Teaching Model 1.5

Module5

2023-04-20

5 Batch Effect and Factor Analysis

In this module, we will consider batch effects, which are used to explain the differences in effects between groups and within groups for model construction. We still want to construct a linear model, but this model will be different from the previous linear models (fit and fit_) where we found high errors during cross-validation, leading to uncertainty about the models we have constructed.In this section, we will verify how this error occur using batch effect analysis. Additionally, we will also investigate what caused the big difference in variances that we observed in Module 2.1 in terms of gender.

```
library(rafalib)
library(dplyr)
```

```
## Warning: package ’dplyr’ was built under R version 4.2.2
```

```
##
## Attaching package: ’dplyr’
```

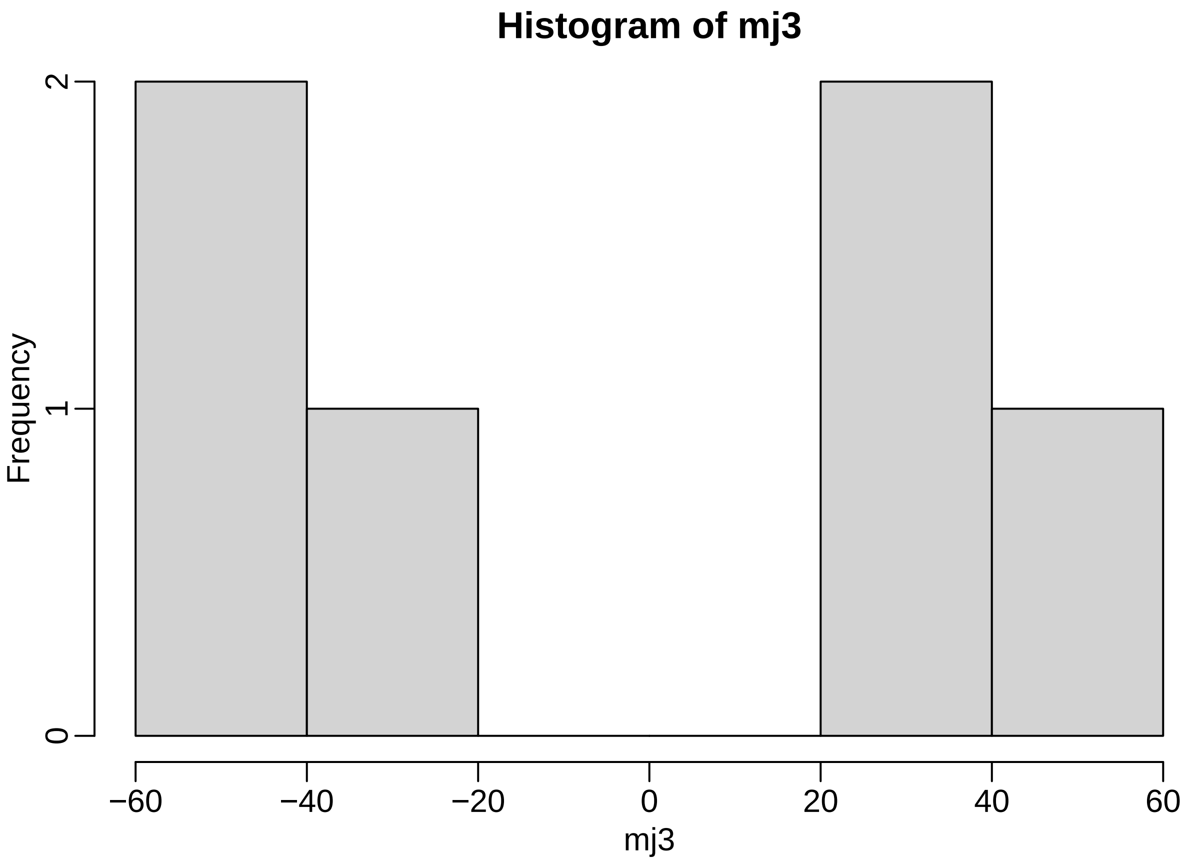
```
## The following objects are masked from ’package:stats’:
##
##     filter, lag
```

```
## The following objects are masked from ’package:base’:
##
##     intersect, setdiff, setequal, union
```

```
colgtotal<- read.csv("C:/Users/Jing Xie/Documents/R/Teaching Project/Proj 1/Data/colgtotal.csv")

colgmx <- apply(as.matrix(colgtotal), 2, as.numeric)
colgmx <- colgmx[,6:11]

colgmx <- colgmx - rowMeans(colgmx)
colgmj3 <- colgmx[which(colgtotal$Major == '3'),]
mj3 <- colMeans(colgmx[which(colgtotal$Major == '3'),])
```

