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.

[state question of interest]. 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. . . .

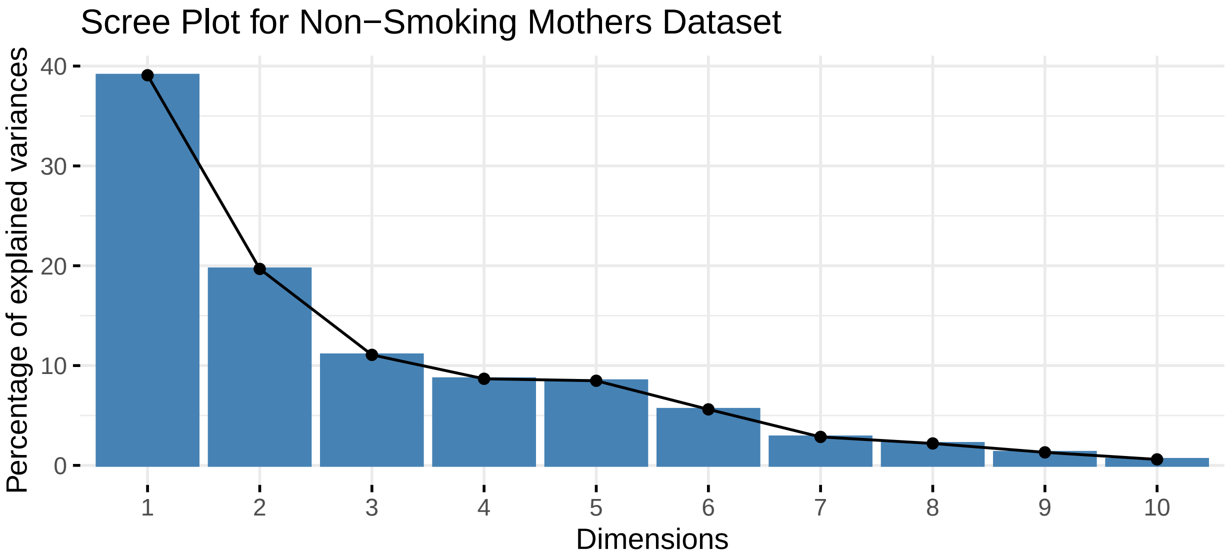
When conducting the analysis, we also want to know if we want to use the covariance matrix or the correlation matrix. [Discuss the positive and negative aspects of using the covariance matrix for a PCA rather than the correlation matrix.]

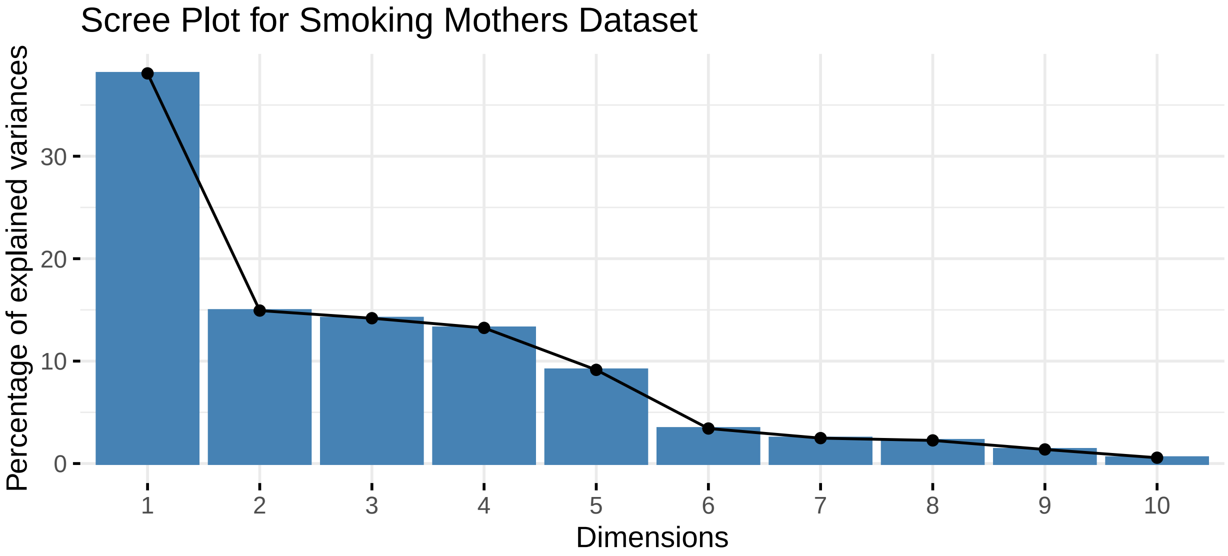
III. Analysis

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

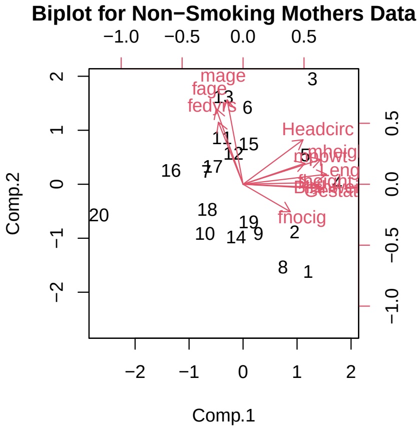
If we compared babies according the their mother's smoking status, we can see what factors influence each other in both groups and compare them to see if they are different or similar. If there is a difference, it shows that a mother's smoking status does play a role in influencing their infant's growth.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

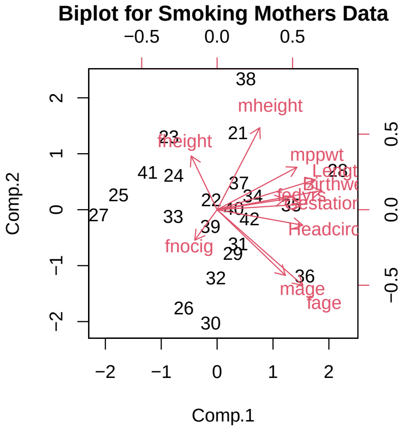
When analyzing the data via PCA, we found that both datasets only needed the first 6 PCs to achieve over 90% explanation of the variance. The majority of the variance is explained by the first PC at about 40% as seen from both scree plots below. From there it tapers off and after PC 6, each PC explains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

[**Question of Interest Analysis Caption Title**]

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

2. [Determine the number of PCs needed when using the correlation matrix.]
3. [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.]

3. [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.]

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 (Baby).]

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)
#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)
```