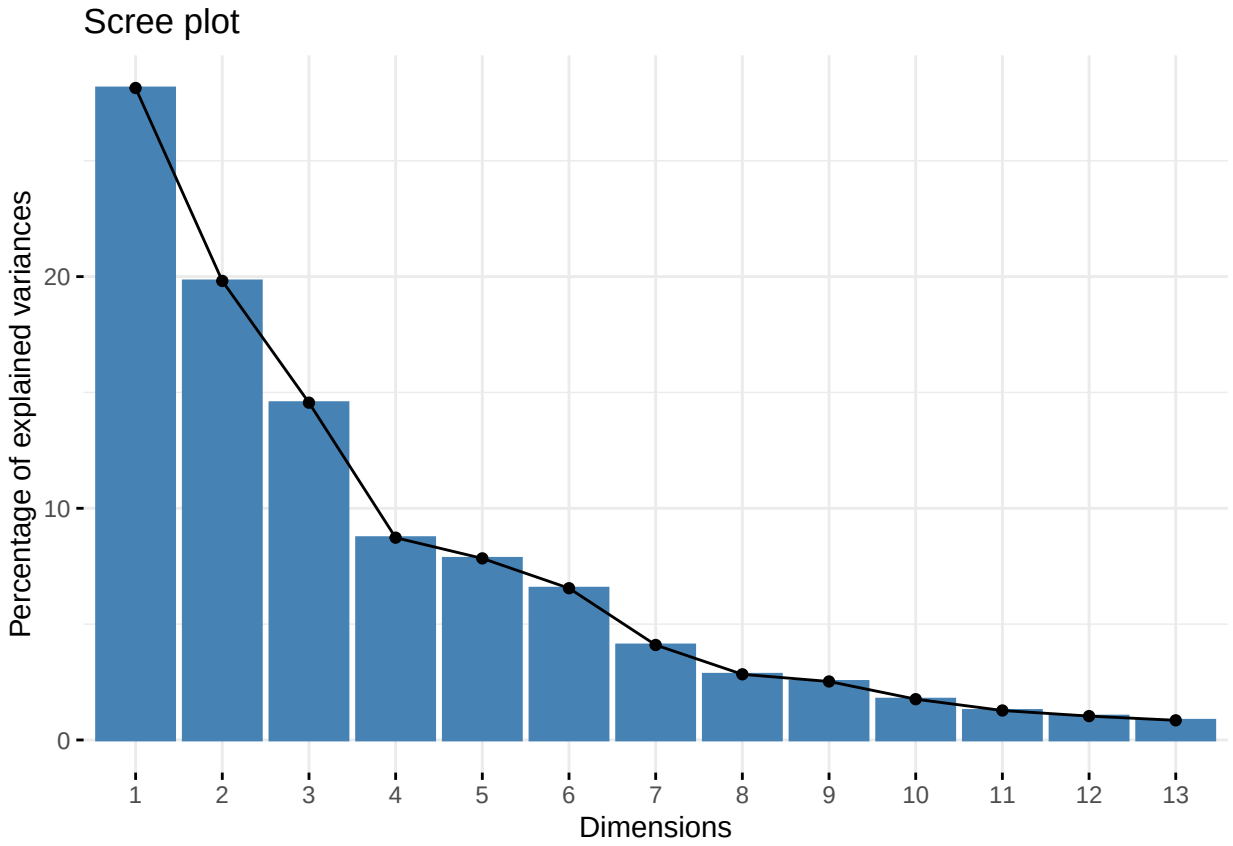
2. [Determine the number of PCs needed when using the correlation matrix.]

```
## Importance of components:
##                           Comp.1   Comp.2    Comp.3    Comp.4    Comp.5
## Standard deviation     1.912632 1.604867 1.3756776 1.0653791 1.0094043
## Proportion of Variance 0.281397 0.198123 0.1455761 0.0873102 0.0783767
## Cumulative Proportion  0.281397 0.479520 0.6250961 0.7124063 0.7907830
##                            Comp.6     Comp.7     Comp.8     Comp.9    Comp.10
```

```
## Standard deviation     0.92276769 0.73001989 0.60705191 0.57293253 0.47850636
## Proportion of Variance 0.06550002 0.04099454 0.02834708 0.02525013 0.01761295
## Cumulative Proportion  0.85628301 0.89727755 0.92562463 0.95087476 0.96848771
##                           Comp.11    Comp.12     Comp.13
## Standard deviation     0.40676905 0.36594910 0.332084370
## Proportion of Variance 0.01272777 0.01030144 0.008483079
## Cumulative Proportion  0.98121548 0.99151692 1.000000000
##
## Loadings:
##             Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7 Comp.8 Comp.9
## Length       0.444                                          0.224  0.543
## Birthweight  0.447                0.212                    -0.335
## Headcirc     0.389                0.203               -0.767
## Gestation    0.382                0.272  0.432         0.387
## smoker             -0.445  0.329                0.280 -0.232        -0.202
## mage               -0.476 -0.289               -0.230         0.362
## mnocig             -0.474  0.224         0.279  0.405
## mheight      0.317         0.304 -0.493 -0.254                0.254  0.264
## mppwt        0.340               -0.442 -0.327         0.280 -0.301 -0.439
## fage               -0.423 -0.321  0.261                             -0.246
## fedyrs             -0.222 -0.293 -0.381  0.512 -0.340        -0.429  0.202
## fnocig             -0.231  0.439  0.375 -0.319 -0.347        -0.485  0.303
## fheight                    0.497         0.306 -0.610         0.288 -0.311
##             Comp.10 Comp.11 Comp.12 Comp.13
## Length       0.396   0.458
## Birthweight -0.236  -0.276   0.660
## Headcirc                    -0.334
## Gestation           -0.289  -0.530
## smoker       0.564  -0.287
## mage        -0.201                  -0.592
## mnocig      -0.544   0.382
## mheight             -0.465           0.319
## mppwt        0.203   0.315
## fage                                 0.660
## fedyrs       0.207
## fnocig                      -0.206
## fheight                      0.217
```