The follow histogram describes the performance of female and male regarding the mj3 group.

```
mypar()
hist(mj3)
```



This plot can help explain the variance difference that we mentioned in module 1.1 for major 3 where we found the negative change was predominately coming from male,

```
colgtotalmx <- as.matrix(na.omit(colgtotal[,2:12]))
colgtotalmx<-colgtotalmx[,1:11] - rowMeans(colgtotalmx[,1:11])
```

```
gd<- colgtotal$Gender
s <- svd(colgmx)
st <- svd(colgtotalmx)
sv<- svd(colgmj3)
```

```
What <- t(st$v[,1:2])
colnames(What)<-colnames(colgtotalmx)
round(What,2)
```

```
##      Gender  Race Goal Major Project quiz1 Mid.Test   HW Participation
## [1,]   0.25  0.23 0.23  0.23    0.24 -0.56    -0.40 0.08          0.20
## [2,]  -0.02 -0.05 0.01 -0.05   -0.03  0.53    -0.71 0.01          0.01
##      Final.Score    fm
## [1,]       -0.45 -0.05
## [2,]       -0.14  0.43
```

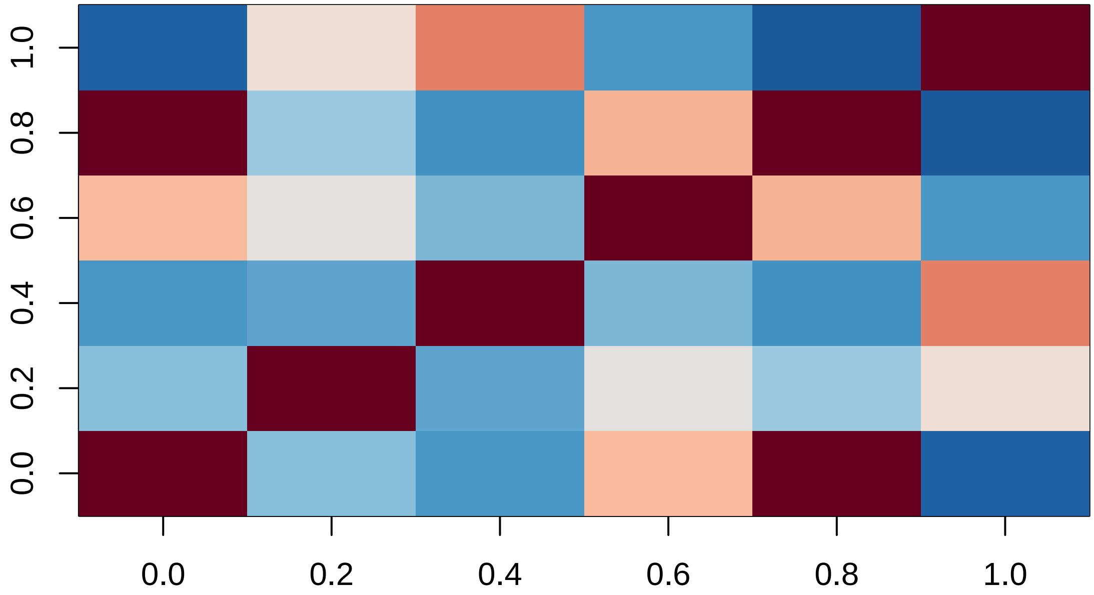
The table shows the correlation between the abilities of all the features and their contribution to the project, as well as the correlation between the difference in abilities based on gender and their contribution to the project.

```
colgtotalmx <- as.matrix(na.omit(colgtotal[,2:12]))
colg<-colgtotalmx[,2:11] - rowMeans(colgtotalmx[,2:11])
fitup = st$u[,1:2]%*% (st$d[1:2]*What)
var(as.vector(fitup))/var(as.vector(colgtotalmx))
```

