Project 1

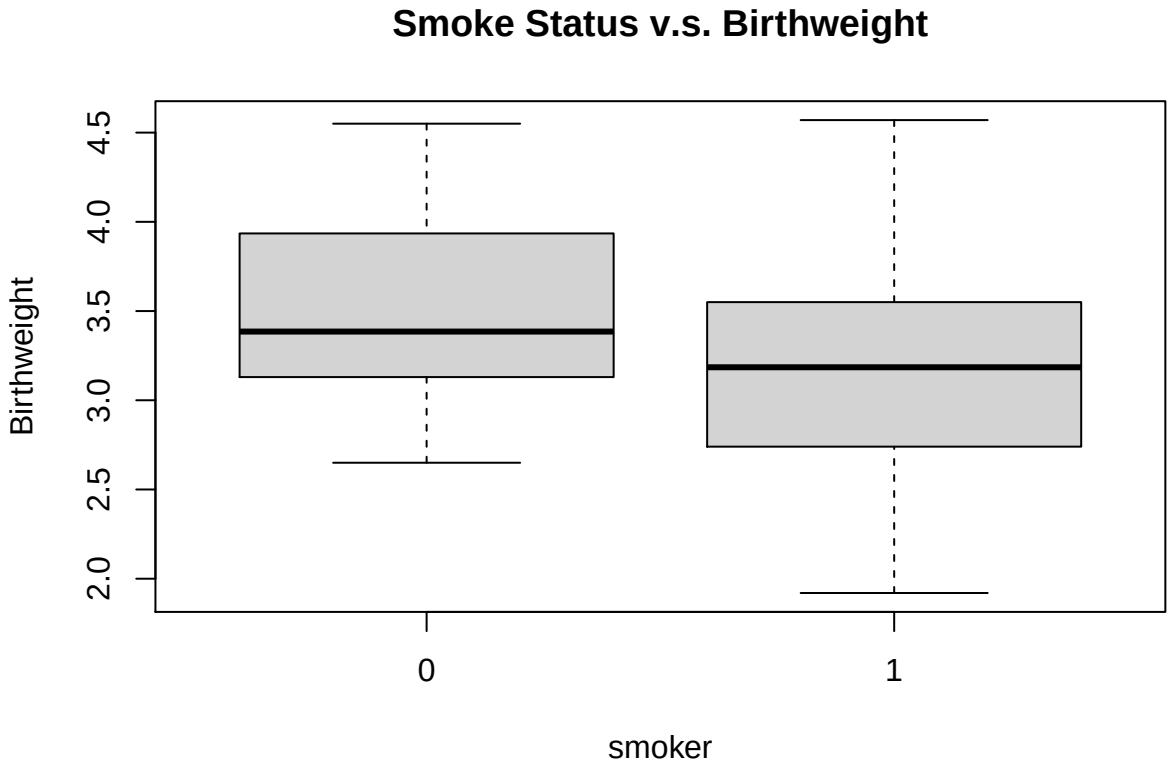
YiJin Chi, Youwei Wang, Shih-Chi Chen

2021/5/2

I. Introduction

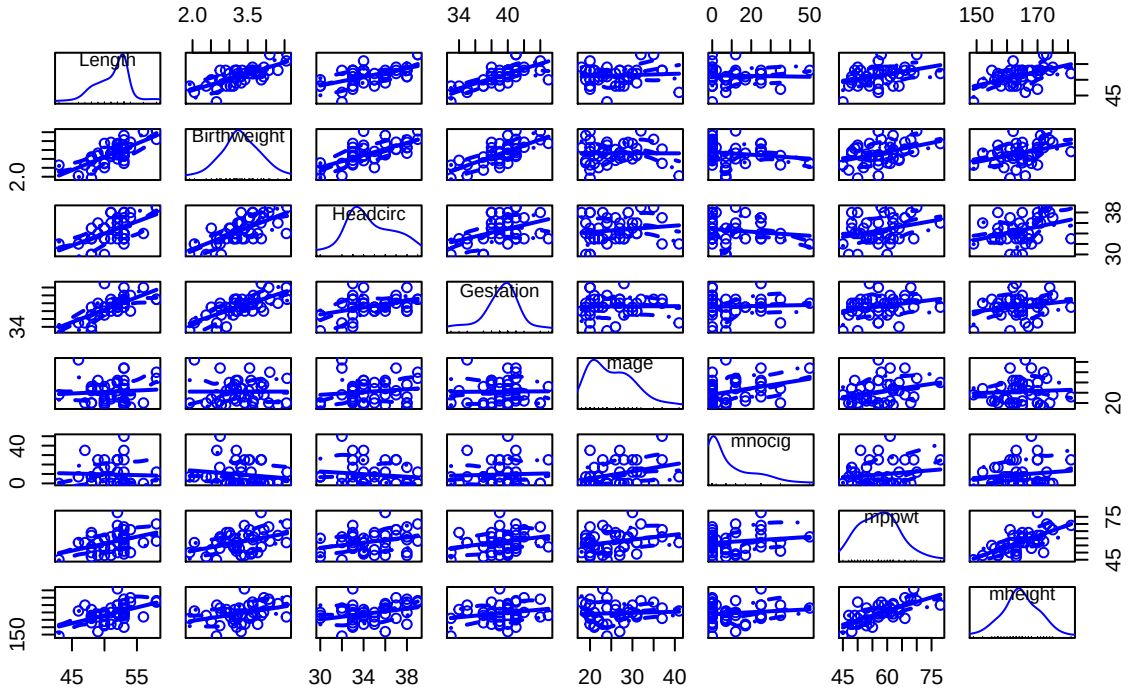
For this project, our goal is to find if there are any parents-related variables that will affect the babies' weight. Also, we are going to use the smoking status: 0 and 1 to split the data and run PCs on them to find the similarities and difference by comparing the PCs and plots.

