Presentation

Gordon Dri

December 4, 2016

Priniciple Component Analysis Tutorial:

- The data set shows Per Capita Income of 77 neighborhoods in Chicago
- There are 31 predictor varaibles available
- The predictor variables have been combined from two public datasets from the City of Chicago's Data Portal:

1. "Census Data - Selected socioeconomic indicators in Chicago, 2008 - 2012" (https://data.cityofchicago.org/Health-Human-Services/Census-Data-Selected-socioeconomic-indicators-in-C/kn9c-c2s2)
2. "Public Health Statistics- Selected public health indicators by Chicago community" (https://data.cityofchicago.org/Health-Human-Services/Public-Health-Statistics-Selected-public-health-in/iqnk-2tcu)

Part A: Load in Data and Data Preparation

```
# set the working directory to the file path in which the data sets are stored
datapath <- "/Users/gordondri/Downloads"

# dataset 1: public health indicators by Chicago community
public.health.indicators <- read.csv(file=paste(datapath,
"Public_health_indicators_by_Chicago_community_area.csv", sep="/"), header=TRUE, sep=",",
stringsAsFactors = FALSE)

# dataset 2: socioeconomic indicators by Chicago community
socioeconomic.indicators <- read.csv(file=paste(datapath,  "Socioeconomic_indicators_in_Chicago_2008_20
stringsAsFactors = FALSE)

# combine the two datasets, don't repeat redundant columns

# compare the number of rows in both datasets
nrow(public.health.indicators) > nrow(socioeconomic.indicators)
```

```
## [1] FALSE
```

```
# check the end of the socioeconomic indicators dataset
tail(socioeconomic.indicators)
```

```
##    Community.Area.Number COMMUNITY.AREA.NAME PERCENT.OF.HOUSING.CROWDED
## 73                    73   Washington Height                        1.1
## 74                    74     Mount Greenwood                        1.0
## 75                    75         Morgan Park                        0.8
## 76                    76              O'Hare                        3.6
## 77                    77           Edgewater                        4.1
## 78                    NA             CHICAGO                        4.7
##    PERCENT.HOUSEHOLDS.BELOW.POVERTY PERCENT.AGED.16..UNEMPLOYED
```

```
## 73                              16.9                         20.8
## 74                               3.4                          8.7
## 75                              13.2                         15.0
## 76                              15.4                          7.1
## 77                              18.2                          9.2
## 78                              19.7                         12.9
##    PERCENT.AGED.25..WITHOUT.HIGH.SCHOOL.DIPLOMA
## 73                                         13.7
## 74                                          4.3
## 75                                         10.8
## 76                                         10.9
## 77                                          9.7
## 78                                         19.5
##    PERCENT.AGED.UNDER.18.OR.OVER.64 PER.CAPITA.INCOME HARDSHIP.INDEX
## 73                             42.6             19713             48
## 74                             36.8             34381             16
## 75                             40.3             27149             30
## 76                             30.3             25828             24
## 77                             23.8             33385             19
## 78                             33.5             28202             NA
```

```
# remove the last row of this table to ensure both tables have equal number of rows
socioeconomic.indicators <- socioeconomic.indicators[1:77, ]

# check the first two column headers of both datasets for potential discrepancies
colnames(public.health.indicators)[1:2]
```

```
## [1] "Community.Area"      "Community.Area.Name"
```

```
colnames(socioeconomic.indicators)[1:2]
```

```
## [1] "Community.Area.Number" "COMMUNITY.AREA.NAME"
```

```
# make these headers the same in both datasets
colnames(socioeconomic.indicators)[1:2] <- colnames(public.health.indicators)[1:2]

# check the column headers
colnames(public.health.indicators)[1:2] ==
colnames(socioeconomic.indicators)[1:2]
```

```
## [1] TRUE TRUE
```

```
# do the same for per capita income
col.index.public <- which(colnames(public.health.indicators) == 'Per.Capita.Income')

col.index.socio <- which(colnames(socioeconomic.indicators) == 'PER.CAPITA.INCOME')

colnames(socioeconomic.indicators)[col.index.socio] <- colnames(public.health.indicators)[col.index.pub

colnames(public.health.indicators)
```