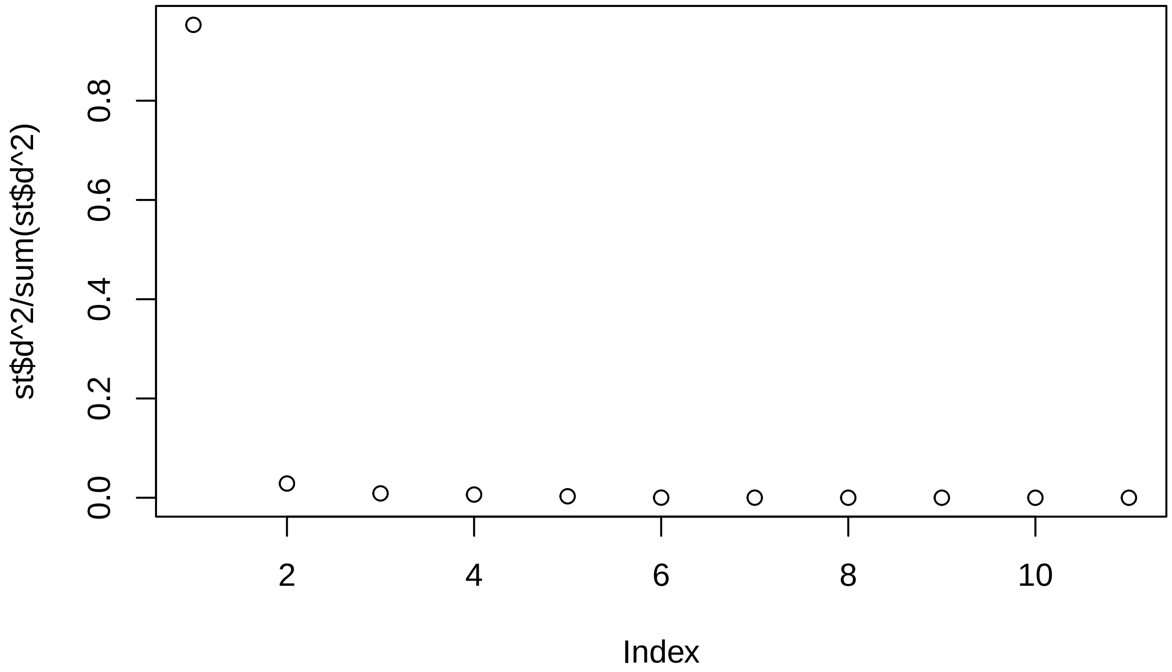
```
## [1] 0.9577919
```

To construct a linear model, we look at the first two components, which help explain around 95.8 percent of the data.That is, we use factor analysis and PCA both to construct the model fitup. We can also add other remaining components to improve the model accuracy.

```
library(RColorBrewer)
cols=colorRampPalette(rev(brewer.pal(10,"RdBu")))(70)
image ( cor(colgmx) ,col=cols,zlim=c(-1,1))
```