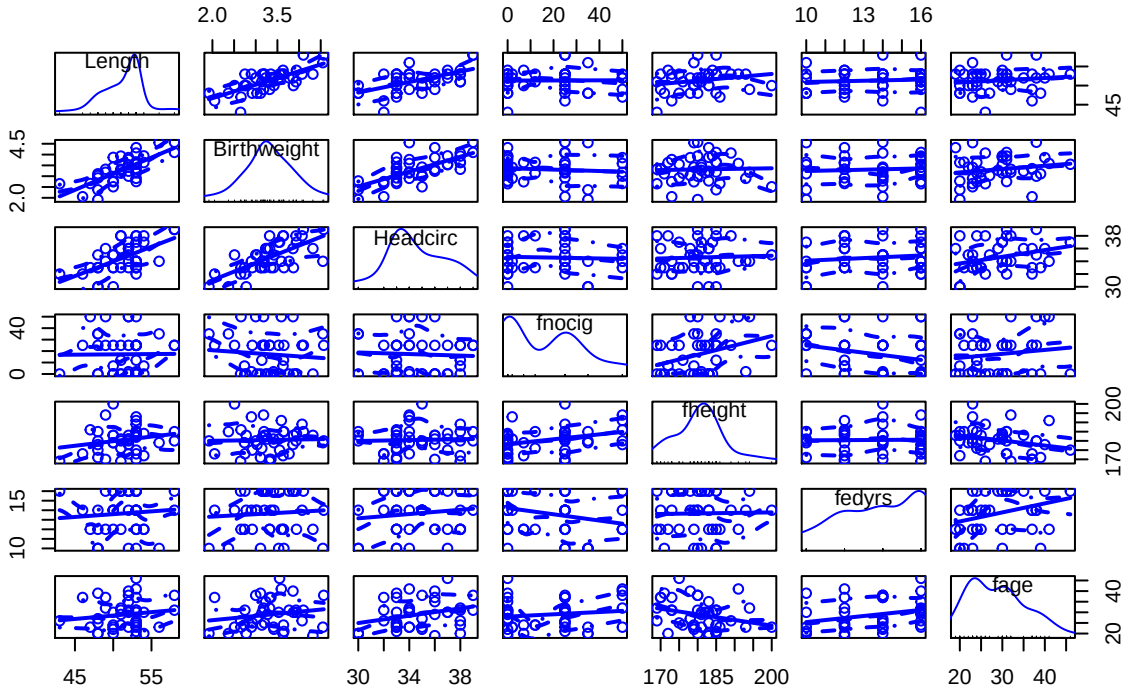
II. Summary of your data



As we can see from the above boxplot, mother's smoking status has an obvious effect on babies' birthweight. Here we want to explore what and how other parents' characteristics affect babies' birthweight.

Correlation matrix for Mother





The above two graphs are the correlation matrix for both Mother and Father to find if there are correlation between variables. From correlation matrix for Mother, the variables Gestation, mppwt (Mothers pre-pregnancy weight), and mheight (Mothers height) have obvious positive correlation with babies' Birthweight. From correlation matrix for Father, there is no obvious correlation found with babies' Birthweight.

III / IV. Analysis and Interpretation

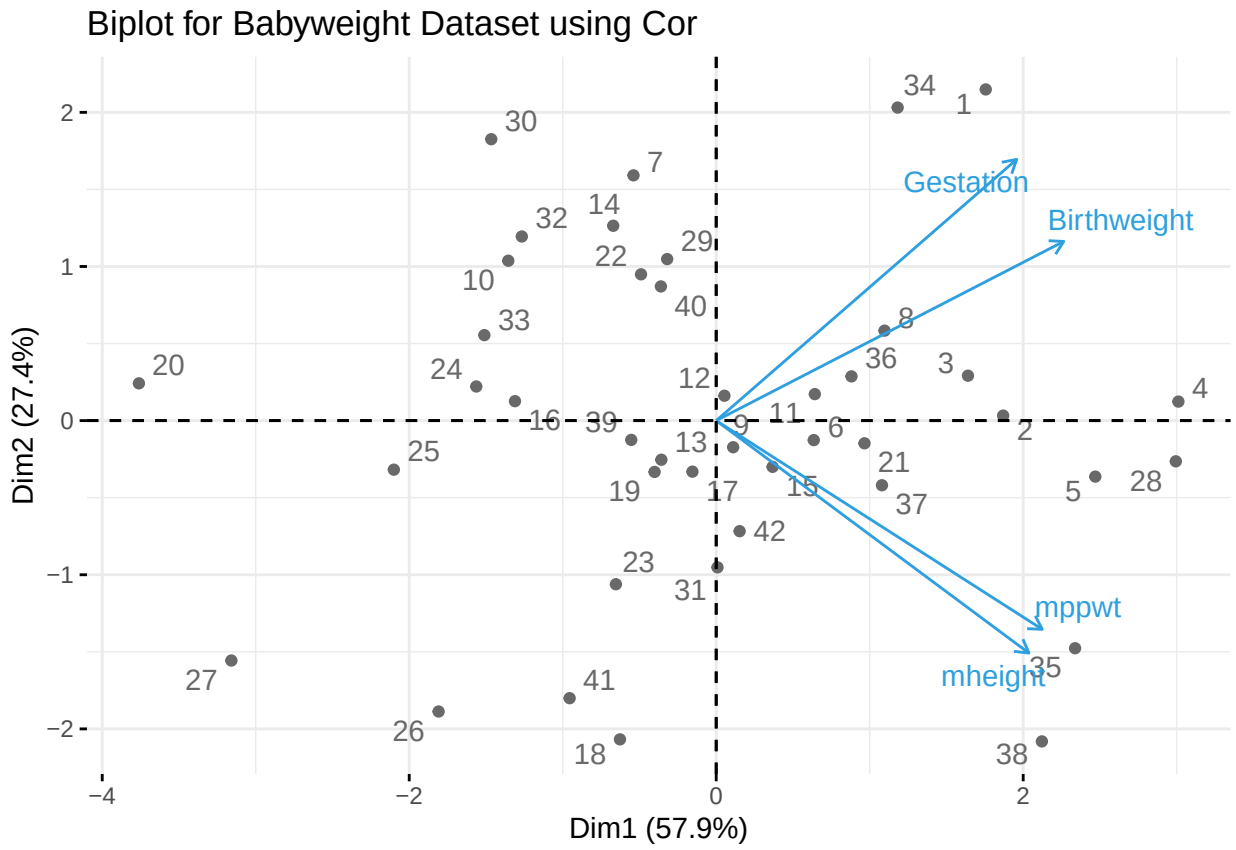
A. Run PCA for the variables: Birthweight, Gestation, mppwt, mheight

```
## Importance of components:
##                           Comp.1    Comp.2     Comp.3     Comp.4
## Standard deviation     1.5215805 1.0462713 0.56379933 0.52176575
## Proportion of Variance 0.5788018 0.2736709 0.07946742 0.06805987
## Cumulative Proportion  0.5788018 0.8524727 0.93194013 1.00000000
##
## Loadings:
##             Comp.1 Comp.2 Comp.3 Comp.4
## Birthweight  0.539  0.403  0.040  0.738
## Gestation    0.466  0.587  0.014 -0.662
## mppwt        0.506 -0.469 -0.720 -0.074
## mheight      0.485 -0.523  0.693 -0.107
```

