# 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.

**Scree Plot for Non–Smoking Mothers Dataset**



**Scree Plot for Smoking Mothers Dataset**
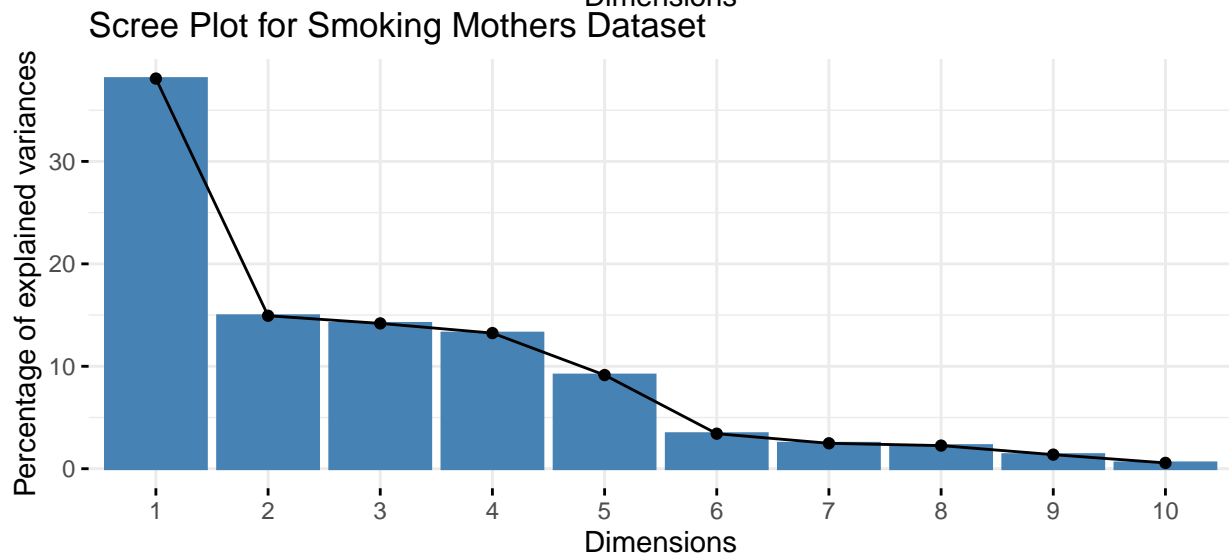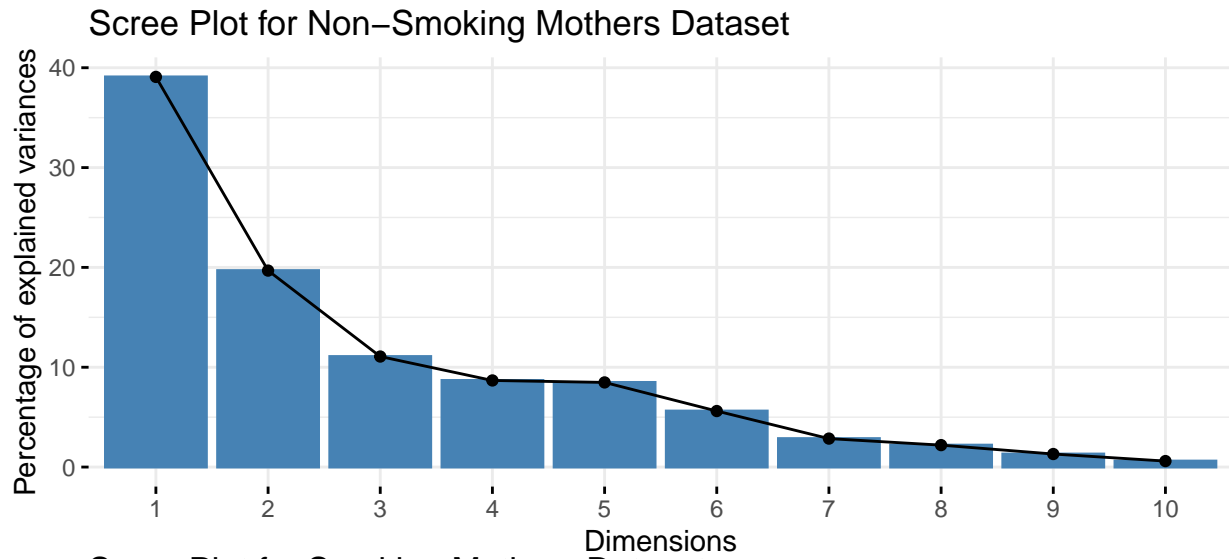
The non-smoking mothers dataset shows that the first PC is based mainly on body measurements of the baby, mother, and father. The gestation period and the infant's weight and length predominantly postively influences this component. This is similar to the smoking mothers dataset, however, there is also 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.