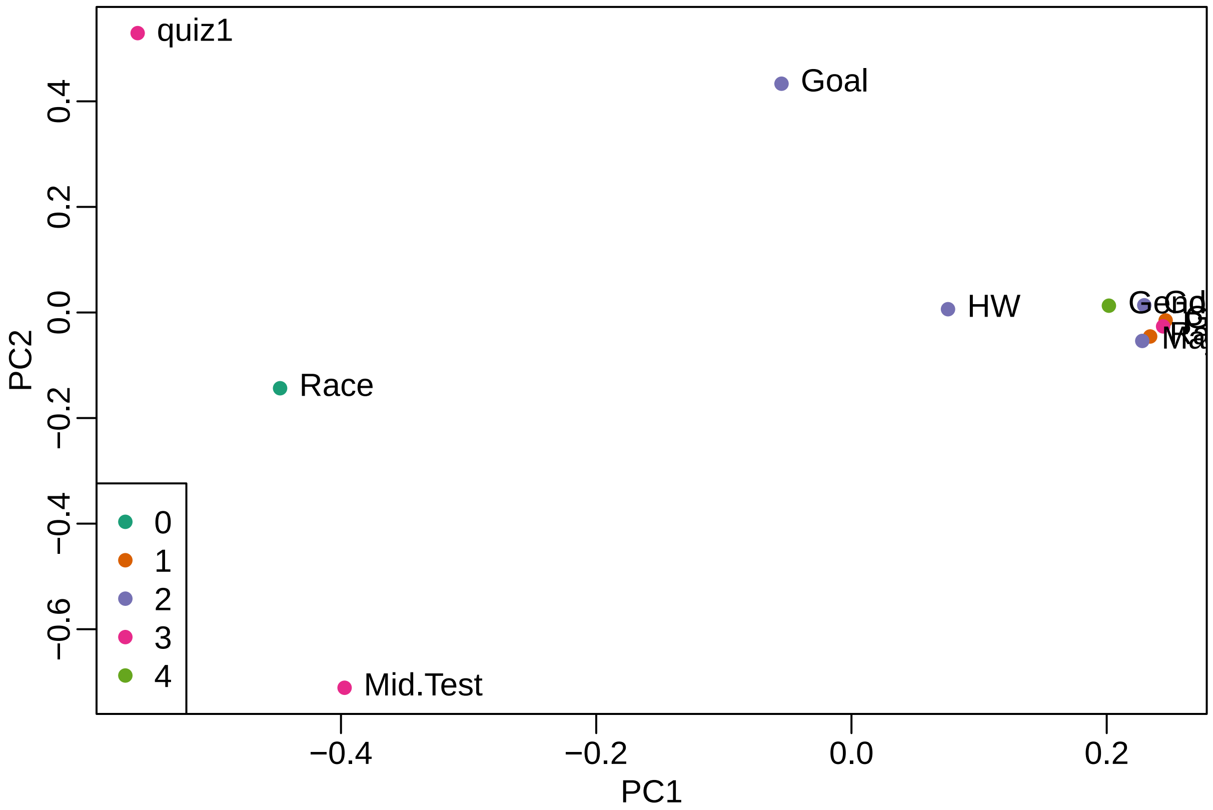


```
plot(st$d^2/sum(st$d^2))
```



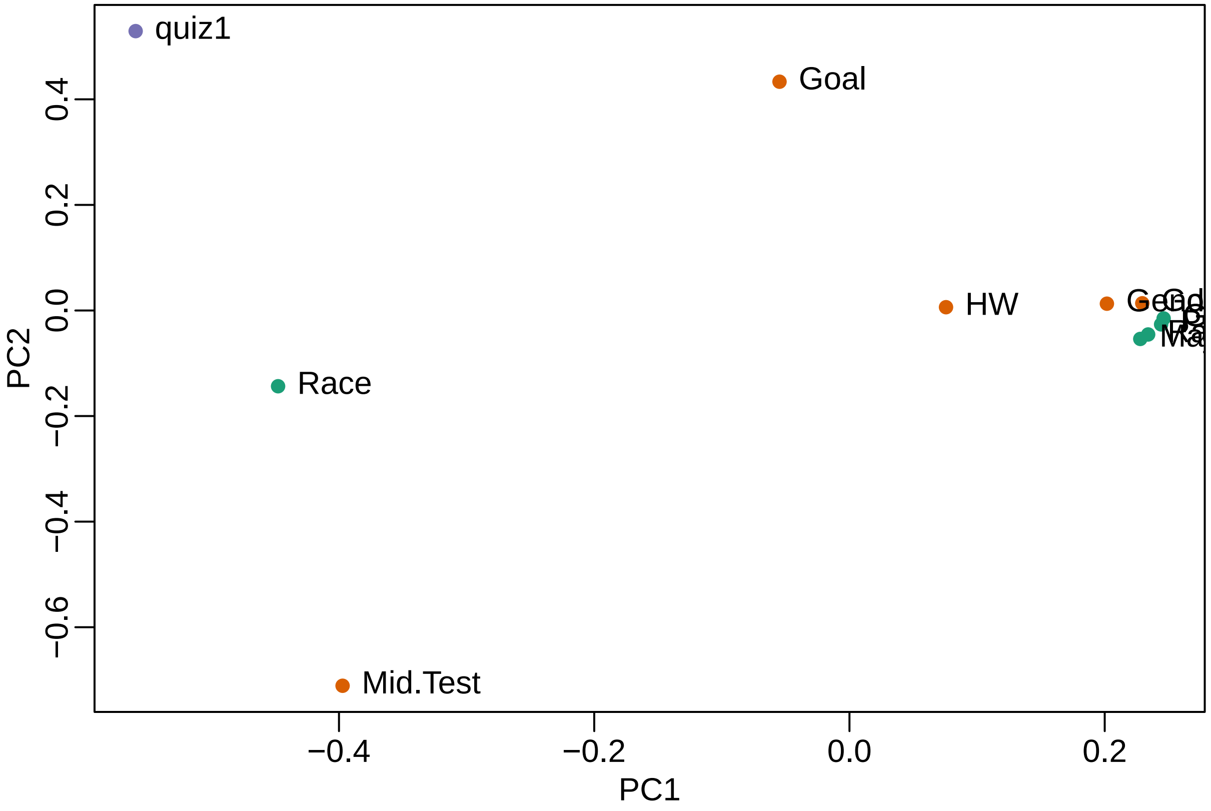
The following are the demonstration of MDS to our interest variable and the principal components(PCs) to observe their relationship.

```
mj <- factor(colgtotal$Major)
cols = as.numeric(mj)
library(rafalib)
mypar()
plot(st$v[,1],st$v[,2],col=cols,pch=16,
     xlab="PC1",ylab="PC2")
Features <- c("Gender","Race","Goal", "Major", "Project", "quiz1", "Mid.Test", "HW")
text(st$v[,1],st$v[,2], pos = 4, labels = Features)

legend("bottomleft",levels(mj),col=seq(along=levels(mj)),pch=16)
```