Since the first three principle components explained over 90% of the total variation of sample data, we only maintain the first three here.

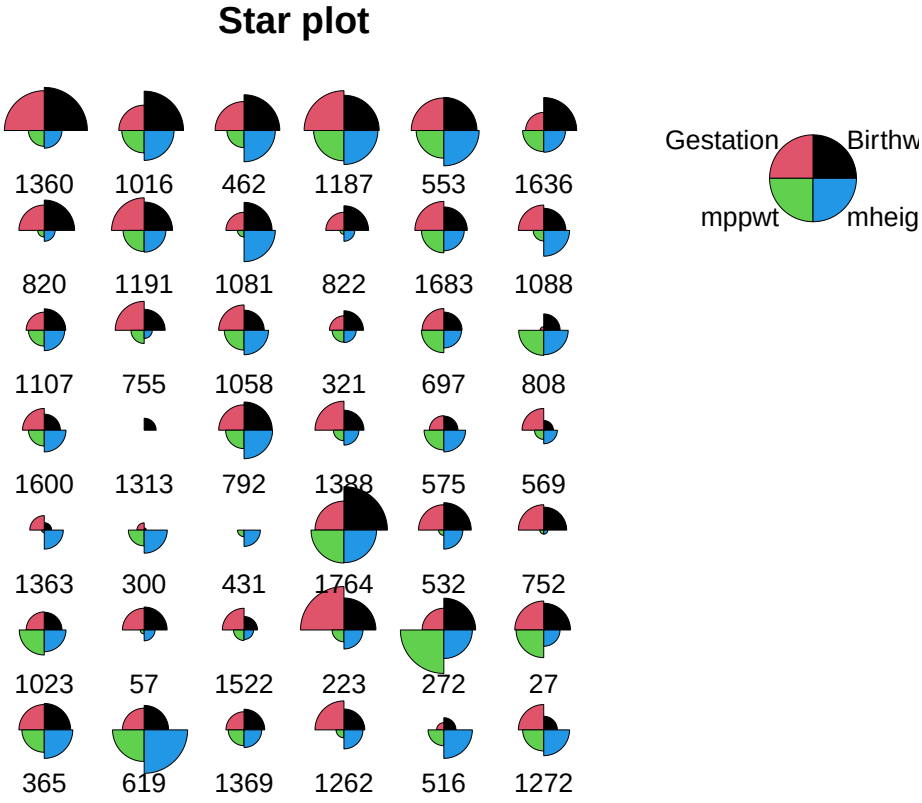
Let discuss the first three principle components. This is done using the outputs of the loadings. The first PC seems to be a weighted average of the four variables. However, the second PC shows an interesting result. It seems to compare Birthweight and Gestation to mppwt and mheight. The third PC put more weight on variable mppwt and mheight, while Birthweight and Gestation does not really play a large role in explaining the variation on PC3.

Biplot for the four variables



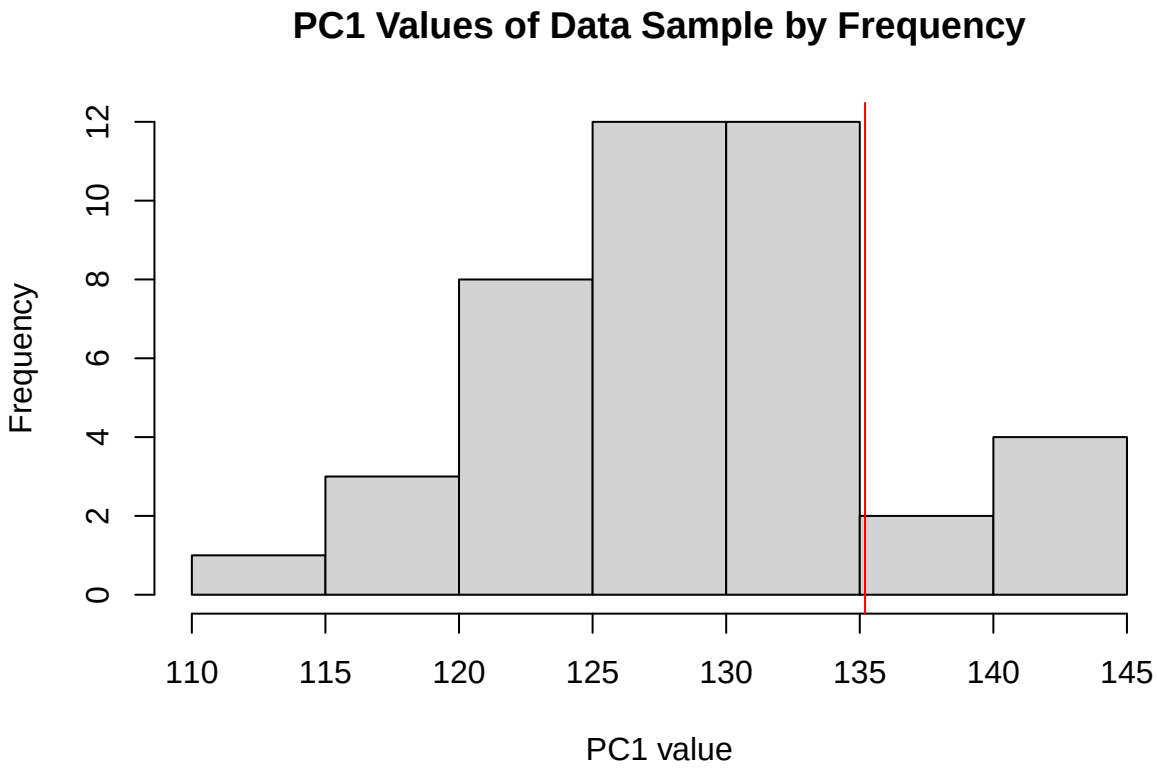
In the first dimension, all variables have similar attributes since they point to the same direction. In the second dimension, gestation and birthweight have the similar attributes while mppwt and mheight have the similar attributes in the opposite direction compared to the first two. We can find there is Birthweight has stronger correlation with Gestation compare to the other teo variables, which that the higher Gestation tend to have higher Birthweight since the eigenvalues give the same direction. Also, there is strong correlation between variables mppwt (Mothers pre-pregnancy weight) and mheight (Mothers height).