```
##  [1] "Community.Area"
##  [2] "Community.Area.Name"
##  [3] "Birth.Rate"
##  [4] "General.Fertility.Rate"
##  [5] "Low.Birth.Weight"
##  [6] "Prenatal.Care.Beginning.in.First.Trimester"
##  [7] "Preterm.Births"
##  [8] "Teen.Birth.Rate"
##  [9] "Assault..Homicide."
## [10] "Breast.cancer.in.females"
## [11] "Cancer..All.Sites."
## [12] "Colorectal.Cancer"
## [13] "Diabetes.related"
## [14] "Firearm.related"
## [15] "Infant.Mortality.Rate"
## [16] "Lung.Cancer"
## [17] "Prostate.Cancer.in.Males"
## [18] "Stroke..Cerebrovascular.Disease."
## [19] "Childhood.Blood.Lead.Level.Screening"
## [20] "Childhood.Lead.Poisoning"
## [21] "Gonorrhea.in.Females"
## [22] "Gonorrhea.in.Males"
## [23] "Tuberculosis"
## [24] "Below.Poverty.Level"
## [25] "Crowded.Housing"
## [26] "Dependency"
## [27] "No.High.School.Diploma"
## [28] "Per.Capita.Income"
## [29] "Unemployment"
```

```
colnames(socioeconomic.indicators)
```

```
## [1] "Community.Area"
## [2] "Community.Area.Name"
## [3] "PERCENT.OF.HOUSING.CROWDED"
## [4] "PERCENT.HOUSEHOLDS.BELOW.POVERTY"
## [5] "PERCENT.AGED.16..UNEMPLOYED"
## [6] "PERCENT.AGED.25..WITHOUT.HIGH.SCHOOL.DIPLOMA"
## [7] "PERCENT.AGED.UNDER.18.OR.OVER.64"
## [8] "Per.Capita.Income"
## [9] "HARDSHIP.INDEX"
```

```
# merge the two datasets
combined.data <- merge(public.health.indicators,
                       socioeconomic.indicators,
                       all = TRUE)

# check the combined dataset
colnames(combined.data)
```

```
##  [1] "Community.Area"
##  [2] "Community.Area.Name"
##  [3] "Per.Capita.Income"
```