**Biplot for Non−Smoking Mothers Data**



**Biplot for Smoking Mothers Data**



Comp.1
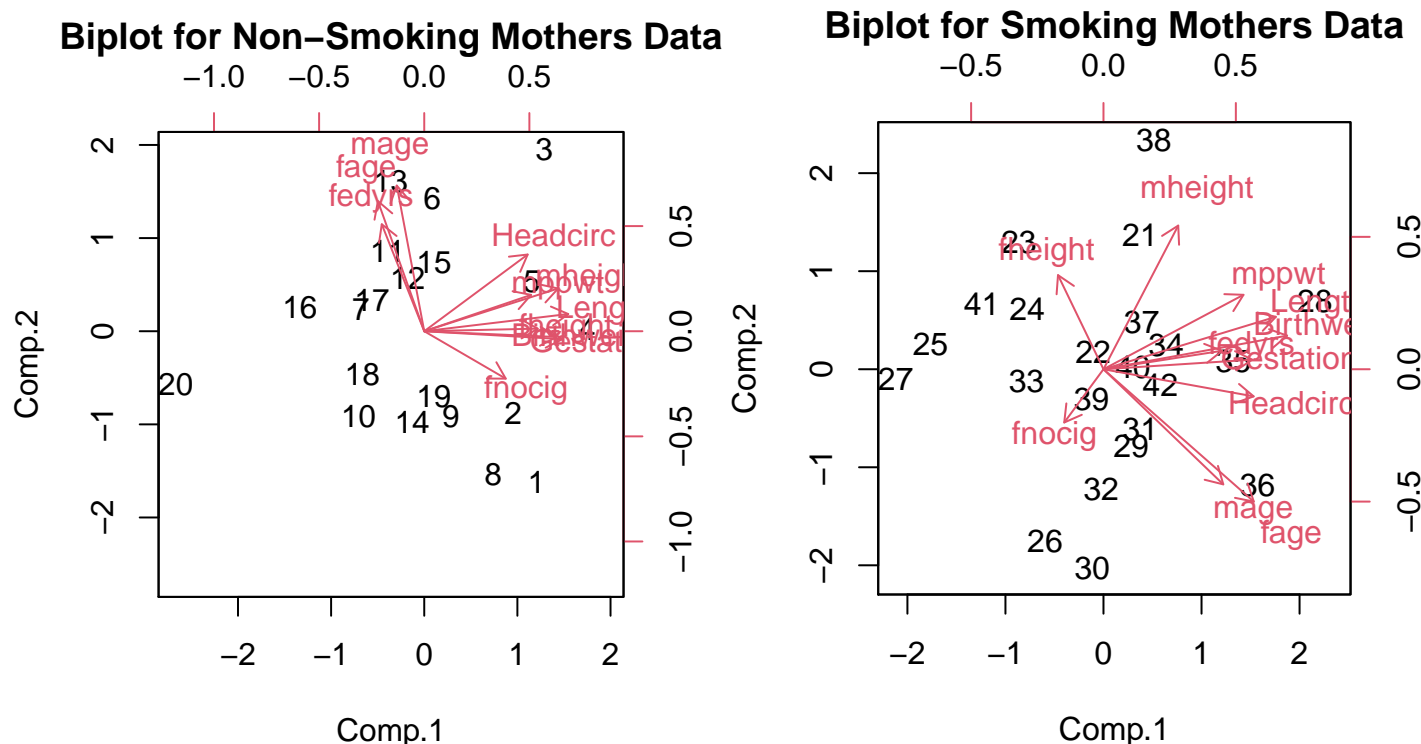
For PCs 3-6, the most notable similarity between the two groups is that the parent's age becomes less of an influence on the PCs. In the smoking dataset, the father's characteristics—education and number of cigarettes—negatively impacts the PCs. Whereas in the non-smoking dataset, the factors seem relatively 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.

