Module\_03 Solution

Romil Shah

June 5, 2016

### Load BLS data and needed additional packages

require(ggplot2)

## Loading required package: ggplot2

## Warning: package 'ggplot2' was built under R version 3.2.5

require(reshape2)

## Loading required package: reshape2

## Warning: package 'reshape2' was built under R version 3.2.5

require(psych)

## Loading required package: psych

## Warning: package 'psych' was built under R version 3.2.5

##   
## Attaching package: 'psych'

## The following objects are masked from 'package:ggplot2':  
##   
## %+%, alpha

memory.limit(10000)

## [1] 10000

BLS<-read.csv("C:/Users/rams1/Desktop/DSCS6030/Module\_03/2014.annual.singlefile.csv")  
head(BLS)

## area\_fips own\_code industry\_code agglvl\_code size\_code year qtr  
## 1 01000 0 10 50 0 2014 A  
## 2 01000 1 10 51 0 2014 A  
## 3 01000 1 102 52 0 2014 A  
## 4 01000 1 1021 53 0 2014 A  
## 5 01000 1 1022 53 0 2014 A  
## 6 01000 1 1023 53 0 2014 A  
## disclosure\_code annual\_avg\_estabs annual\_avg\_emplvl total\_annual\_wages  
## 1 117452 1863561 80668352987  
## 2 1186 53491 4148191291  
## 3 1186 53491 4148191291  
## 4 587 11462 719550831  
## 5 2 13 430575  
## 6 17 147 11630538  
## taxable\_annual\_wages annual\_contributions annual\_avg\_wkly\_wage  
## 1 13917605638 316127565 832  
## 2 0 0 1491  
## 3 0 0 1491  
## 4 0 0 1207  
## 5 0 0 649  
## 6 0 0 1527  
## avg\_annual\_pay lq\_disclosure\_code lq\_annual\_avg\_estabs  
## 1 43287 1.00  
## 2 77550 1.55  
## 3 77550 1.55  
## 4 62776 1.55  
## 5 33771 1.24  
## 6 79389 1.70  
## lq\_annual\_avg\_emplvl lq\_total\_annual\_wages lq\_taxable\_annual\_wages  
## 1 1.00 1.00 1  
## 2 1.44 1.74 0  
## 3 1.46 1.77 0  
## 4 1.26 1.60 0  
## 5 0.14 0.08 0  
## 6 0.79 0.69 0  
## lq\_annual\_contributions lq\_annual\_avg\_wkly\_wage lq\_avg\_annual\_pay  
## 1 1 1.00 1.00  
## 2 0 1.21 1.21  
## 3 0 1.22 1.21  
## 4 0 1.27 1.27  
## 5 0 0.55 0.55  
## 6 0 0.88 0.88  
## oty\_disclosure\_code oty\_annual\_avg\_estabs\_chg  
## 1 1394  
## 2 -10  
## 3 -10  
## 4 0  
## 5 0  
## 6 0  
## oty\_annual\_avg\_estabs\_pct\_chg oty\_annual\_avg\_emplvl\_chg  
## 1 1.2 18475  
## 2 -0.8 -1097  
## 3 -0.8 -1097  
## 4 0.0 -75  
## 5 0.0 0  
## 6 0.0 -7  
## oty\_annual\_avg\_emplvl\_pct\_chg oty\_total\_annual\_wages\_chg  
## 1 1.0 2665745136  
## 2 -2.0 97930469  
## 3 -2.0 97930469  
## 4 -0.7 24761729  
## 5 0.0 -7288  
## 6 -4.5 -263697  
## oty\_total\_annual\_wages\_pct\_chg oty\_taxable\_annual\_wages\_chg  
## 1 3.4 311188704  
## 2 2.4 0  
## 3 2.4 0  
## 4 3.6 0  
## 5 -1.7 0  
## 6 -2.2 0  
## oty\_taxable\_annual\_wages\_pct\_chg oty\_annual\_contributions\_chg  
## 1 2.3 -70421983  
## 2 0.0 0  
## 3 0.0 0  
## 4 0.0 0  
## 5 0.0 0  
## 6 0.0 0  
## oty\_annual\_contributions\_pct\_chg oty\_annual\_avg\_wkly\_wage\_chg  
## 1 -18.2 19  
## 2 0.0 64  
## 3 0.0 64  
## 4 0.0 49  
## 5 0.0 -25  
## 6 0.0 43  
## oty\_annual\_avg\_wkly\_wage\_pct\_chg oty\_avg\_annual\_pay\_chg  
## 1 2.3 1011  
## 2 4.5 3353  
## 3 4.5 3353  
## 4 4.2 2552  
## 5 -3.7 -1258  
## 6 2.9 2195  
## oty\_avg\_annual\_pay\_pct\_chg  
## 1 2.4  
## 2 4.5  
## 3 4.5  
## 4 4.2  
## 5 -3.6  
## 6 2.8