Starplot for the four variables



There are few babies that seem to be distinct from the rest, namely #1187, #553, #1764, #272. #300 and #431 have very small Birthweight for babies.

B. New born baby observation compared with given data samples



From the histogram plot, we can see the distribution of data samples' PC1 values tends to be symmetry. Most of the PC1 values falls into the range between 120 and 135, while the new baby observation is relevantly larger. Therefore, by adding the new baby to the data set, we may see an increasing in weight on the PC1. Note here, we only calculate the PC1 value based on the variables we picked.

C. Split the data by smoking status and run PCs on mother who are Non-smoker and Smoker

PCA Summary for Nonsmoker For running the PCA, we set the cutoff = 0.2 for better observing

```
## Importance of components:
##                           Comp.1    Comp.2    Comp.3     Comp.4
## Standard deviation     1.5973651 0.9083376 0.6335415 0.47113967
## Proportion of Variance 0.6378938 0.2062693 0.1003437 0.05549315
## Cumulative Proportion  0.6378938 0.8441631 0.9445069 1.00000000
##
## Loadings:
##             Comp.1 Comp.2 Comp.3 Comp.4
## Birthweight  0.496  0.543  0.392  0.553
## Gestation    0.511  0.436 -0.549 -0.497
## mppwt        0.478 -0.590 -0.450  0.470
## mheight      0.514 -0.408  0.585 -0.476
```

For the PCA for mothers who are non-smokers, the first 3 PCs are needed since they account for more than 94.45% of the variability in the data.

Since the first three principle components explained over 90% of the total variation of sample data, we only maintain the first three here.

Let discuss the first three principle components. This is done using the outputs of the loadings. The first PC seems to be a weighted average of the four variables. However, the second PC shows an interesting result. It seems to compare Birthweight and Gestation to mppwt and mheight. The third PC seems to compare Birthweight and mheight to mppwt and Gestation.

PCA Summary for Smoker

```
## Importance of components:
##                           Comp.1    Comp.2     Comp.3    Comp.4
## Standard deviation     1.4796028 1.1715641 0.52849666 0.3986282
## Proportion of Variance 0.5473061 0.3431406 0.06982718 0.0397261
## Cumulative Proportion  0.5473061 0.8904467 0.96027390 1.0000000
##
## Loadings:
##             Comp.1 Comp.2 Comp.3 Comp.4
## Birthweight  0.584  0.350         0.729
## Gestation    0.405  0.643 -0.225 -0.610
## mppwt        0.540 -0.400  0.672 -0.311
## mheight      0.451 -0.551 -0.702
```

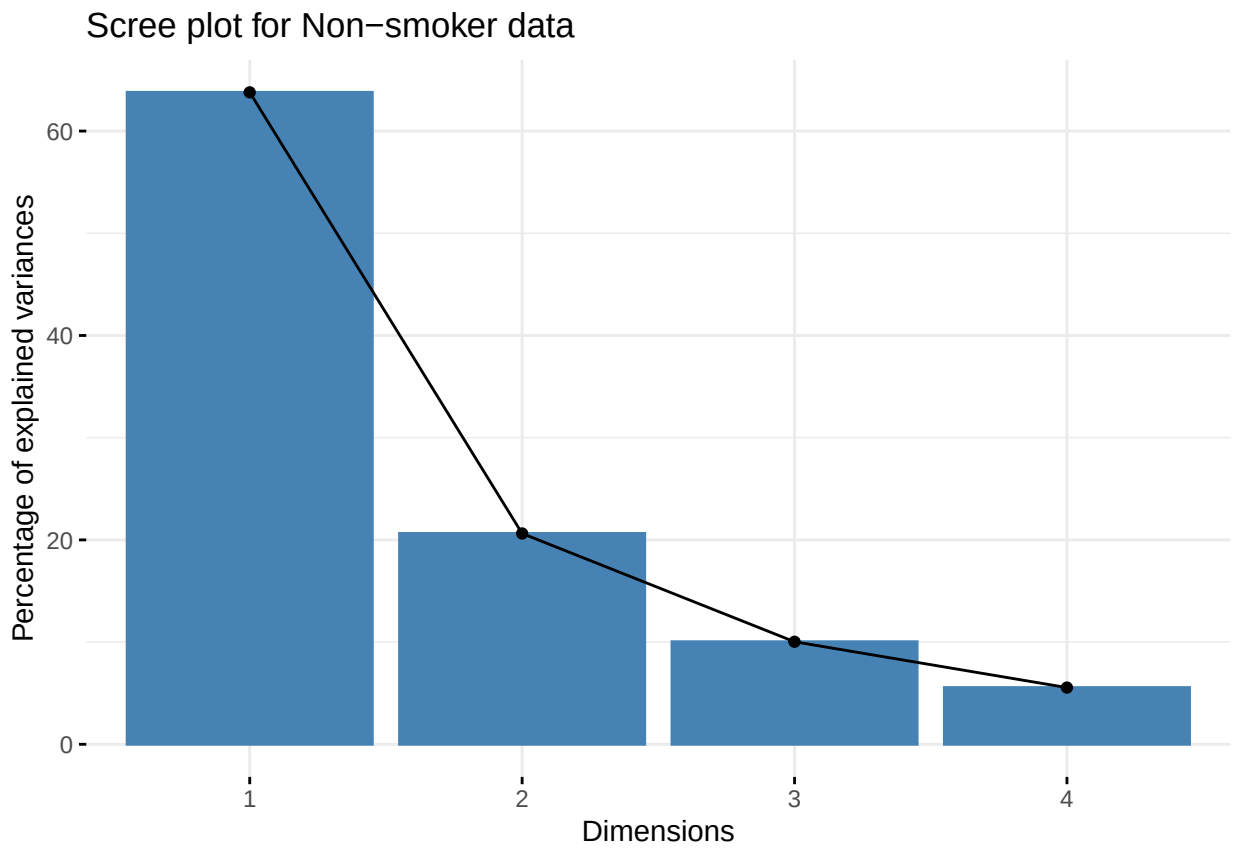
For the PCA for mothers who are smokers, the first 3 PCs are needed since they account for more than 96.02% of the variability in the data, which explains more of the variation of the data than mothers who are non-smokers