The number of PCs needed using the correlation matrix is only the first 8. Using only these 8 we are able to explain 92.56% of the variance of the data.

This is further visualized using the following scree plot:



In this plot we are able to see that the elbow curve at the 8th dimension reflects the calculations results showing a need of only 8 dimensions to preserve at least 90% of the variance of the data.

Based on the loadings also seen above, we can see that when both birthweight and another score are positive, as seen in Comp. 1 with birthweight and headcirc,

```
## Birthweight    Headcirc
##   0.4465604   0.3891197
```

that means that a large head circumference is beneficial for the baby.

That said, when birthweight is positive and another score in the component is negative, as seen with smoker in Comp. 2:

```
## Birthweight      smoker
##   0.1655971  -0.4446873
```

that is a sign that the attribute of being a smoker is not beneficial for the baby.

As seen based on the PC analysis, these correlations become weaker as we iterate through the components This means that the later components contradicting that birthweight and smoking are actually the same sign score such as in Comp.11:

```
## Birthweight      smoker
##  -0.2764383  -0.4446873
```

are negligible, since the later components explain very little proportion of the overall data variance.

3. Based on the PCs from the summary,

[Interpret the PCs chosen from the previous part. Make sure to specifically comment on whether positive or negative scores (or scores close to 0) for a PC would likely be beneficial for baby.

Hint: To help see which of the loadings are "away from zero", one could set the cutoff argument value in summary()to something other than 0.0. For example one can use a cutoff of 0.2.] Based on the PCs we can see that the most beneficial variables are

4. [Examine plots of the PC scores and interpret them in the context of the problem. For example, what do you think of a particular patient (~~Bebe~~).]

IV. Interpretation

Rehash and summarize analysis section here.

[Suppose a new baby is born after the PCA has been completed. The newly born has the characteristics as indicated in Table 1. Through using the previous PCA results, discuss how this particular individual/baby would compare to the other babies.]

V. Conclusion

[What is your general conclusion and possible future work. By future work, I mean if you have more time what possible questions would you want to investigate ?]

Appendix: R Script

```
knitr::opts_chunk$set(echo = F)
library(factoextra)
smok = read.csv("data/Birthweight_reduced_kg_R.csv")
smok = smok[,2:14]
#Plots




nonsmoke = subset(smok, smoker == 0)
smoke = subset(smok, smoker == 1)
#use: select = c("Length", "Birthweight", "other variables") in subset fcn for specific columns
c = princomp(smok, cor = TRUE)
l = summary(c, loadings = TRUE, cutoff = 0.2)$loadings
summary(c, loadings = TRUE, cutoff = 0.2)
fviz_eig(c, ncp = 13)
c(l[,1]['Birthweight'], l[, 1]['Headcirc'])
c(l[,2]['Birthweight'], l[, 2]['smoker'])
c(l[,11]['Birthweight'], l[, 2]['smoker'])
```