There are only some variables out of 38 variables that actually have an effect on the data. There are some which have no data at all. There are others whose numeric value has no meaning to the data. All such columns can be removed which would be useless. Hence after removing the unimportant data, I kept about 20 variables features that would be useful for the principal component analysis. I am using two methods: princomp() and principal(). Also the 'NA' values are removed before parsing the data. The included columns are: ("annual\_avg\_estabs","annual\_avg\_emplvl","total\_annual\_wages","annual\_avg\_wkly\_wage","avg\_annual\_pay","lq\_annual\_avg\_estabs","lq\_annual\_avg\_emplvl","lq\_total\_annual\_wages","lq\_annual\_avg\_wkly\_wage","lq\_avg\_annual\_pay","oty\_annual\_avg\_estabs\_chg","oty\_annual\_avg\_estabs\_pct\_chg","oty\_annual\_avg\_emplvl\_chg","oty\_annual\_avg\_emplvl\_pct\_chg","oty\_total\_annual\_wages\_chg","oty\_total\_annual\_wages\_pct\_chg","oty\_annual\_avg\_wkly\_wage\_chg","oty\_annual\_avg\_wkly\_wage\_pct\_chg","oty\_avg\_annual\_pay\_chg","oty\_avg\_annual\_pay\_pct\_chg")

keep <- c("annual\_avg\_estabs","annual\_avg\_emplvl","total\_annual\_wages","annual\_avg\_wkly\_wage","avg\_annual\_pay","lq\_annual\_avg\_estabs","lq\_annual\_avg\_emplvl","lq\_total\_annual\_wages","lq\_annual\_avg\_wkly\_wage","lq\_avg\_annual\_pay","oty\_annual\_avg\_estabs\_chg","oty\_annual\_avg\_estabs\_pct\_chg","oty\_annual\_avg\_emplvl\_chg","oty\_annual\_avg\_emplvl\_pct\_chg","oty\_total\_annual\_wages\_chg","oty\_total\_annual\_wages\_pct\_chg","oty\_annual\_avg\_wkly\_wage\_chg","oty\_annual\_avg\_wkly\_wage\_pct\_chg","oty\_avg\_annual\_pay\_chg","oty\_avg\_annual\_pay\_pct\_chg")  
  
BLS\_reduced = BLS[keep]  
head(BLS\_reduced)

## annual\_avg\_estabs annual\_avg\_emplvl total\_annual\_wages  
## 1 117452 1863561 80668352987  
## 2 1186 53491 4148191291  
## 3 1186 53491 4148191291  
## 4 587 11462 719550831  
## 5 2 13 430575  
## 6 17 147 11630538  
## annual\_avg\_wkly\_wage avg\_annual\_pay lq\_annual\_avg\_estabs  
## 1 832 43287 1.00  
## 2 1491 77550 1.55  
## 3 1491 77550 1.55  
## 4 1207 62776 1.55  
## 5 649 33771 1.24  
## 6 1527 79389 1.70  
## lq\_annual\_avg\_emplvl lq\_total\_annual\_wages lq\_annual\_avg\_wkly\_wage  
## 1 1.00 1.00 1.00  
## 2 1.44 1.74 1.21  
## 3 1.46 1.77 1.22  
## 4 1.26 1.60 1.27  
## 5 0.14 0.08 0.55  
## 6 0.79 0.69 0.88  
## lq\_avg\_annual\_pay oty\_annual\_avg\_estabs\_chg  
## 1 1.00 1394  
## 2 1.21 -10  
## 3 1.21 -10  
## 4 1.27 0  
## 5 0.55 0  
## 6 0.88 0  
## oty\_annual\_avg\_estabs\_pct\_chg oty\_annual\_avg\_emplvl\_chg  
## 1 1.2 18475  
## 2 -0.8 -1097  
## 3 -0.8 -1097  
## 4 0.0 -75  
## 5 0.0 0  
## 6 0.0 -7  
## oty\_annual\_avg\_emplvl\_pct\_chg oty\_total\_annual\_wages\_chg  
## 1 1.0 2665745136  
## 2 -2.0 97930469  
## 3 -2.0 97930469  
## 4 -0.7 24761729  
## 5 0.0 -7288  
## 6 -4.5 -263697  
## oty\_total\_annual\_wages\_pct\_chg oty\_annual\_avg\_wkly\_wage\_chg  
## 1 3.4 19  
## 2 2.4 64  
## 3 2.4 64  
## 4 3.6 49  
## 5 -1.7 -25  
## 6 -2.2 43  
## oty\_annual\_avg\_wkly\_wage\_pct\_chg oty\_avg\_annual\_pay\_chg  
## 1 2.3 1011  
## 2 4.5 3353  
## 3 4.5 3353  
## 4 4.2 2552  
## 5 -3.7 -1258  
## 6 2.9 2195  
## oty\_avg\_annual\_pay\_pct\_chg  
## 1 2.4  
## 2 4.5  
## 3 4.5  
## 4 4.2  
## 5 -3.6  
## 6 2.8