Since the first three principle components explained over 90% of the total variation of sample data, we only maintain the first three here.

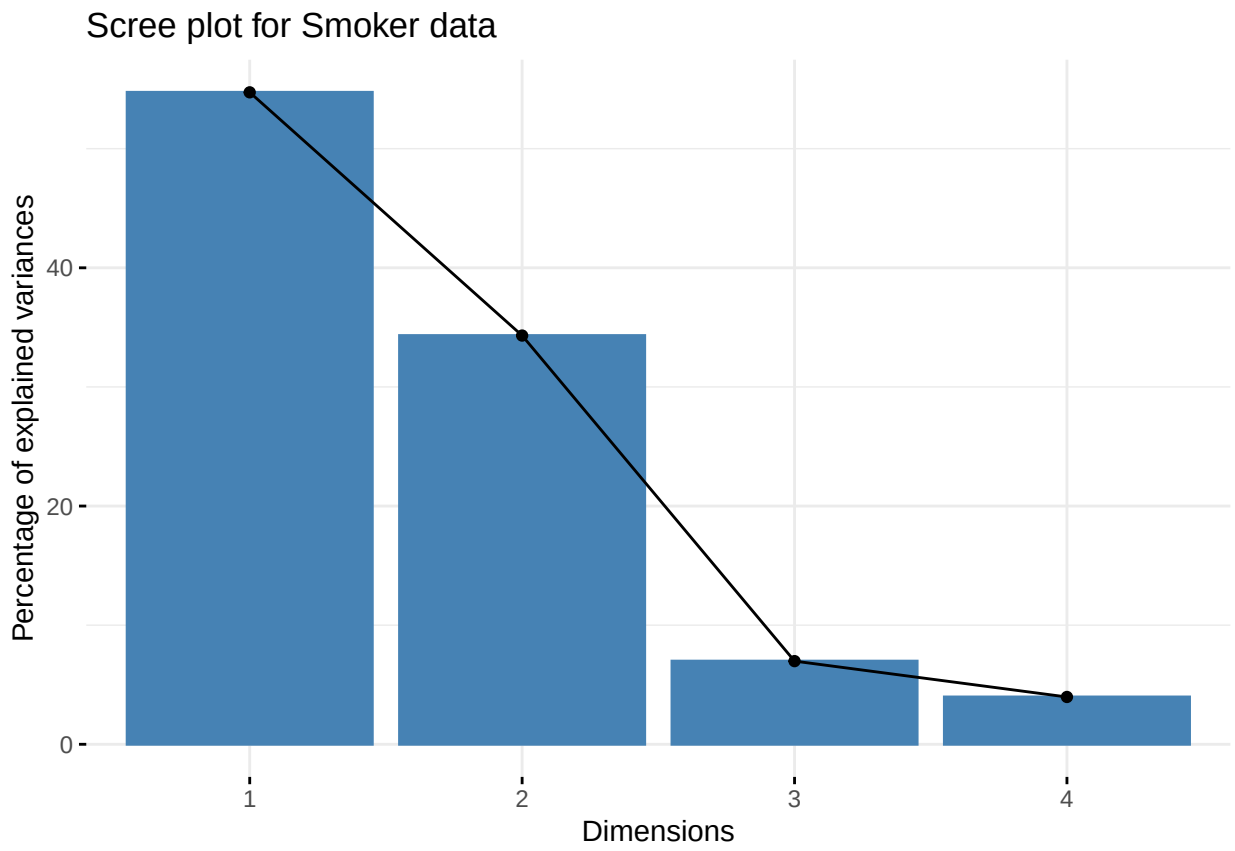
Let discuss the first three principle components. This is done using the outputs of the loadings. The first PC seems to be a weighted average of the four variables. However, the second PC shows an interesting result. It seems to compare Birthweight and Gestation to mppwt and mheight. The third PC put more weight on variable mppwt and mheight, while Birthweight does not really play a large role in explaining the variation on PC3.

Since the loading for birthweight in the third component is zero, other variables will not vary along with birthweight in the 3rd PC. Within each of the first two PCs, the magnitude of birthweight loading is very close the others', which means birthweight is sensitive to other variables variation.

