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

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. . . .

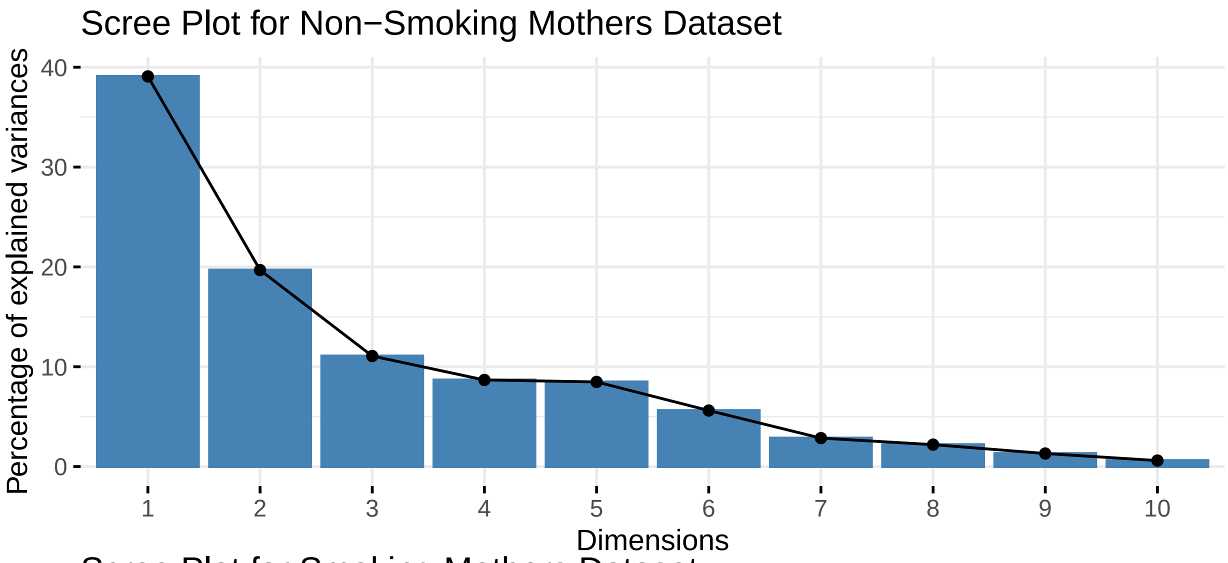
When conducting the analysis, we also want to know if we want to use the covariance matrix or the correlation matrix. [Discuss the positive and negative aspects of using the covariance matrix for a PCA rather than the correlation matrix.]

