Module 3

Neha Parulekar

February 8, 2016

## Additional packages needed

To run the code in M03\_Lesson\_02.Rmd you may need additional packages.

* If necessary install the following packages.
* install.packages("ggplot2");
* install.packages("reshape2");
* install.packages("psych");

require(ggplot2)

## Loading required package: ggplot2

require(reshape2)

## Loading required package: reshape2

require(psych)

## Loading required package: psych

##   
## Attaching package: 'psych'

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

## Assignment

* Download the compressed data from the U.S. [Bureau of Labor Statistics](https://en.wikipedia.org/wiki/Bureau_of_Labor_Statistics) [<http://www.bls.gov/>](http://www.bls.gov/) @ <http://www.bls.gov/cew/data/files/2014/csv/2014_annual_singlefile.zip>, and extract the .csv file.
* Run Principal Components Analysis on the BLS data and answer the following questions. (You can use any PCA function you wish, i.e. princomp(), prcomp(), principal() or by hand.)  
  **Reading the file**

# set the working directory  
setwd( "C:/Users/Neha/Desktop")  
# loading the data  
LaborStatistics <- read.csv("2014.annual.singlefile.csv")  
  
# Checking the data   
names(LaborStatistics)

## [1] "area\_fips" "own\_code"   
## [3] "industry\_code" "agglvl\_code"   
## [5] "size\_code" "year"   
## [7] "qtr" "disclosure\_code"   
## [9] "annual\_avg\_estabs" "annual\_avg\_emplvl"   
## [11] "total\_annual\_wages" "taxable\_annual\_wages"   
## [13] "annual\_contributions" "annual\_avg\_wkly\_wage"   
## [15] "avg\_annual\_pay" "lq\_disclosure\_code"   
## [17] "lq\_annual\_avg\_estabs" "lq\_annual\_avg\_emplvl"   
## [19] "lq\_total\_annual\_wages" "lq\_taxable\_annual\_wages"   
## [21] "lq\_annual\_contributions" "lq\_annual\_avg\_wkly\_wage"   
## [23] "lq\_avg\_annual\_pay" "oty\_disclosure\_code"   
## [25] "oty\_annual\_avg\_estabs\_chg" "oty\_annual\_avg\_estabs\_pct\_chg"   
## [27] "oty\_annual\_avg\_emplvl\_chg" "oty\_annual\_avg\_emplvl\_pct\_chg"   
## [29] "oty\_total\_annual\_wages\_chg" "oty\_total\_annual\_wages\_pct\_chg"   
## [31] "oty\_taxable\_annual\_wages\_chg" "oty\_taxable\_annual\_wages\_pct\_chg"  
## [33] "oty\_annual\_contributions\_chg" "oty\_annual\_contributions\_pct\_chg"  
## [35] "oty\_annual\_avg\_wkly\_wage\_chg" "oty\_annual\_avg\_wkly\_wage\_pct\_chg"  
## [37] "oty\_avg\_annual\_pay\_chg" "oty\_avg\_annual\_pay\_pct\_chg"

# retriving the dimensions of the data  
dim(LaborStatistics)

## [1] 3569127 38

# Removing the first eight columns from the data   
RemoveCols <- c("area\_fips","industry\_code","size\_code","qtr","own\_code","agglvl\_code","year","disclosure\_code","lq\_disclosure\_code","oty\_disclosure\_code")  
LaborStatistics <- LaborStatistics[,!(names(LaborStatistics) %in% RemoveCols)]  
  
# checking the dimensions of the data  
dim(LaborStatistics)

## [1] 3569127 28

* **Run Principal Components Analysis on the BLS data and answer the following questions. (You can use any PCA function you wish, i.e. princomp(), prcomp(), principal() or by hand.)**

LaborStatistic.fit.A <- princomp(formula = ~., data = LaborStatistics, cor = TRUE, na.action=na.exclude)  
print(LaborStatistic.fit.A)

## Call:  
## princomp(formula = ~., data = LaborStatistics, na.action = na.exclude,   
## cor = TRUE)  
##   
## Standard deviations:  
## Comp.1 Comp.2 Comp.3 Comp.4 Comp.5   
## 3.0871243914 1.9539344628 1.6826950969 1.5562368831 1.3363812066   
## Comp.6 Comp.7 Comp.8 Comp.9 Comp.10   
## 1.1830067372 1.1629781715 1.0401246533 0.9855732862 0.9342121508   
## Comp.11 Comp.12 Comp.13 Comp.14 Comp.15   
## 0.7988757665 0.6495354134 0.5003750633 0.4017578339 0.3875207081   
## Comp.16 Comp.17 Comp.18 Comp.19 Comp.20   
## 0.3630700797 0.3044679826 0.1757438733 0.1341101339 0.1268132207   
## Comp.21 Comp.22 Comp.23 Comp.24 Comp.25   
## 0.1142661371 0.0838781775 0.0480066349 0.0250238568 0.0023137519   
## Comp.26 Comp.27 Comp.28   
## 0.0020174013 0.0009981984 0.0001649247   
##   
## 28 variables and 3569127 observations.