Scree Plot for Nonsmoker

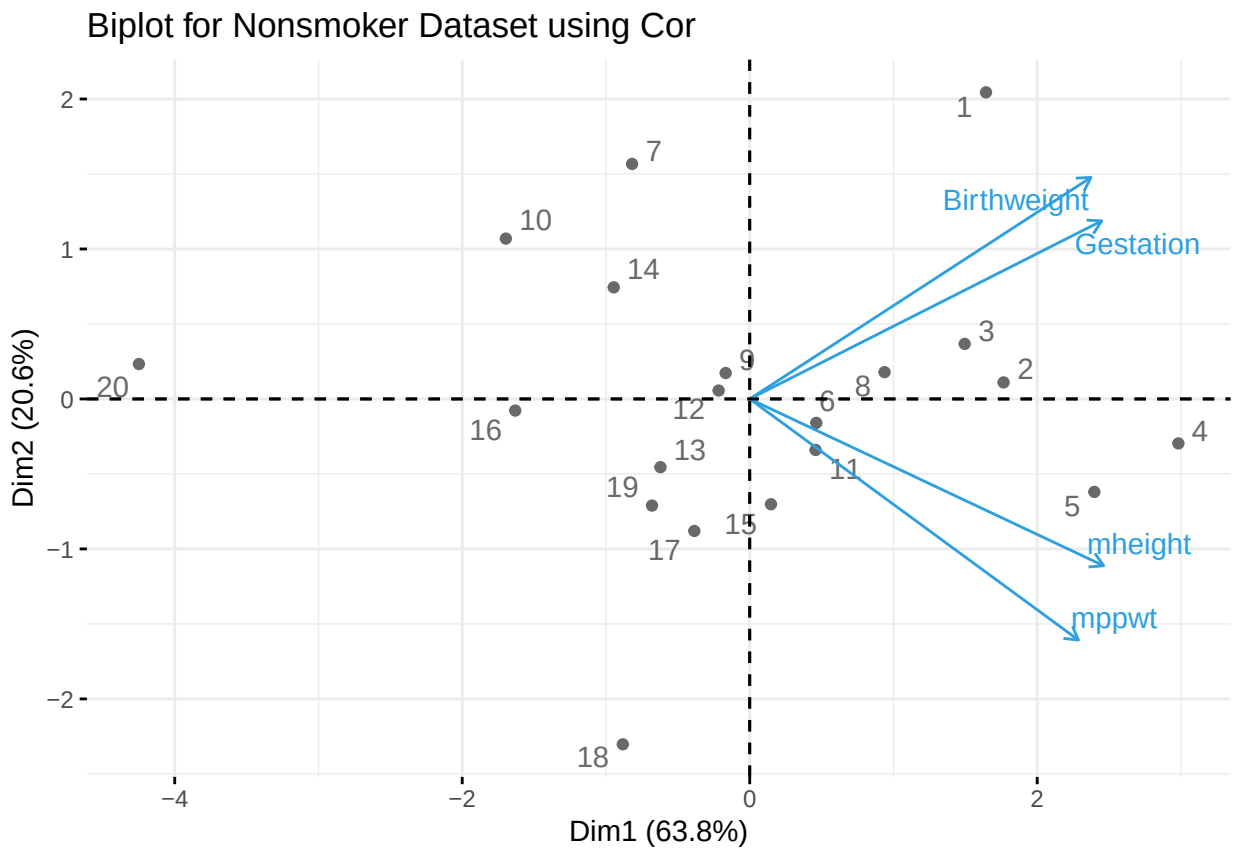


Scree Plot for Smoker



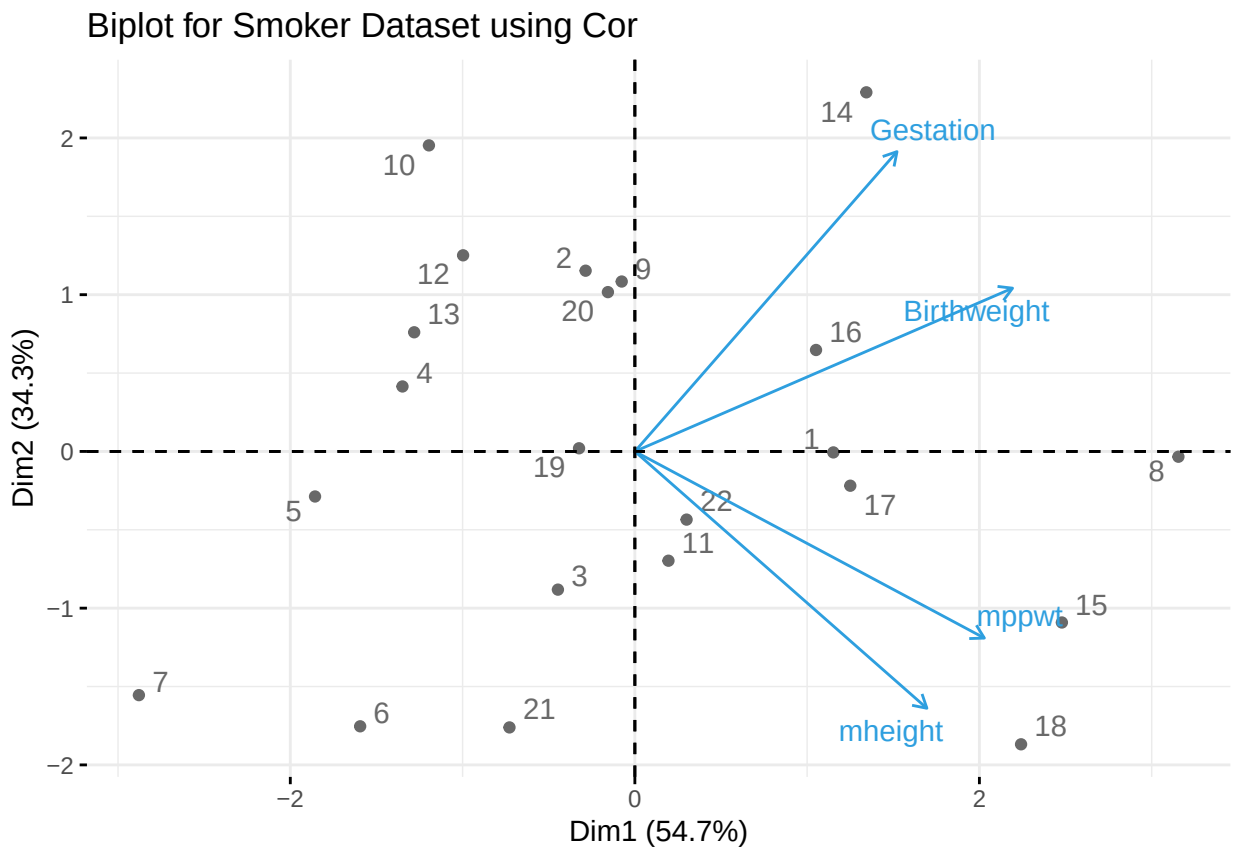
By comparing the two scree plot: first for non-smoker and second for smoker, we can find that for the non-smoker, the PC1 account for more variation of the data (nearly 63.79%), while for the smoker, PC1 account for only nearly 54.73% variation of the data. However, PC2 explain less variation (20.63%) of the data for non-smoker compare to smoker (34.31%).