```
##  [4] "Birth.Rate"
##  [5] "General.Fertility.Rate"
##  [6] "Low.Birth.Weight"
##  [7] "Prenatal.Care.Beginning.in.First.Trimester"
##  [8] "Preterm.Births"
##  [9] "Teen.Birth.Rate"
## [10] "Assault..Homicide."
## [11] "Breast.cancer.in.females"
## [12] "Cancer..All.Sites."
## [13] "Colorectal.Cancer"
## [14] "Diabetes.related"
## [15] "Firearm.related"
## [16] "Infant.Mortality.Rate"
## [17] "Lung.Cancer"
## [18] "Prostate.Cancer.in.Males"
## [19] "Stroke..Cerebrovascular.Disease."
## [20] "Childhood.Blood.Lead.Level.Screening"
## [21] "Childhood.Lead.Poisoning"
## [22] "Gonorrhea.in.Females"
## [23] "Gonorrhea.in.Males"
## [24] "Tuberculosis"
## [25] "Below.Poverty.Level"
## [26] "Crowded.Housing"
## [27] "Dependency"
## [28] "No.High.School.Diploma"
## [29] "Unemployment"
## [30] "PERCENT.OF.HOUSING.CROWDED"
## [31] "PERCENT.HOUSEHOLDS.BELOW.POVERTY"
## [32] "PERCENT.AGED.16..UNEMPLOYED"
## [33] "PERCENT.AGED.25..WITHOUT.HIGH.SCHOOL.DIPLOMA"
## [34] "PERCENT.AGED.UNDER.18.OR.OVER.64"
## [35] "HARDSHIP.INDEX"
```

```
head(combined.data)
```

```
##   Community.Area Community.Area.Name Per.Capita.Income Birth.Rate
## 1              1         Rogers Park             23714       16.4
## 2              1         Rogers Park             23939         NA
## 3              2          West Ridge             21375       17.3
## 4              2          West Ridge             23040         NA
## 5              3              Uptown             32355       13.1
## 6              3              Uptown             35787         NA
##   General.Fertility.Rate Low.Birth.Weight
## 1                   62.0             11.0
## 2                     NA               NA
## 3                   83.3              8.1
## 4                     NA               NA
## 5                   50.5              8.3
## 6                     NA               NA
##   Prenatal.Care.Beginning.in.First.Trimester Preterm.Births
## 1                                       73.0           11.2
## 2                                         NA             NA
## 3                                       71.1            8.3
## 4                                         NA             NA
```

```
## 5                                     77.7           10.3
## 6                                       NA             NA
##   Teen.Birth.Rate Assault..Homicide. Breast.cancer.in.females
## 1            40.8                7.7                     23.3
## 2              NA                 NA                       NA
## 3            29.9                5.8                     20.2
## 4              NA                 NA                       NA
## 5            35.1                5.4                     21.3
## 6              NA                 NA                       NA
##   Cancer..All.Sites. Colorectal.Cancer Diabetes.related Firearm.related
## 1              176.9              25.3             77.1             5.2
## 2                 NA                NA               NA              NA
## 3              155.9              17.3             60.5             3.7
## 4                 NA                NA               NA              NA
## 5              183.3              20.5             80.0             4.6
## 6                 NA                NA               NA              NA
##   Infant.Mortality.Rate Lung.Cancer Prostate.Cancer.in.Males
## 1                   6.4        36.7                     21.7
## 2                    NA          NA                       NA
## 3                   5.1        36.0                     14.2
## 4                    NA          NA                       NA
## 5                   6.5        50.5                     25.2
## 6                    NA          NA                       NA
##   Stroke..Cerebrovascular.Disease. Childhood.Blood.Lead.Level.Screening
## 1                             33.7                                364.7
## 2                               NA                                   NA
## 3                             34.7                                331.4
## 4                               NA                                   NA
## 5                             41.7                                353.7
## 6                               NA                                   NA
##   Childhood.Lead.Poisoning Gonorrhea.in.Females Gonorrhea.in.Males
## 1                      0.5                322.5              423.3
## 2                       NA                   NA               <NA>
## 3                      1.0                141.0              205.7
## 4                       NA                   NA               <NA>
## 5                      0.5                170.8              468.7
## 6                       NA                   NA               <NA>
##   Tuberculosis Below.Poverty.Level Crowded.Housing Dependency
## 1         11.4                22.7             7.9       28.8
## 2           NA                  NA              NA         NA
## 3          8.9                15.1             7.0       38.3
## 4           NA                  NA              NA         NA
## 5         13.6                22.7             4.6       22.2
## 6           NA                  NA              NA         NA
##   No.High.School.Diploma Unemployment PERCENT.OF.HOUSING.CROWDED
## 1                   18.1          7.5                         NA
## 2                     NA           NA                        7.7
## 3                   19.6          7.9                         NA
## 4                     NA           NA                        7.8
## 5                   13.6          7.7                         NA
## 6                     NA           NA                        3.8
##   PERCENT.HOUSEHOLDS.BELOW.POVERTY PERCENT.AGED.16..UNEMPLOYED
## 1                               NA                         NA
## 2                             23.6                        8.7
```