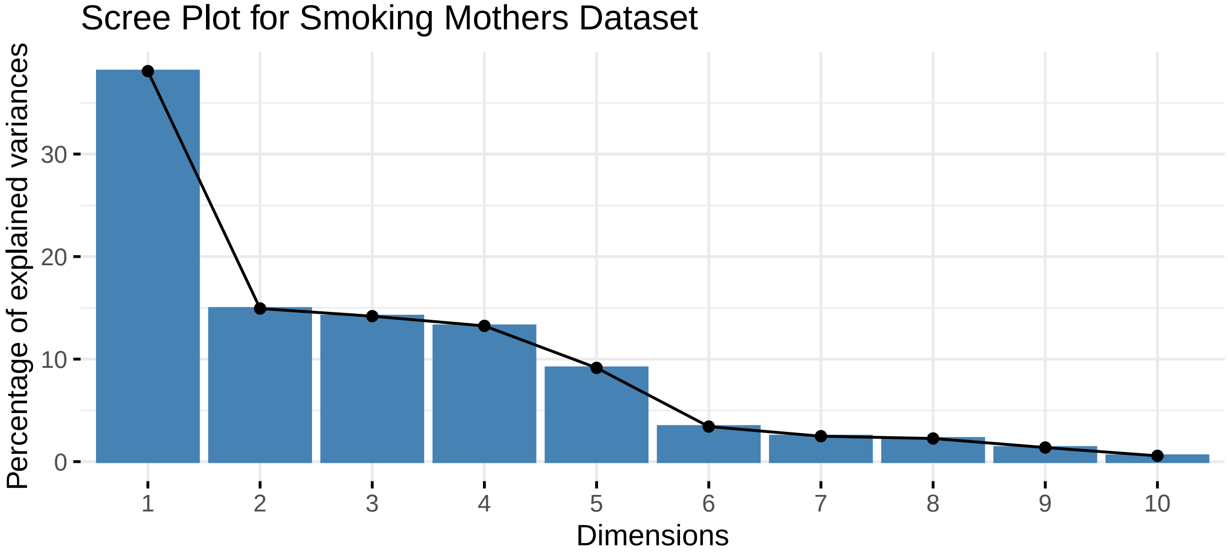
III. Analysis

Does a Mother's Smoking Status have an Effect on the Infant?

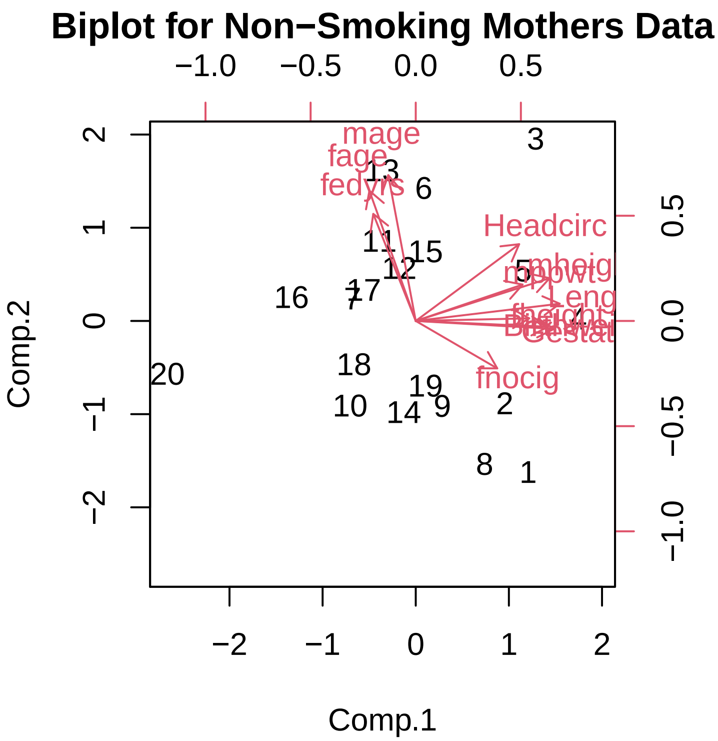
If we compared babies according the their mother's smoking status, we can see what factors influence each other in both groups and compare them to see if they are different or similar. If there is a difference, it shows that a mother's smoking status does play a role in influencing their infant's growth. However, if they are similar, then something other than the mother's smoking status is causing a difference in an infant's growth.