Plot the first two PCs: Nonsmoker



In the first dimension, all variables have similar attributes since they point to the same direction. In the second dimension, gestation and birthweight have the similar attributes while mppwt and mheight have the similar attributes in the opposite direction compared to the first two. We can find there is Birthweight has stronger correlation with Gestation compare to the other two variables, which that the higher Gestation tend to have higher Birthweight since the eigenvalues give the same direction. Also, there is strong correlation between variables mppwt (Mothers pre-pregnancy weight) and mheight (Mothers height).

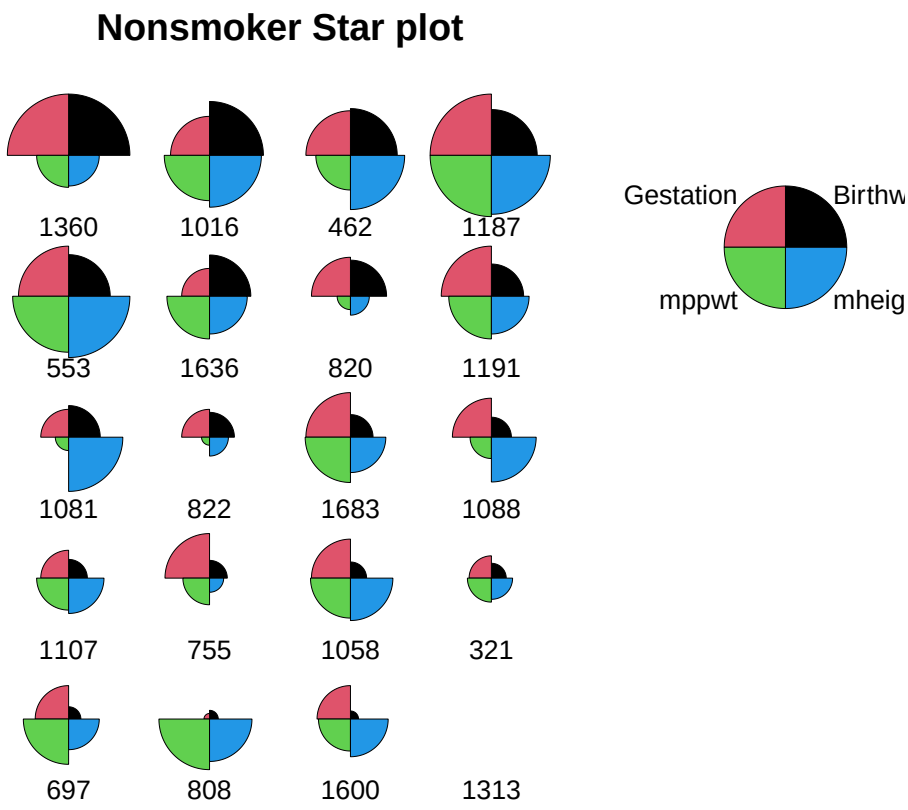
Plot the first two PCs: Smoker



The interpretation of the plot for smoker is similar to the plot for non-smoker.

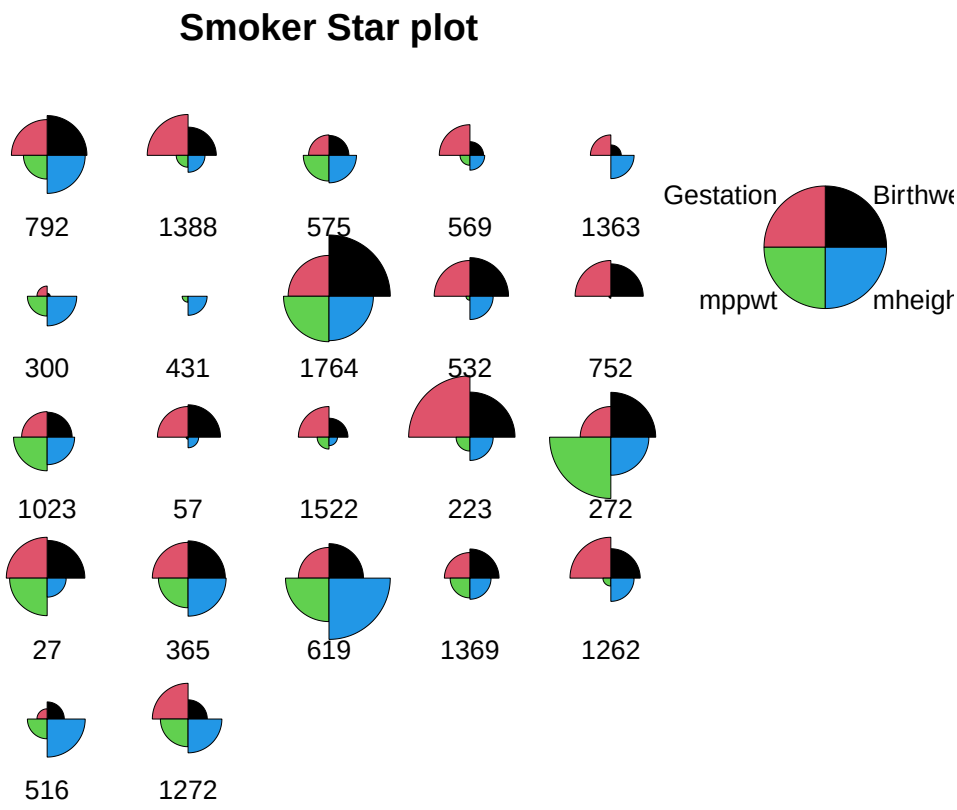
However, by comparing the Biplot for Nonsmoker and Smoker, we can find the correlation between Gestation and Birthweight for nonsmoker is stronger than that for smoker because the angle between the vectors for Gestation and Birthweight is narrower in the plot for nonsmoker. Also, higher Gestation tend to have higher Birthweight for babies since the eigenvectors gives the same direction. On the other hand, Gestation is most correlated to Birthweight for both nonsmoker and smoker, while for nonsmoker, mheight has the second strongest correlation with Birthweight and for smoker, mppwt has the second strongest correlation with Birthweight.