```
## 3                              NA                          NA
## 4                            17.2                         8.8
## 5                              NA                          NA
## 6                            24.0                         8.9
##   PERCENT.AGED.25..WITHOUT.HIGH.SCHOOL.DIPLOMA
## 1                                           NA
## 2                                         18.2
## 3                                           NA
## 4                                         20.8
## 5                                           NA
## 6                                         11.8
##   PERCENT.AGED.UNDER.18.OR.OVER.64 HARDSHIP.INDEX
## 1                               NA             NA
## 2                             27.5             39
## 3                               NA             NA
## 4                             38.5             46
## 5                               NA             NA
## 6                             22.2             20
```

```
# we see that the two datasets are using different per capita incomes for each Chicago community
cbind(Dataset1.Income.Per.Capita = public.health.indicators[1:10, 'Per.Capita.Income'],
Dataset2.Income.Per.Capita = socioeconomic.indicators[1:10, 'Per.Capita.Income'])
```

```
##       Dataset1.Income.Per.Capita Dataset2.Income.Per.Capita
##  [1,]                      23714                      23939
##  [2,]                      21375                      23040
##  [3,]                      32355                      35787
##  [4,]                      35503                      37524
##  [5,]                      51615                      57123
##  [6,]                      58227                      60058
##  [7,]                      71403                      71551
##  [8,]                      87163                      88669
##  [9,]                      38337                      40959
## [10,]                      31659                      32875
```

```
# merge the two datasets using only community area and community area name
combined.data <- merge(public.health.indicators,
                       socioeconomic.indicators,
                       by = c('Community.Area', 'Community.Area.Name'))
nrow(combined.data)
```

```
## [1] 75
```

```
nrow(public.health.indicators)
```

```
## [1] 77
```

```
nrow(socioeconomic.indicators)
```

```
## [1] 77
```

```r
# we see that the combined data set now has 75 rows instead of 77 which means that not all of the commu

# check the 1st column for differences
public.health.indicators[,1] == socioeconomic.indicators[,1]
```

```
##  [1] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## [15] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## [29] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## [43] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## [57] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## [71] TRUE TRUE TRUE TRUE TRUE TRUE TRUE
```

```r
# check the 2nd column for differences
public.health.indicators[,2] == socioeconomic.indicators[,2]
```

```
##  [1]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
## [12]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
## [23] FALSE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
## [34]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
## [45]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
## [56]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
## [67]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE  TRUE  TRUE
```

```r
equalities <- (public.health.indicators[,2] == socioeconomic.indicators[,2])

# analyze the community area names that are different in the two data sets
cbind(public.health.indicators[!equalities,2],
      socioeconomic.indicators[!equalities,2])
```

```
##      [,1]                 [,2]
## [1,] "Humboldt Park"      "Humboldt park"
## [2,] "Washington Heights" "Washington Height"
```

```r
# reconcile the differences
socioeconomic.indicators[!equalities,2] <- public.health.indicators[!equalities,2]

# check the 2nd column for differences
public.health.indicators[,2] == socioeconomic.indicators[,2]
```

```
##  [1] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## [15] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## [29] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## [43] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## [57] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## [71] TRUE TRUE TRUE TRUE TRUE TRUE TRUE
```

```r
# merge the two datasets using only community area and community area name
combined.data <- merge(public.health.indicators,
                       socioeconomic.indicators,
                       by = c('Community.Area', 'Community.Area.Name'))
colnames(combined.data)
```