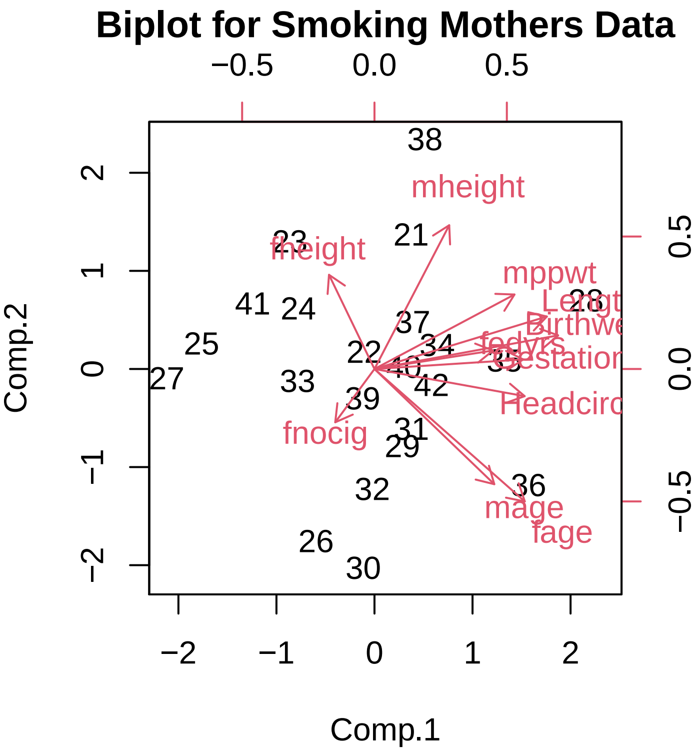
When analyzing the data via PCA, we found that both datasets only needed the first 6 PCs to achieve over 90% explanation of the variance. The majority of the variance is explained by the first PC at about 40% as seen from both scree plots below. From there it tapers off and after PC 6, each PC explains less than 5% of the total variance.





The non-smoking mothers dataset shows that the first PC is based mainly on body measurements of the baby, mother, and father. The gestation period and the infant's weight and length predominantly postively influences this component. This is similar to the smoking mothers dataset, however, there is also an emphasis on the parent's age that was not seen in the non-smoking dataset. For the second PC, in the non-smoking dataset, it was postively influenced by the parent's age. Whereas the other dataset was the opposite in that the parent's age had a negative impact and the parent's height had a postive impact.





For PCs 3-6, the most notable similarity between the two groups is that the parent's age becomes less of an influence on the PCs. In the smoking dataset, the father's characteristics—education and number of cigarettes—negatively impacts the PCs. Whereas in the non-smoking dataset, the factors seem relatively spread out across the PCs and its hard to interpret them.

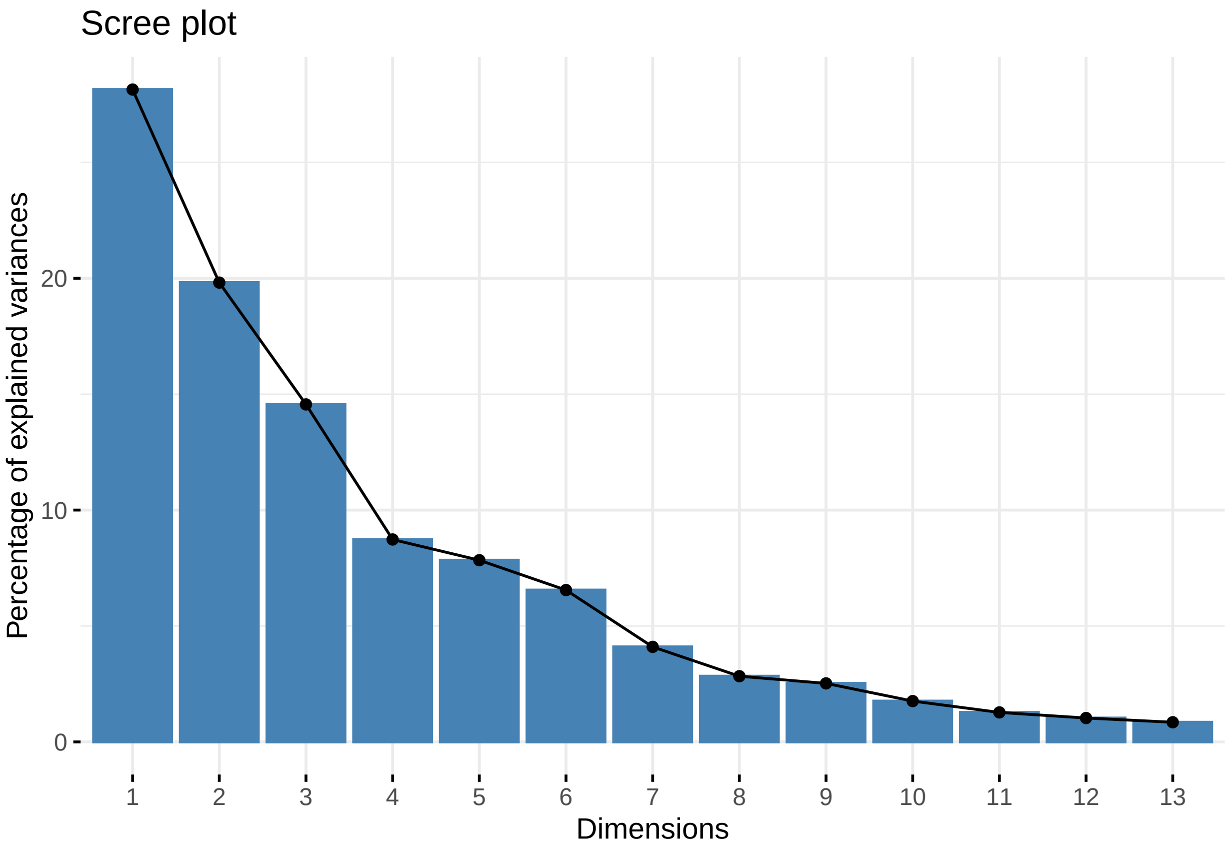
Overall, because there is a difference in which variables impact the PCs, it highlights that a mother's smoking status does impact an infant's growth. The following section will explore what other variables correlate with a mother's smoking status such that it influences an infant's growth.