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:

|  | Comp.1 | Comp.2 | Comp.3 | Comp.4 | Comp.5 | Comp.6 | Comp.7 |
|---|---|---|---|---|---|---|---|
| Standard deviation | 1.912632 | 1.604867 | 1.3756776 | 1.0653791 | 1.0094043 | 0.9227677 | 0.7300199 |
| Proportion of Variance | 0.281397 | 0.198123 | 0.1455761 | 0.0873102 | 0.0783767 | 0.0655000 | 0.0409945 |
| Cumulative Proportion | 0.281397 | 0.479520 | 0.6250961 | 0.7124063 | 0.7907830 | 0.8562830 | 0.8972775 |

|  | Comp.8 | Comp.9 | Comp.10 | Comp.11 | Comp.12 | Comp.13 |
|---|---|---|---|---|---|---|
| Standard deviation | 0.6070519 | 0.5729325 | 0.4785064 | 0.4067690 | 0.3659491 | 0.3320844 |
| Proportion of Variance | 0.0283471 | 0.0252501 | 0.0176129 | 0.0127278 | 0.0103014 | 0.0084831 |
| Cumulative Proportion | 0.9256246 | 0.9508748 | 0.9684877 | 0.9812155 | 0.9915169 | 1.0000000 |

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:

4

## Scree plot
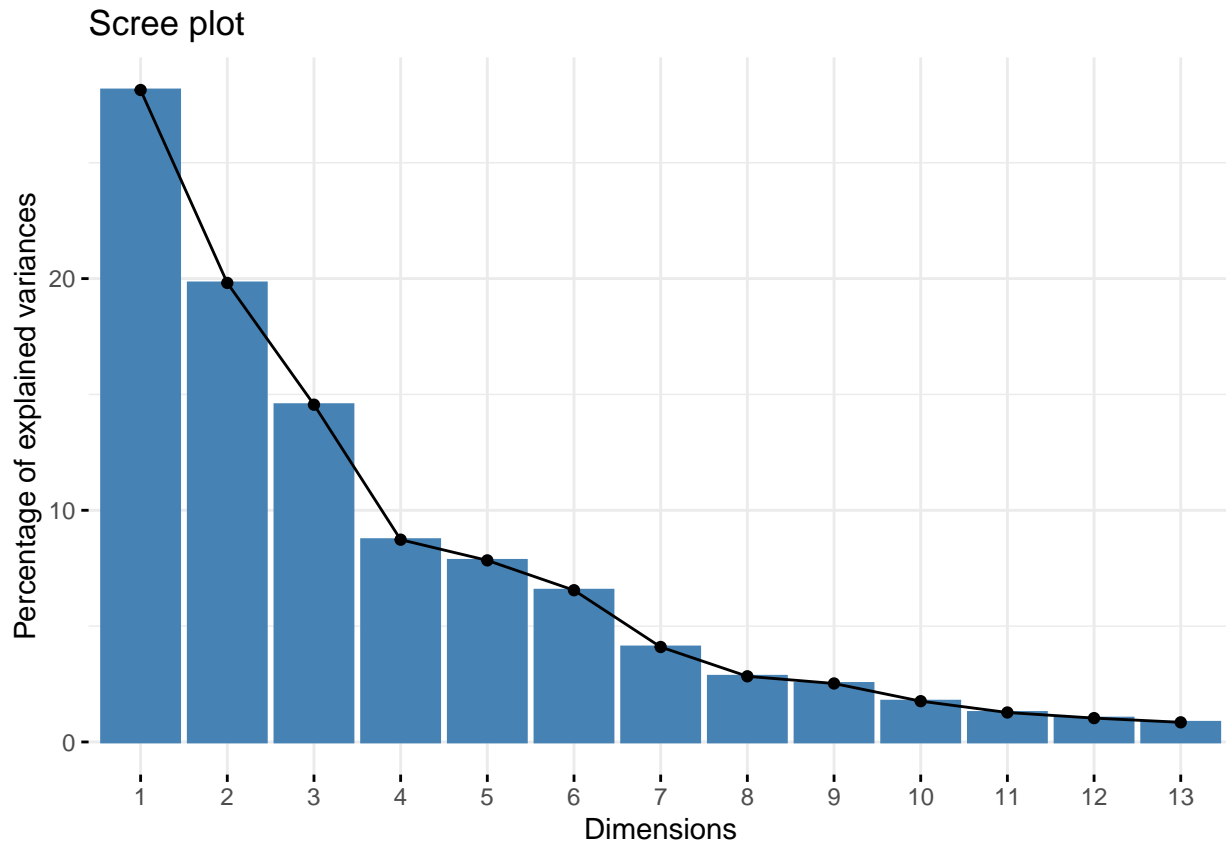


In this plot we are able to see that the elbow curve at the 8th dimension reflects the calculations results showing a need of only 8 dimensions to preserve at least 90% of the variance of the data.