```
##  [1] "Community.Area"
##  [2] "Community.Area.Name"
##  [3] "Birth.Rate"
##  [4] "General.Fertility.Rate"
##  [5] "Low.Birth.Weight"
##  [6] "Prenatal.Care.Beginning.in.First.Trimester"
##  [7] "Preterm.Births"
##  [8] "Teen.Birth.Rate"
##  [9] "Assault..Homicide."
## [10] "Breast.cancer.in.females"
## [11] "Cancer..All.Sites."
## [12] "Colorectal.Cancer"
## [13] "Diabetes.related"
## [14] "Firearm.related"
## [15] "Infant.Mortality.Rate"
## [16] "Lung.Cancer"
## [17] "Prostate.Cancer.in.Males"
## [18] "Stroke..Cerebrovascular.Disease."
## [19] "Childhood.Blood.Lead.Level.Screening"
## [20] "Childhood.Lead.Poisoning"
## [21] "Gonorrhea.in.Females"
## [22] "Gonorrhea.in.Males"
## [23] "Tuberculosis"
## [24] "Below.Poverty.Level"
## [25] "Crowded.Housing"
## [26] "Dependency"
## [27] "No.High.School.Diploma"
## [28] "Per.Capita.Income.x"
## [29] "Unemployment"
## [30] "PERCENT.OF.HOUSING.CROWDED"
## [31] "PERCENT.HOUSEHOLDS.BELOW.POVERTY"
## [32] "PERCENT.AGED.16..UNEMPLOYED"
## [33] "PERCENT.AGED.25..WITHOUT.HIGH.SCHOOL.DIPLOMA"
## [34] "PERCENT.AGED.UNDER.18.OR.OVER.64"
## [35] "Per.Capita.Income.y"
## [36] "HARDSHIP.INDEX"
```

```r
# delete one column of per capita income and keep the other
combined.data <- combined.data[,-which(colnames(combined.data) == 'Per.Capita.Income.y')]

colnames(combined.data)
```

```
##  [1] "Community.Area"
##  [2] "Community.Area.Name"
##  [3] "Birth.Rate"
##  [4] "General.Fertility.Rate"
##  [5] "Low.Birth.Weight"
##  [6] "Prenatal.Care.Beginning.in.First.Trimester"
##  [7] "Preterm.Births"
##  [8] "Teen.Birth.Rate"
##  [9] "Assault..Homicide."
## [10] "Breast.cancer.in.females"
## [11] "Cancer..All.Sites."
## [12] "Colorectal.Cancer"
```

```
## [13] "Diabetes.related"
## [14] "Firearm.related"
## [15] "Infant.Mortality.Rate"
## [16] "Lung.Cancer"
## [17] "Prostate.Cancer.in.Males"
## [18] "Stroke..Cerebrovascular.Disease."
## [19] "Childhood.Blood.Lead.Level.Screening"
## [20] "Childhood.Lead.Poisoning"
## [21] "Gonorrhea.in.Females"
## [22] "Gonorrhea.in.Males"
## [23] "Tuberculosis"
## [24] "Below.Poverty.Level"
## [25] "Crowded.Housing"
## [26] "Dependency"
## [27] "No.High.School.Diploma"
## [28] "Per.Capita.Income.x"
## [29] "Unemployment"
## [30] "PERCENT.OF.HOUSING.CROWDED"
## [31] "PERCENT.HOUSEHOLDS.BELOW.POVERTY"
## [32] "PERCENT.AGED.16..UNEMPLOYED"
## [33] "PERCENT.AGED.25..WITHOUT.HIGH.SCHOOL.DIPLOMA"
## [34] "PERCENT.AGED.UNDER.18.OR.OVER.64"
## [35] "HARDSHIP.INDEX"
```

```r
# look for any missing values in the data
row.has.na <- apply(combined.data, 1, function(x){any(is.na(x))})

# replace missing values in the data

# remove the 'Gonorrhea in Females' and 'Gonorrhea in Males' columns since they have many missing value
combined.data <- combined.data[,-c(which(colnames(combined.data) == 'Gonorrhea.in.Females'),which(colna

# look for any additional missing values in the data
row.has.na <- apply(combined.data, 1, function(x){any(is.na(x))})
combined.data[row.has.na, ]
```