Star Plot for Non-Smoker



For Nonsmoker mother, there are few babies that seem to be distinct from the rest, namely #1360 which babies seems to have large Birthweight, and for #1313, babies seem to have small Birthweight.

Star Plot for Smoker



For Smoker mother, there are few babies that seem to be distinct from the rest, namely #1764 which babies seems to have large Birthweight, and for #431, babies seem to have small Birthweight.

V. Conclusion

In summary, we find out that variables Gestation has the strongest positive correlation with Birthweight for babies discard mother's smoke status. However, by comparing the three Biplot: one both smoke and nonsmoke, one for nonsmoker, and one for smoker, one interesting thing that worth to notice is that, among those babies given birth by non-smoker mothers, the correlation between gestation and birthweight is stronger that those given birth by smoker mothers.

One possible future work that we want to investigate is that for mothers who are nonsmokers, we find there is very small correlation between Mothers pre-pregnancy weig (mppwt) and Birthweight. It is worth to further run statistical test to check the accurancy of this finding.

Appendix Code

```
knitr::opts_chunk$set(echo = F, fig.align = 'center')
# read data
library(dplyr)
Birthweight_data = read.csv("C:/Users/youwe/Desktop/STA 135/project/project 1/Birthweight_reduced_kg_R.
boxplot(Birthweight~smoker,data=Birthweight_data, main="Smoke Status v.s. Birthweight")
```

```r
library(car)
# Mother
scatterplotMatrix(formula = ~ Length+Birthweight+Headcirc+Gestation+mage+mnocig+mppwt+mheight,
                  data = Birthweight_data, reg.line = lm, smooth = TRUE, span = 0.5,
                  diagonal = "histogram", main="Correlation matrix for Mother" )

# Father
scatterplotMatrix(formula = ~ Length+Birthweight+Headcirc+fnocig+fheight+fedyrs+fage , data = Birthweigh
diagonal = "histogram", main="Correlation matrix for Father")
pca.save <- princomp(formula = ~ Birthweight + Gestation + mppwt + mheight,
                     data = Birthweight_data,
                     cor = TRUE, scores = TRUE)
summary(pca.save, loadings = TRUE, cutoff = 0.0)
library(factoextra)
fviz_pca_biplot(pca.save, repel = TRUE,
                col.var = "#2E9FDF", # Variables color
                col.ind = "#696969", # Individuals color
                title = "Biplot for Babyweight Dataset using Cor"
                )
the.data2 = Birthweight_data %>%
  select(Birthweight,Gestation, mppwt, mheight)

stars(x = the.data2, draw.segments = TRUE, key.loc =
c(20,15), main = "Star plot", labels = Birthweight_data$ID)
x1=Birthweight_data$Birthweight
x2=Birthweight_data$Gestation
x3=Birthweight_data$mppwt
x4=Birthweight_data$mheight

y=0.539*x1+0.466*x2+0.506*x3+0.485*x4

y_new = 0.539*5.1 + 0.466*43 + 0.506*64 + 0.485*165

hist(y, xlab = "PC1 value", main = "PC1 Values of Data Sample by Frequency")
abline(v = y_new, col = "red")
nonsmoker_mom = Birthweight_data %>%
  filter(smoker == "0") %>%
  select(Birthweight,Gestation, mppwt, mheight)
smoker_mom = Birthweight_data %>%
  filter(smoker == "1") %>%
  select(Birthweight,Gestation, mppwt, mheight)
pca.save_nonsmoker <- princomp(formula = ~  Birthweight + Gestation + mppwt+ mheight,
                               data = nonsmoker_mom, cor = TRUE, scores = TRUE)
summary(pca.save_nonsmoker, loadings = TRUE, cutoff = 0.2)
pca.save_smoker <- princomp(formula = ~ Birthweight + Gestation + mppwt+ mheight,
                            data = smoker_mom, cor = TRUE, scores = TRUE)
summary(pca.save_smoker, loadings = TRUE, cutoff = 0.2)
# plot(pca.save, type = "lines", main = "Scree plot for Non-smoker data")
fviz_eig(pca.save_nonsmoker, main = "Scree plot for Non-smoker data")
# plot(pca.save, type = "lines", main = "Scree plot for Smoker data")
library(factoextra)
fviz_eig(pca.save_smoker, main = "Scree plot for Smoker data")
library(factoextra)
```

```
fviz_pca_biplot(pca.save_nonsmoker, repel = TRUE,
                col.var = "#2E9FDF", # Variables color
                col.ind = "#696969", # Individuals color
                title = "Biplot for Nonsmoker Dataset using Cor"
                )
library(factoextra)
fviz_pca_biplot(pca.save_smoker, repel = TRUE,
                col.var = "#2E9FDF", # Variables color
                col.ind = "#696969", # Individuals color
                title = "Biplot for Smoker Dataset using Cor"
                )
the.data_nonsmoker = Birthweight_data %>%
  filter(smoker == "0")
stars(x = nonsmoker_mom, draw.segments = TRUE, key.loc =
c(14,10), main = "Nonsmoker Star plot", labels = the.data_nonsmoker$ID)
the.data_smoker = Birthweight_data %>%
  filter(smoker == "1")
stars(x = smoker_mom, draw.segments = TRUE, key.loc =
c(15,10), main = "Smoker Star plot", labels = the.data_smoker$ID)
```