## PCA using princomp() and principal()

bls.fit.A <- princomp(formula = ~., data=BLS\_reduced, cor=TRUE, na.action=na.exclude)  
bls.fit.A

## Call:  
## princomp(formula = ~., data = BLS\_reduced, na.action = na.exclude,   
## cor = TRUE)  
##   
## Standard deviations:  
## Comp.1 Comp.2 Comp.3 Comp.4 Comp.5   
## 2.4133123121 1.9509518995 1.6825759361 1.3894174406 1.3356196928   
## Comp.6 Comp.7 Comp.8 Comp.9 Comp.10   
## 1.0547330642 0.9865858335 0.9386482159 0.6498237062 0.4017623660   
## Comp.11 Comp.12 Comp.13 Comp.14 Comp.15   
## 0.3489667405 0.3110178295 0.1659228595 0.1271441265 0.1039242789   
## Comp.16 Comp.17 Comp.18 Comp.19 Comp.20   
## 0.0267101313 0.0023142136 0.0020174100 0.0009982061 0.0001649254   
##   
## 20 variables and 3569127 observations.

library(psych)  
bls.fit.B <- principal(BLS\_reduced, nfactors=10, rotate="varimax")

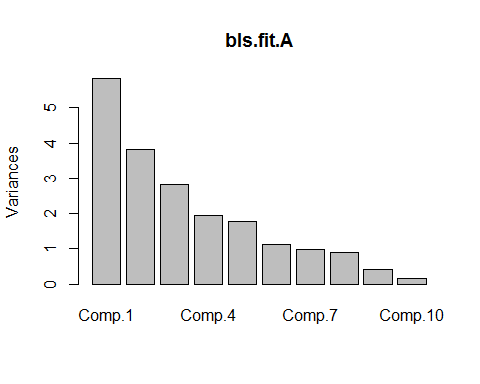
## Warning in pchisq(chi.sq.statistic, df, ncp = lam): pnchisq(x=2.58796e  
## +07, ..): not converged in 1000000 iter.  
  