```
##    Community.Area Community.Area.Name Birth.Rate General.Fertility.Rate
## 50             54           Riverdale       12.5                    46.1
##    Low.Birth.Weight Prenatal.Care.Beginning.in.First.Trimester
## 50             15.3                                      74.1
##    Preterm.Births Teen.Birth.Rate Assault..Homicide.
## 50           16.5            64.5                33
##    Breast.cancer.in.females Cancer..All.Sites. Colorectal.Cancer
## 50                       25              258.3              39.4
##    Diabetes.related Firearm.related Infant.Mortality.Rate Lung.Cancer
## 50            115.9            32.8                   8.7        86.1
##    Prostate.Cancer.in.Males Stroke..Cerebrovascular.Disease.
## 50                     42.5                              80.6
##    Childhood.Blood.Lead.Level.Screening Childhood.Lead.Poisoning
## 50                                   NA                      NA
##    Tuberculosis Below.Poverty.Level Crowded.Housing Dependency
## 50          5.8                61.4             5.1       50.2
##    No.High.School.Diploma Per.Capita.Income.x Unemployment
## 50                   24.6                8535         26.4
```

```
##    PERCENT.OF.HOUSING.CROWDED PERCENT.HOUSEHOLDS.BELOW.POVERTY
## 50                        5.8                               56.5
##    PERCENT.AGED.16..UNEMPLOYED
## 50                        34.6
##    PERCENT.AGED.25..WITHOUT.HIGH.SCHOOL.DIPLOMA
## 50                                         27.5
##    PERCENT.AGED.UNDER.18.OR.OVER.64 HARDSHIP.INDEX
## 50                             51.5             98
```

```
# replace the remaining missing values with column means

# find the index(es) at which there are NA values
indx <- which(is.na(combined.data), arr.ind=TRUE); indx
```

```
##      row col
## [1,]  50  19
## [2,]  50  20
```

```
# change NA values to 0
combined.data[is.na(combined.data)] <- 0
# change the columns to numeric
combined.data[3:ncol(combined.data)] <- sapply(combined.data[,3:ncol(combined.data)], as.numeric)
# find columns means
cM <- colMeans(combined.data[3:ncol(combined.data)], na.rm=TRUE)
# replace the index(es) with their column mean(s). Subtract 2 to the columns from indx since we only ca
combined.data[indx] <- cM[(indx[,2] - 2)]

# define the 31 predictors. Remove the first two columns as they pertain to the community names and rem
dataPredictors <- combined.data[, -c(1, 2, 26)]

# standardize every column of the predictor data since some are given as percents and some are given as
dataPredictors <- apply(dataPredictors, 2, scale)

# define new data set with normalized columns
combined.data <- cbind(combined.data[,1:2], Per.Capita.Income = combined.data[,26], dataPredictors)
```

Part B: View the data

```
summary(combined.data)
```

```
##  Community.Area Community.Area.Name Per.Capita.Income   Birth.Rate
##  Min.   : 1     Length:77           Min.   : 8535     Min.   :-1.784974
##  1st Qu.:20     Class :character    1st Qu.:15467     1st Qu.:-0.793117
##  Median :39     Mode  :character    Median :20489     Median : 0.000368
##  Mean   :39                         Mean   :25107     Mean   : 0.000000
##  3rd Qu.:58                         3rd Qu.:29026     3rd Qu.: 0.793853
##  Max.   :77                         Max.   :87163     Max.   : 1.899065
##  General.Fertility.Rate Low.Birth.Weight
##  Min.   :-2.667206      Min.   :-1.6866
##  1st Qu.:-0.543655      1st Qu.:-0.7157
##  Median :-0.006214      Median :-0.3580
##  Mean   : 0.000000      Mean   : 0.0000
```