LaborStatistic.fit.B <- prcomp(LaborStatistics,center = TRUE, scale.= TRUE)  
#print(LaborStatistic.fit.B)  
  
# getting the summary  
summary(LaborStatistic.fit.A)

## Importance of components:  
## Comp.1 Comp.2 Comp.3 Comp.4 Comp.5  
## Standard deviation 3.0871244 1.9539345 1.6826951 1.55623688 1.33638121  
## Proportion of Variance 0.3403692 0.1363521 0.1011237 0.08649547 0.06378267  
## Cumulative Proportion 0.3403692 0.4767213 0.5778450 0.66434046 0.72812313  
## Comp.6 Comp.7 Comp.8 Comp.9  
## Standard deviation 1.18300674 1.16297817 1.04012465 0.98557329  
## Proportion of Variance 0.04998232 0.04830422 0.03863783 0.03469124  
## Cumulative Proportion 0.77810545 0.82640967 0.86504750 0.89973874  
## Comp.10 Comp.11 Comp.12 Comp.13  
## Standard deviation 0.93421215 0.79887577 0.64953541 0.500375063  
## Proportion of Variance 0.03116973 0.02279295 0.01506772 0.008941972  
## Cumulative Proportion 0.93090847 0.95370142 0.96876914 0.977711111  
## Comp.14 Comp.15 Comp.16 Comp.17  
## Standard deviation 0.40175783 0.387520708 0.363070080 0.304467983  
## Proportion of Variance 0.00576462 0.005363296 0.004707853 0.003310741  
## Cumulative Proportion 0.98347573 0.988839027 0.993546880 0.996857621  
## Comp.18 Comp.19 Comp.20 Comp.21  
## Standard deviation 0.175743873 0.1341101339 0.1268132207 0.1142661371  
## Proportion of Variance 0.001103068 0.0006423403 0.0005743426 0.0004663125  
## Cumulative Proportion 0.997960689 0.9986030297 0.9991773723 0.9996436848  
## Comp.22 Comp.23 Comp.24 Comp.25  
## Standard deviation 0.0838781775 4.800663e-02 2.502386e-02 2.313752e-03  
## Proportion of Variance 0.0002512696 8.230846e-05 2.236405e-05 1.911946e-07  
## Cumulative Proportion 0.9998949544 9.999773e-01 9.999996e-01 9.999998e-01  
## Comp.26 Comp.27 Comp.28  
## Standard deviation 2.017401e-03 9.981984e-04 1.649247e-04  
## Proportion of Variance 1.453539e-07 3.558571e-08 9.714336e-10  
## Cumulative Proportion 1.000000e+00 1.000000e+00 1.000000e+00

names(LaborStatistic.fit.A)

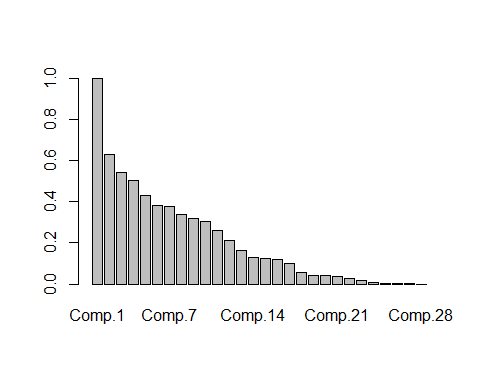
## [1] "sdev" "loadings" "center" "scale" "n.obs" "scores"   
## [7] "call"

summary(LaborStatistic.fit.B)

## Importance of components:  
## PC1 PC2 PC3 PC4 PC5 PC6 PC7  
## Standard deviation 3.0871 1.9539 1.6827 1.5562 1.33638 1.18301 1.1630  
## Proportion of Variance 0.3404 0.1363 0.1011 0.0865 0.06378 0.04998 0.0483  
## Cumulative Proportion 0.3404 0.4767 0.5778 0.6643 0.72812 0.77811 0.8264  
## PC8 PC9 PC10 PC11 PC12 PC13  
## Standard deviation 1.04012 0.98557 0.93421 0.79888 0.64954 0.50038  
## Proportion of Variance 0.03864 0.03469 0.03117 0.02279 0.01507 0.00894  
## Cumulative Proportion 0.86505 0.89974 0.93091 0.95370 0.96877 0.97771  
## PC14 PC15 PC16 PC17 PC18 PC19  
## Standard deviation 0.40176 0.38752 0.36307 0.30447 0.1757 0.13411  
## Proportion of Variance 0.00576 0.00536 0.00471 0.00331 0.0011 0.00064  
## Cumulative Proportion 0.98348 0.98884 0.99355 0.99686 0.9980 0.99860  
## PC20 PC21 PC22 PC23 PC24 PC25  
## Standard deviation 0.12681 0.11427 0.08388 0.04801 0.02502 0.002314  
## Proportion of Variance 0.00057 0.00047 0.00025 0.00008 0.00002 0.000000  
## Cumulative Proportion 0.99918 0.99964 0.99989 0.99998 1.00000 1.000000  
## PC26 PC27 PC28  
## Standard deviation 0.002017 0.0009981 0.000165  
## Proportion of Variance 0.000000 0.0000000 0.000000  
## Cumulative Proportion 1.000000 1.0000000 1.000000