Based on the loadings:

|             | Comp.1 | Comp.2 | Comp.3 | Comp.4 | Comp.5 | Comp.6 | Comp.7 |
|-------------|--------|--------|--------|--------|--------|--------|--------|
| Length      | 0.444  |        |        |        |        |        |        |
| Birthweight | 0.447  |        |        | 0.212  |        |        |        |
| 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 |
| mage        |        | -0.476 | -0.289 |        |        | -0.230 |        |
| mnocig      |        | -0.474 | 0.224  |        | 0.279  | 0.405  |        |
| mheight     | 0.317  |        | 0.304  | -0.493 | -0.254 |        |        |
| mppwt       | 0.340  |        |        | -0.442 | -0.327 |        | 0.280  |
| fage        |        | -0.423 | -0.321 | 0.261  |        |        |        |
| fedyrs      |        | -0.222 | -0.293 | -0.381 | 0.512  | -0.340 |        |
| fnocig      |        | -0.231 | 0.439  | 0.375  | -0.319 | -0.347 |        |
| fheight     |        |        | 0.497  |        | 0.306  | -0.610 |        |

|             | Comp.8 | Comp.9 | Comp.10 | Comp.11 | Comp.12 | Comp.13 |
|-------------|--------|--------|---------|---------|---------|---------|
| Length      | 0.224  | 0.543  | 0.396   | 0.458   |         |         |
| Birthweight | -0.335 |        | -0.236  | -0.276  | 0.660   |         |
| Headcirc    |        |        |         |         | -0.334  |         |
| Gestation   |        |        |         | -0.289  | -0.530  |         |
| smoker      |        | -0.202 | 0.564   | -0.287  |         |         |

|          | Comp.8 | Comp.9 | Comp.10 | Comp.11 | Comp.12 | Comp.13 |
|----------|--------|--------|---------|---------|---------|---------|
| mage     | 0.362  |        | -0.201  |         |         | -0.592  |
| mnocig   |        |        | -0.544  | 0.382   |         |         |
| mheight  | 0.254  | 0.264  |         | -0.465  |         | 0.319   |
| mppwt    | -0.301 | -0.439 | 0.203   | 0.315   |         |         |
| fage     |        | -0.246 |         |         |         | 0.660   |
| fedyrs   | -0.429 | 0.202  | 0.207   |         |         |         |
| fnocig   | -0.485 | 0.303  |         |         | -0.206  |         |
| fheight  | 0.288  | -0.311 |         |         | 0.217   |         |

we can see that when the physical features of the baby are all positive, as seen in the first four features of Comp. 1,

| Length    | Birthweight | Headcirc  | Gestation |
|-----------|-------------|-----------|-----------|
| 0.4442198 | 0.4465604   | 0.3891197 | 0.3822305 |

that means that these features all correlate with eachother, meaning that a long gestation period is well correlated with a larger and healthier 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, since it will impact the physical features.

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  | -0.2871868 |

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

We can also see that the number of cigarettes smoked per day by mother's and father's are both, as expected, well correlated with the smoking status of the mother. This is also visible with the mother and father's ages, as seen by the data in Comp.2:

| smoker     | mage       | mnocig     | fage       | fnocig     |
|------------|------------|------------|------------|------------|
| -0.4446873 | -0.4756988 | -0.4735367 | -0.4231726 | -0.2310304 |

Based on these Principle Components of the correlation matrix, it is clear that a larger gestational period is a beneficial thing for a baby's health, however a greater count of smoked cigarettes is not. It is also clear that the father's smoking status, as shown by the number of cigarettes he smokes per day, is well correlated with the mothers.

What is further interesting about this analysis is the lack of correlation of the physical attributes of the parents with the physical attributes of their newborn in comparison to that of the smoking features. While in Comp.1 we see that the physical features of the child and the physical features of the mother are similar scores:

| Birthweight | Headcirc | Gestation | mheight | mppwt |
|---|---|---|---|---|
| 0.4465604 | 0.3891197 | 0.3822305 | 0.3169582 | 0.3404202 |

we see only three principle components later in Comp.4 a contradiction:

| Birthweight | Headcirc | Gestation | mheight | mppwt |
|---|---|---|---|---|
| 0.2124146 | 0.2030967 | 0.2719839 | -0.4928765 | -0.4417855 |

This contradiction may be in a later component, however it is still a component that explains 8.7% of the overall variance of the data, and is relatively significant. It shows a reduced importance of the mother's physical features in explaining the physical features of her child, and a comparative increase in the impact of smoking.

## IV. Interpretation

Based on the analysis we can understand an estimation of what the physical features of a new baby not considered in this analysis will be using only a few features of the parents. For example, if the father is older he is more likely to smoke multiple cigarettes per day. In this case, based on the Principle Component Analysis we could expect that the mother also smokes multiple per day, and as such is marked as a smoker in the data. If this is in fact true we can predict that the gestation period will be shorter and the head circumference, birth weight, and length of this newborn baby will all be lower than expected.