```
##  3rd Qu.: 0.806503      3rd Qu.: 0.6640
##  Max.   : 1.737195      Max.   : 2.4524
##  Prenatal.Care.Beginning.in.First.Trimester Preterm.Births
##  Min.   :-2.5458                           Min.   :-2.0766
##  1st Qu.:-0.6967                           1st Qu.:-0.8170
##  Median :-0.1629                           Median :-0.1541
##  Mean   : 0.0000                           Mean   : 0.0000
##  3rd Qu.: 0.6568                           3rd Qu.: 0.8071
##  Max.   : 3.3447                           Max.   : 2.0667
##  Teen.Birth.Rate    Assault..Homicide. Breast.cancer.in.females
##  Min.   :-1.73554   Min.   :-1.0910    Min.   :-1.9201
##  1st Qu.:-0.58243   1st Qu.:-0.7952    1st Qu.:-0.6018
##  Median :-0.03078   Median :-0.4389    Median :-0.2042
##  Mean   : 0.00000   Mean   : 0.0000    Mean   : 0.0000
##  3rd Qu.: 0.63475   3rd Qu.: 0.8533    3rd Qu.: 0.7060
##  Max.   : 2.37866   Max.   : 3.1539    Max.   : 3.0079
##  Cancer..All.Sites. Colorectal.Cancer  Diabetes.related
##  Min.   :-1.6249    Min.   :-1.68731   Min.   :-2.09962
##  1st Qu.:-0.9064    1st Qu.:-0.84264   1st Qu.:-0.62041
##  Median :-0.1069    Median : 0.04101   Median : 0.04942
##  Mean   : 0.0000    Mean   : 0.00000   Mean   : 0.00000
##  3rd Qu.: 0.8964    3rd Qu.: 0.71674   3rd Qu.: 0.55644
##  Max.   : 2.1296    Max.   : 2.31512   Max.   : 2.19380
##  Firearm.related   Infant.Mortality.Rate  Lung.Cancer
##  Min.   :-1.0877   Min.   :-1.6045       Min.   :-2.16435
##  1st Qu.:-0.7766   1st Qu.:-0.7880       1st Qu.:-0.83831
##  Median :-0.4239   Median :-0.1075       Median :-0.09014
##  Mean   : 0.0000   Mean   : 0.0000       Mean   : 0.00000
##  3rd Qu.: 0.6271   3rd Qu.: 0.6410       3rd Qu.: 0.74928
##  Max.   : 3.7041   Max.   : 3.1814       Max.   : 2.31863
##  Prostate.Cancer.in.Males Stroke..Cerebrovascular.Disease.
##  Min.   :-1.7871          Min.   :-1.6987
##  1st Qu.:-0.8305          1st Qu.:-0.5983
##  Median :-0.2285          Median :-0.2384
##  Mean   : 0.0000          Mean   : 0.0000
##  3rd Qu.: 0.7087          3rd Qu.: 0.4814
##  Max.   : 2.7237          Max.   : 3.6374
##  Childhood.Blood.Lead.Level.Screening Childhood.Lead.Poisoning
##  Min.   :-2.113128                    Min.   :-1.1241
##  1st Qu.:-0.621685                    1st Qu.:-0.7333
##  Median :-0.002769                    Median :-0.2123
##  Mean   : 0.000000                    Mean   : 0.0000
##  3rd Qu.: 0.814334                    3rd Qu.: 0.3087
##  Max.   : 1.853137                    Max.   : 3.6952
##   Tuberculosis      Below.Poverty.Level Crowded.Housing
##  Min.   :-1.49177   Min.   :-1.4954     Min.   :-1.2886
##  1st Qu.:-0.83788   1st Qu.:-0.7212     1st Qu.:-0.7965
##  Median :-0.07501   Median :-0.1820     Median :-0.1949
##  Mean   : 0.00000   Mean   : 0.0000     Mean   : 0.0000
##  3rd Qu.: 0.55708   3rd Qu.: 0.5052     3rd Qu.: 0.5160
##  Max.   : 3.45597   Max.   : 3.5755     Max.   : 3.4689
##    Dependency      No.High.School.Diploma  Unemployment
##  Min.   :-2.7965   Min.   :-1.5132        Min.   :-1.2946
##  1st Qu.:-0.4856   1st Qu.:-0.6634        1st Qu.:-0.7827
```

```