## Warning in pchisq(chi.sq.statistic, df, ncp = lam): pnchisq(x=2.58796e  
## +07, ..): not converged in 1000000 iter.  
  
## Warning in pchisq(chi.sq.statistic, df, ncp = lam): pnchisq(x=2.58796e  
## +07, ..): not converged in 1000000 iter.  
  
## Warning in pchisq(chi.sq.statistic, df, ncp = lam): pnchisq(x=2.58796e  
## +07, ..): not converged in 1000000 iter.  
  
## Warning in pchisq(chi.sq.statistic, df, ncp = lam): pnchisq(x=2.58796e  
## +07, ..): not converged in 1000000 iter.  
  
## Warning in pchisq(chi.sq.statistic, df, ncp = lam): pnchisq(x=2.58796e  
## +07, ..): not converged in 1000000 iter.  
  
## Warning in pchisq(chi.sq.statistic, df, ncp = lam): pnchisq(x=2.58796e  
## +07, ..): not converged in 1000000 iter.  
  
## Warning in pchisq(chi.sq.statistic, df, ncp = lam): pnchisq(x=2.58796e  
## +07, ..): not converged in 1000000 iter.  
  
## Warning in pchisq(chi.sq.statistic, df, ncp = lam): pnchisq(x=2.58796e  
## +07, ..): not converged in 1000000 iter.  
  
## Warning in pchisq(chi.sq.statistic, df, ncp = lam): pnchisq(x=2.58796e  
## +07, ..): not converged in 1000000 iter.  
  
## Warning in pchisq(chi.sq.statistic, df, ncp = lam): pnchisq(x=2.58796e  
## +07, ..): not converged in 1000000 iter.  
  
## Warning in pchisq(chi.sq.statistic, df, ncp = lam): pnchisq(x=2.58796e  
## +07, ..): not converged in 1000000 iter.  
  
## Warning in pchisq(chi.sq.statistic, df, ncp = lam): pnchisq(x=2.58796e  
## +07, ..): not converged in 1000000 iter.  
  
## Warning in pchisq(chi.sq.statistic, df, ncp = lam): pnchisq(x=2.58796e  
## +07, ..): not converged in 1000000 iter.  
  
## Warning in pchisq(chi.sq.statistic, df, ncp = lam): pnchisq(x=2.58796e  
## +07, ..): not converged in 1000000 iter.  
  
## Warning in pchisq(chi.sq.statistic, df, ncp = lam): pnchisq(x=2.58796e  
## +07, ..): not converged in 1000000 iter.  
  
## Warning in pchisq(chi.sq.statistic, df, ncp = lam): pnchisq(x=2.58796e  
## +07, ..): not converged in 1000000 iter.  
  
## Warning in pchisq(chi.sq.statistic, df, ncp = lam): pnchisq(x=2.58796e  
## +07, ..): not converged in 1000000 iter.  
  
## Warning in pchisq(chi.sq.statistic, df, ncp = lam): pnchisq(x=2.58796e  
## +07, ..): not converged in 1000000 iter.  
  
## Warning in pchisq(chi.sq.statistic, df, ncp = lam): pnchisq(x=2.58796e  
## +07, ..): not converged in 1000000 iter.  
  
## Warning in pchisq(chi.sq.statistic, df, ncp = lam): pnchisq(x=2.58796e  
## +07, ..): not converged in 1000000 iter.  
  
## Warning in pchisq(chi.sq.statistic, df, ncp = lam): pnchisq(x=2.58796e  
## +07, ..): not converged in 1000000 iter.  
  
## Warning in pchisq(chi.sq.statistic, df, ncp = lam): pnchisq(x=2.58796e  
## +07, ..): not converged in 1000000 iter.  
  
## Warning in pchisq(chi.sq.statistic, df, ncp = lam): pnchisq(x=2.58796e  
## +07, ..): not converged in 1000000 iter.  
  
## Warning in pchisq(chi.sq.statistic, df, ncp = lam): pnchisq(x=2.58796e  
## +07, ..): not converged in 1000000 iter.  
  
## Warning in pchisq(chi.sq.statistic, df, ncp = lam): pnchisq(x=2.58796e  
## +07, ..): not converged in 1000000 iter.  
  
## Warning in pchisq(chi.sq.statistic, df, ncp = lam): pnchisq(x=2.58796e  
## +07, ..): not converged in 1000000 iter.  
  
## Warning in pchisq(chi.sq.statistic, df, ncp = lam): pnchisq(x=2.58796e  
## +07, ..): not converged in 1000000 iter.  
  
## Warning in pchisq(chi.sq.statistic, df, ncp = lam): pnchisq(x=2.58796e  
## +07, ..): not converged in 1000000 iter.  
  
## Warning in pchisq(chi.sq.statistic, df, ncp = lam): pnchisq(x=2.58796e  
## +07, ..): not converged in 1000000 iter.  
  
## Warning in pchisq(chi.sq.statistic, df, ncp = lam): pnchisq(x=2.58796e  
## +07, ..): not converged in 1000000 iter.  
  
## Warning in pchisq(chi.sq.statistic, df, ncp = lam): pnchisq(x=2.58796e  
## +07, ..): not converged in 1000000 iter.  
  
## Warning in pchisq(chi.sq.statistic, df, ncp = lam): pnchisq(x=2.58796e  
## +07, ..): not converged in 1000000 iter.

bls.fit.B