It is further notable that smoking is not beneficial for babies based on the PC analysis of the dataset as a whole. When the Principle Components are found on the dataset using the correlation matrix, we find the following:

```
## 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.

Based on these Principle Components of the correlation matrix, it is clear that attributes such as head circumference are beneficial for the baby, however smoking is not.

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.]

Based on the

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)
#data
smok = read.csv("data/Birthweight_reduced_kg_R.csv")
smok = smok[,2:14]

#Plots




######[If the overall goal is to compare babies according with their
###### mother’s smoking status, how would you proceed ?]

#split the data
nonsmoke = subset(smok, smoker == 0, select = -c(5))
smoke = subset(smok, smoker == 1, select = -c(5))

#pca and summary
##nonsmoking mothers
pca.nonsmoke =  princomp(formula = ~ Length + Birthweight + Headcirc +
                           Gestation + mage + mheight + mppwt + fage +
                           fedyrs + fnocig + fheight,
                         data = nonsmoke, cor = TRUE, scores = TRUE)
summary(pca.nonsmoke, loadings = TRUE, cutoff = 0.0)



##smoking mothers
pca.smoke = princomp(formula = ~ Length + Birthweight + Headcirc +
                           Gestation + mage + mheight + mppwt + fage +
                           fedyrs + fnocig + fheight,
                         data = smoke, cor = TRUE, scores = TRUE)
summary(pca.smoke, loadings = TRUE, cutoff = 0.0)

#scree plots for nonsmoking vs. smoking mothers datasets
library(factoextra)

fviz_eig(pca.nonsmoke, main = "Scree Plot for Non-Smoking Mothers Dataset")
fviz_eig(pca.smoke, main = "Scree Plot for Smoking Mothers Dataset")

#biplots
biplot(x = pca.nonsmoke, main = "Biplot for Non-Smoking Mothers Data",
       pc.biplot = TRUE)
biplot(x = pca.smoke, main = "Biplot for Smoking Mothers Data",
       pc.biplot = TRUE)
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'])
```

```r
c(l[,2]['Birthweight'], l[, 2]['smoker'])
c(l[,11]['Birthweight'], l[, 2]['smoker'])
```