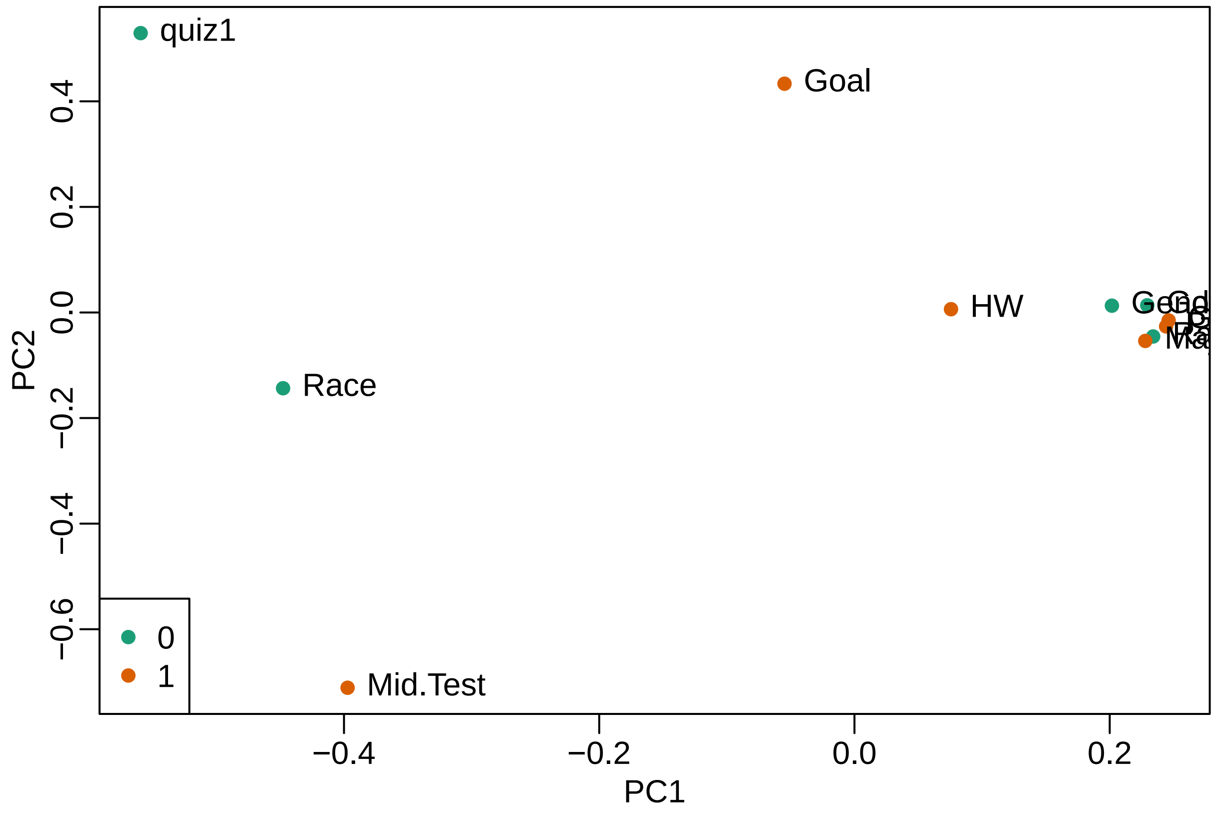
This plot above shows apparent negative relationship between the feature variables and pcs including quiz1, Race, Mid.Test, while positive in Gender, Goal, and other varietals. Goal and Major have the similarity.

```
rc <- factor(colgtotal$Race)
cols = as.numeric(rc)
mypar()
plot(st$v[,1],st$v[,2],col=cols,pch=16,
     xlab="PC1",ylab="PC2")
Features <- c("Gender","Race","Goal", "Major", "Project", "quiz1", "Mid.Test", "HW")
text(st$v[,1],st$v[,2], pos = 4, labels = Features)
```



```
#legend("bottomleft",levels(rc),col=seq(along=levels(rc)),pch=16)
```

```
gd <- factor(colgtotal$Gender)
cols = as.numeric(gd)
mypar()
plot(st$v[,1],st$v[,2],col=cols,pch=16,
     xlab="PC1",ylab="PC2")
Features <- c("Gender","Race","Goal", "Major", "Project", "quiz1", "Mid.Test", "HW")
text(st$v[,1],st$v[,2], pos = 4, labels = Features)
legend("bottomleft",levels(gd),col=seq(along=levels(gd)),pch=16)
```