## Principal Components Analysis  
## Call: principal(r = BLS\_reduced, nfactors = 10, rotate = "varimax")  
## Standardized loadings (pattern matrix) based upon correlation matrix  
## RC1 RC2 RC3 RC5 RC4 RC8 RC6  
## annual\_avg\_estabs 0.99 0.00 0.00 0.00 0.00 0.00 0.00  
## annual\_avg\_emplvl 0.99 0.00 0.00 0.00 0.00 0.00 0.00  
## total\_annual\_wages 0.99 0.01 0.00 0.00 0.00 0.00 0.00  
## annual\_avg\_wkly\_wage 0.01 0.89 0.01 0.21 0.01 0.02 0.00  
## avg\_annual\_pay 0.01 0.89 0.01 0.21 0.01 0.02 0.00  
## lq\_annual\_avg\_estabs 0.00 -0.01 0.00 0.00 0.07 0.00 0.00  
## lq\_annual\_avg\_emplvl 0.00 0.02 0.00 0.00 0.97 0.00 0.00  
## lq\_total\_annual\_wages 0.00 0.02 0.00 0.00 0.97 0.00 0.00  
## lq\_annual\_avg\_wkly\_wage 0.01 0.97 0.01 0.02 0.02 0.02 -0.01  
## lq\_avg\_annual\_pay 0.01 0.97 0.01 0.02 0.02 0.02 -0.01  
## oty\_annual\_avg\_estabs\_chg 0.95 0.00 0.00 0.00 0.00 0.00 0.01  
## oty\_annual\_avg\_estabs\_pct\_chg 0.00 -0.01 0.00 0.00 0.00 0.06 1.00  
## oty\_annual\_avg\_emplvl\_chg 0.99 0.00 0.00 0.00 0.00 0.00 0.00  
## oty\_annual\_avg\_emplvl\_pct\_chg 0.00 0.05 0.02 -0.01 0.00 1.00 0.06  
## oty\_total\_annual\_wages\_chg 0.99 0.01 0.00 0.00 0.00 0.00 0.00  
## oty\_total\_annual\_wages\_pct\_chg 0.00 0.01 0.93 0.01 0.00 0.06 0.01  
## oty\_annual\_avg\_wkly\_wage\_chg 0.00 0.14 0.02 0.99 0.00 -0.01 0.00  
## oty\_annual\_avg\_wkly\_wage\_pct\_chg 0.00 0.01 0.99 0.02 0.00 -0.01 0.00  
## oty\_avg\_annual\_pay\_chg 0.00 0.14 0.02 0.99 0.00 -0.01 0.00  
## oty\_avg\_annual\_pay\_pct\_chg 0.00 0.01 0.99 0.02 0.00 -0.01 0.00  
## RC7 RC9 RC10 h2 u2 com  
## annual\_avg\_estabs 0.00 0.00 0.00 0.99 1.2e-02 1.0  
## annual\_avg\_emplvl 0.00 0.00 0.00 0.98 1.7e-02 1.0  
## total\_annual\_wages 0.00 0.00 0.00 0.98 1.8e-02 1.0  
## annual\_avg\_wkly\_wage 0.00 0.41 0.00 1.00 2.4e-06 1.5  
## avg\_annual\_pay 0.00 0.41 0.00 1.00 2.4e-06 1.5  
## lq\_annual\_avg\_estabs 1.00 0.00 0.00 1.00 4.2e-06 1.0  
## lq\_annual\_avg\_emplvl 0.03 0.00 0.00 0.95 4.8e-02 1.0  
## lq\_total\_annual\_wages 0.04 0.00 0.00 0.95 4.8e-02 1.0  
## lq\_annual\_avg\_wkly\_wage 0.00 -0.25 0.00 1.00 3.2e-06 1.1  
## lq\_avg\_annual\_pay 0.00 -0.25 0.00 1.00 3.2e-06 1.1  
## oty\_annual\_avg\_estabs\_chg 0.00 0.00 0.00 0.90 9.5e-02 1.0  
## oty\_annual\_avg\_estabs\_pct\_chg 0.00 0.00 0.00 1.00 6.4e-06 1.0  
## oty\_annual\_avg\_emplvl\_chg 0.00 0.00 0.00 0.98 1.9e-02 1.0  
## oty\_annual\_avg\_emplvl\_pct\_chg 0.00 0.00 0.00 1.00 9.1e-07 1.0  
## oty\_total\_annual\_wages\_chg 0.00 0.01 0.00 0.98 1.6e-02 1.0  
## oty\_total\_annual\_wages\_pct\_chg 0.00 0.00 0.37 1.00 1.9e-07 1.3  
## oty\_annual\_avg\_wkly\_wage\_chg 0.00 0.02 0.00 1.00 6.0e-07 1.0  
## oty\_annual\_avg\_wkly\_wage\_pct\_chg 0.00 0.00 -0.12 1.00 2.7e-06 1.0  
## oty\_avg\_annual\_pay\_chg 0.00 0.02 0.00 1.00 5.9e-07 1.0  
## oty\_avg\_annual\_pay\_pct\_chg 0.00 0.00 -0.12 1.00 2.7e-06 1.0  
##   
## RC1 RC2 RC3 RC5 RC4 RC8 RC6 RC7 RC9 RC10  
## SS loadings 5.82 3.49 2.83 2.05 1.91 1.00 1.00 1.00 0.46 0.17  
## Proportion Var 0.29 0.17 0.14 0.10 0.10 0.05 0.05 0.05 0.02 0.01  
## Cumulative Var 0.29 0.47 0.61 0.71 0.81 0.86 0.91 0.96 0.98 0.99  
## Proportion Explained 0.30 0.18 0.14 0.10 0.10 0.05 0.05 0.05 0.02 0.01  
## Cumulative Proportion 0.30 0.47 0.62 0.72 0.82 0.87 0.92 0.97 0.99 1.00  
##   
## Mean item complexity = 1.1  
## Test of the hypothesis that 10 components are sufficient.  
##   
## The root mean square of the residuals (RMSR) is 0.01   
## with the empirical chi square 36298.7 with prob < 0   
##   
## Fit based upon off diagonal values = 1