* **Barplot**

#plotting the screeplot, barplot, biplot  
#screeplot(LaborStatistic.fit.A)  
barplot(LaborStatistic.fit.A$sdev/LaborStatistic.fit.A$sdev[1])



#biplot(LaborStatistic.fit.A, expand=100, xlim=c(-0.1, 0.1), ylim=c(-0.1, 0.1))

* **ggbiplot**

#loadint the required package and plotting ggbiplot  
  
library(devtools)  
install\_github("vqv/ggbiplot")

## Skipping install for github remote, the SHA1 (7325e880) has not changed since last install.  
## Use `force = TRUE` to force installation

library(ggbiplot)

## Loading required package: plyr

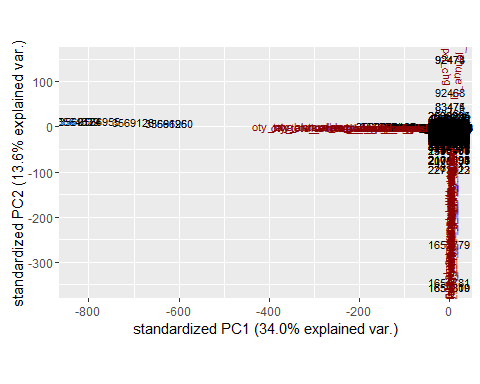
## Loading required package: scales

##   
## Attaching package: 'scales'

## The following objects are masked from 'package:psych':  
##   
## alpha, rescale

## Loading required package: grid

fit<- princomp(LaborStatistics, cor=TRUE)  
ggbiplot(fit, labels = rownames(LaborStatistics))



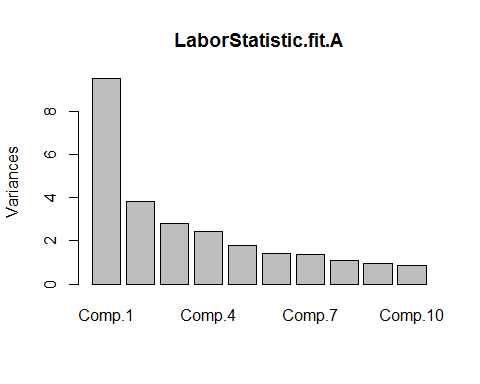
## **Answers**

**(a) What proportion of the total variation in the data is explained by each of the principal components?**

* PC1 PC2 PC3 PC4 PC5 PC6 PC7
* 0.3404 0.1363 0.1011 0.0865 0.06378 0.04998 0.0483
* PC8 PC9 PC10 PC11 PC12 PC13
* 0.03864 0.03469 0.03117 0.02279 0.01507 0.00894
* PC14 PC15 PC16 PC17 PC18 PC19
* 0.00576 0.00536 0.00471 0.00331 0.0011 0.00064
* PC20 PC21 PC22 PC23 PC24 PC25
* 0.00057 0.00047 0.00025 0.00008 0.00002 0.000000
* PC26 PC27 PC28
* 0.000000 0.0000000 0.000000

**(b) Plot a screeplot.**

screeplot(LaborStatistic.fit.A)

 The screeplot is a decreasing plot which shows variance of the first ten PCs.

**(c) Based on the variation explained for each of these components, which, if any, components would you use?** As we can see more than 90% of variance of the data was captured by first 10 PCs. So it would be sufficient to use those components.

**(d) Is there evidence of clustering in the data by creating biplots of the each of the components plotted against one another?** *It takes too long to run.* But running the **ggbiplot** we can see that there is clustering in data.

**(e) Do any of the biplots reveal any interesting structure?** *The biplot takes too long to run* Ruuning **ggbiplot** we can see the points in the labor statistics data would fall when we project from 28 space to 2 space.They are not evenly scattered about the projections.

**(f) How many pcs are required to explain 75% of the variance in the data?** Seven PCs are required to explain 75% of the variance in the data.