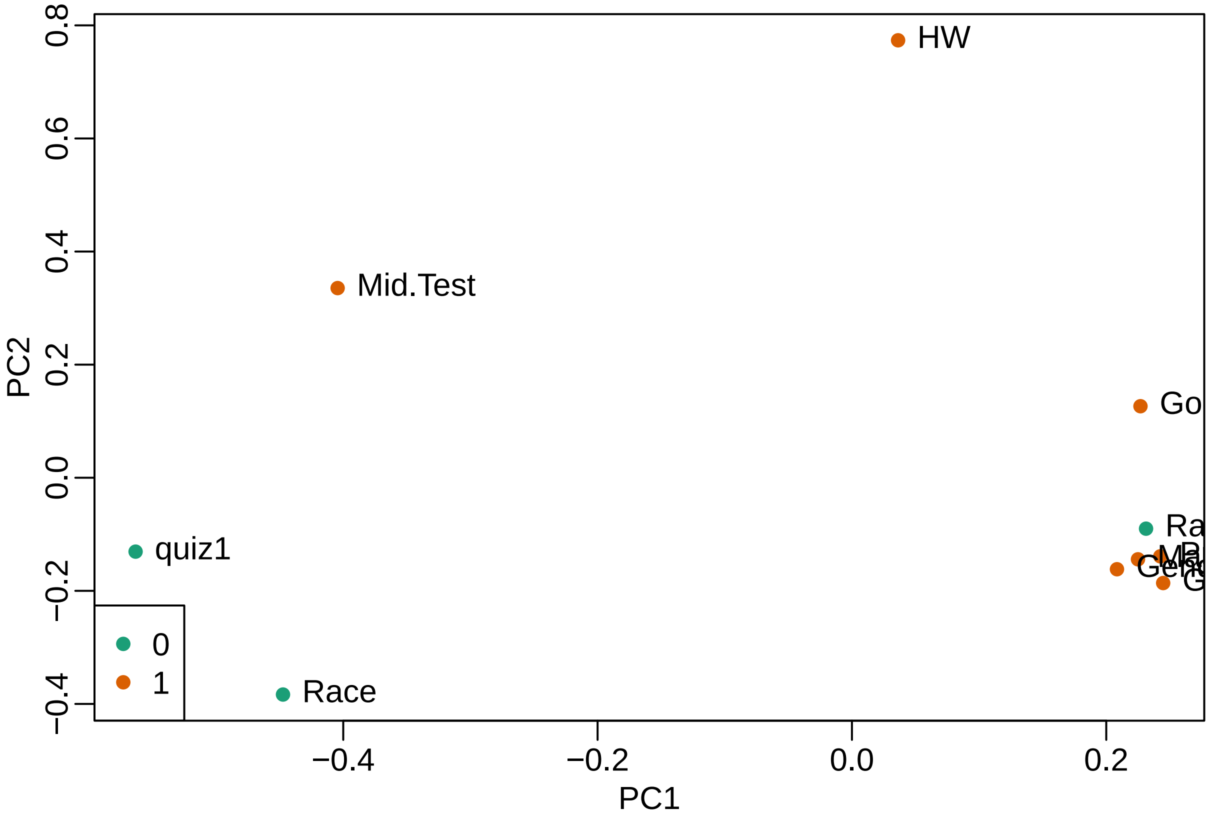
We may have a confounding issue that involves up to five factors, making it difficult to construct a linear model. So far, we have detected errors in the linear models that we previously assumed

```
b <- na.omit(colgtotal[,2:12])
colg <- apply(as.matrix(b), 2, as.numeric)
colg<-colg[,2:11] - rowMeans(colg[,2:11])

sv<- svd(colgmj3)
```

```
colgmj3mt <- filter(colgtotal, Major =='3')

colgmj3 <- apply(as.matrix(colgmj3mt), 2, as.numeric)
colgmj3 <- colgmj3[,2:11]

colgmj3 <- colgmj3 - rowMeans(colgmj3)

sv <- svd(colgmj3)
```

```
gd <- factor(colgmj3mt$Gender)
cols = as.numeric(gd)
library(rafalib)
mypar()
plot(sv$v[,1],sv$v[,2],col=cols,pch=16,
     xlab="PC1",ylab="PC2")
Features <- c("Gender","Race","Goal", "Major", "Project", "quiz1", "Mid.Test", "HW")
text(sv$v[,1],sv$v[,2], pos = 4, labels = Features)
```

```
legend("bottomleft",levels(gd),col=seq(along=levels(gd)),pch=16)
```



In addition, if we look at the mj3, the plot shows an apparent change compared to the entire data set. There is a clear correlation between the first principal component and gender, which can be used to explain the findings in modular1, where significant variance was caused by gender in particular subjects. We can observe that HW, Mid-Test, and Quiz1 are dissimilar.