We would also see that the physical features of the father and mother such as their height and weight do not factor into the features of their child nearly as much as their actions do when it comes to smoking or not. Based on only these Principle Components we are able to understand how the smoking features of the father and mother have a direct impact on the features of their baby, and subsequently the healthiness of their baby.

| length | birthweight | headcirc | gestation | smoker | motherage | mnocig | mheight | mppwt | fage |
|---|---|---|---|---|---|---|---|---|---|
| 61 | 5.1 | 36 | 43 | 0 | 43 | 7 | 165 | 64 | 38 |

| fedyrs | fnocig | fheight |
|---|---|---|
| 19 | 45 | 189 |

## V. Conclusion

Generally, this analysis shows that while physical features of the parents do explain the physical features of their children, features regarding their smoking habits is another very strong influence. We also see that the smoking actions of the father may correlate with the smoking actions of the mother, meaning that it is likely if one smokes so does the other. This is interesting because it disproves any assumption that the actions of the father do not influence the characteristics of his child.

In future analysis of this topic, it would be interesting to include alcohol consumption as another feature, as it has also been shown to have impacts on child gestation and birth features. We could also further analyze any correlation between the mother and father's non-physical features with each other, such as seeing if the more positive actions of the father in activities such as exercise or sleep influence the actions of the mother in similar ways that smoking does. Father's education level is also a feature in this dataset that does not correlate with other features and wasn't heavily analyzed, however maybe the addition of Mother's education level would show some deeper results.

## Appendix: R Script

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

new_print_loadings = function (x, digits = 3L, cutoff = 0.1, sort = FALSE, ...)
{
  ## code from the original print.loadings function, without printing the variances
    Lambda <- unclass(x)
    p <- nrow(Lambda)
    factors <- ncol(Lambda)
    if (sort) {
        mx <- max.col(abs(Lambda))
        ind <- cbind(1L:p, mx)
        mx[abs(Lambda[ind]) < 0.5] <- factors + 1
        Lambda <- Lambda[order(mx, 1L:p), ]
    }
    fx <- setNames(format(round(Lambda, digits)), NULL)
    nc <- nchar(fx[1L], type = "c")
    fx[abs(Lambda) < cutoff] <- strrep(" ", nc)
    return(fx)
}

new_print_sum = function (x, digits = 3L, loadings = x$print.loadings, cutoff = x$cutoff,
    ...)
{
    vars <- x$sdev^2
    vars <- vars/sum(vars)
    return(rbind(`Standard deviation` = x$sdev, `Proportion of Variance` = vars,
        `Cumulative Proportion` = cumsum(vars)))
}

#Plots



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

```r
#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)
## PCA Analysis on whole data
c = princomp(smok, cor = TRUE)
l = summary(c, loadings = TRUE, cutoff = 0.15)$loadings
kable(new_print_sum(c, cutoff = 0.15)[,1:7])
kable(new_print_sum(c, cutoff = 0.15)[,8:13])
fviz_eig(c, ncp = 13)
kable(new_print_loadings(l, cutoff = 0.2)[,1:7])
kable(new_print_loadings(l, cutoff = 0.2)[,8:13])
vars = l[,1][1:4]
kable(matrix(as.numeric(vars), ncol = length(vars)), col.names = names(vars))

vars = c(l[,2]['Birthweight'], l[, 2]['smoker'])
kable(matrix(as.numeric(vars), ncol = length(vars)), col.names = names(vars))

vars = c(l[,11]['Birthweight'], l[, 11]['smoker'])
kable(matrix(as.numeric(vars), ncol = length(vars)), col.names = names(vars))

vars = c(l[,2][5:7], l[, 2][10], l[, 2][12])
kable(matrix(as.numeric(vars), ncol = length(vars)), col.names = names(vars))
vars = c(l[,1][2:4], l[,1][8:9])
kable(matrix(as.numeric(vars), ncol = length(vars)), col.names = names(vars))
```

```r
vars = c(l[,4][2:4], l[,4][8:9])
kable(matrix(as.numeric(vars), ncol = length(vars)), col.names = names(vars))

make_table = function(mat) {
}
kable(matrix(c(61, 5.1, 36, 43, 0, 43, 7, 165, 64, 38), ncol = 10),  col.names = c("length", "birthweigh
kable(matrix(c( 19, 45, 189), ncol = 3), col.names = c("fedyrs", "fnocig", "fheight"))
```