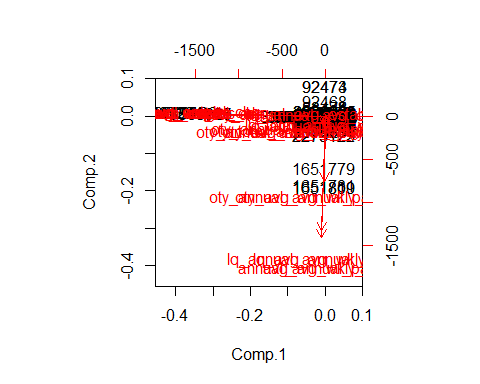
## screeplot()

screeplot(x=bls.fit.A)



## biplot()

biplot(bls.fit.A)



## Answers:

[A1] The proportion of the total variation in the data is explained by the summary:

summary(bls.fit.A)

## Importance of components:  
## Comp.1 Comp.2 Comp.3 Comp.4 Comp.5  
## Standard deviation 2.4133123 1.9509519 1.6825759 1.38941744 1.3356197  
## Proportion of Variance 0.2912038 0.1903107 0.1415531 0.09652404 0.0891940  
## Cumulative Proportion 0.2912038 0.4815145 0.6230676 0.71959161 0.8087856  
## Comp.6 Comp.7 Comp.8 Comp.9  
## Standard deviation 1.05473306 0.98658583 0.93864822 0.64982371  
## Proportion of Variance 0.05562309 0.04866758 0.04405302 0.02111354  
## Cumulative Proportion 0.86440870 0.91307628 0.95712931 0.97824285  
## Comp.10 Comp.11 Comp.12 Comp.13  
## Standard deviation 0.40176237 0.348966740 0.311017829 0.16592286  
## Proportion of Variance 0.00807065 0.006088889 0.004836605 0.00137652  
## Cumulative Proportion 0.98631350 0.992402387 0.997238992 0.99861551  
## Comp.14 Comp.15 Comp.16 Comp.17  
## Standard deviation 0.1271441265 0.1039242789 2.671013e-02 2.314214e-03  
## Proportion of Variance 0.0008082814 0.0005400128 3.567156e-05 2.677792e-07  
## Cumulative Proportion 0.9994237932 0.9999638060 9.999995e-01 9.999997e-01  
## Comp.18 Comp.19 Comp.20  
## Standard deviation 2.017410e-03 9.982061e-04 1.649254e-04  
## Proportion of Variance 2.034972e-07 4.982077e-08 1.360019e-09  
## Cumulative Proportion 9.999999e-01 1.000000e+00 1.000000e+00

[A2] The screeplot() shows the variances for each of the components. The component 1 has the highest variance which decreases with each component. Only 10 components are visible out of 20.

[A3] The first 8 components capture about 90% of the variances. Thus I would use the first 8 components. If more variances are to be captured say about 99% then the first 11 components are to be used.

[A4] Yes. There seems to be good amount of clustering as the data seems to be correlated. There is income information in most of the columns. Thus they are related to each other in the data. Hence the data would be similar for most of the columns.

A[5] The biplot for component 1 vs component 2 shows that the data is highly clustered. The cluster is nearer to the origin, showing that more values are either 0 or very close to zero.

A[6] 5 pcs are required to explain 75% variance of the data. The total variance given by the 6 pcs is about 80.8785%