##  Median : 0.3398   Median :-0.2506        Median :-0.2565
##  Mean   : 0.0000   Mean   : 0.0000        Mean   : 0.0000
##  3rd Qu.: 0.6974   3rd Qu.: 0.6316        3rd Qu.: 0.5825
##  Max.   : 1.9767   Max.   : 3.0031        Max.   : 3.7964
##  PERCENT.OF.HOUSING.CROWDED PERCENT.HOUSEHOLDS.BELOW.POVERTY
##  Min.   :-1.2554            Min.   :-1.6016
##  1st Qu.:-0.7123            1st Qu.:-0.7430
##  Median :-0.3050            Median :-0.2486
##  Mean   : 0.0000            Mean   : 0.0000
##  3rd Qu.: 0.5096            3rd Qu.: 0.6447
##  Max.   : 2.9533            Max.   : 3.0125
##  PERCENT.AGED.16..UNEMPLOYED PERCENT.AGED.25..WITHOUT.HIGH.SCHOOL.DIPLOMA
##  Min.   :-1.4148             Min.   :-1.5090
##  1st Qu.:-0.8183             1st Qu.:-0.7224
##  Median :-0.1952             Median :-0.1558
##  Mean   : 0.0000             Mean   : 0.0000
##  3rd Qu.: 0.6134             3rd Qu.: 0.5293
##  Max.   : 2.7212             Max.   : 2.9145
##  PERCENT.AGED.UNDER.18.OR.OVER.64 HARDSHIP.INDEX
##  Min.   :-3.0360                  Min.   :-1.6907
##  1st Qu.:-0.5113                  1st Qu.:-0.8542
##  Median : 0.3211                  Median : 0.0172
##  Mean   : 0.0000                  Mean   : 0.0000
##  3rd Qu.: 0.6487                  3rd Qu.: 0.8537
##  Max.   : 2.1498                  Max.   : 1.6902
```

Part C: Linear Model Analysis

```r
# fit the data to a linear model
linMod <- lm(Per.Capita.Income~.,data=combined.data[,3:ncol(combined.data)])
summary(linMod)
```

```
##
## Call:
## lm(formula = Per.Capita.Income ~ ., data = combined.data[, 3:ncol(combined.data)])
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10929.6  -2409.1   -685.4   2266.3  23615.1
##
## Coefficients:
##                                              Estimate Std. Error t value
## (Intercept)                                   25106.7      710.4  35.343
## Birth.Rate                                     5203.8     2657.3   1.958
## General.Fertility.Rate                        -6450.7     3778.2  -1.707
## Low.Birth.Weight                              -2672.2     2092.3  -1.277
## Prenatal.Care.Beginning.in.First.Trimester    -4008.1     1240.3  -3.232
## Preterm.Births                                 4276.8     1868.7   2.289
## Teen.Birth.Rate                                 127.2     2591.1   0.049
## Assault..Homicide.                            -1103.4     5489.5  -0.201
## Breast.cancer.in.females                       1227.6     1341.4   0.915
## Cancer..All.Sites.                            -1224.5     4406.0  -0.278
## Colorectal.Cancer                               435.8     1495.4   0.291
```

```
## Diabetes.related                                -5909.0     1555.8  -3.798
## Firearm.related                                  -948.3     4878.0  -0.194
## Infant.Mortality.Rate                           -1085.7     1816.7  -0.598
## Lung.Cancer                                     -1399.8     3066.2  -0.457
## Prostate.Cancer.in.Males                         2038.0     1666.5   1.223
## Stroke..Cerebrovascular.Disease.                 3182.5     1750.4   1.818
## Childhood.Blood.Lead.Level.Screening             -759.9     1898.8  -0.400
## Childhood.Lead.Poisoning                         -949.4     1375.2  -0.690
## Tuberculosis                                    -3377.8     1075.1  -3.142
## Below.Poverty.Level                              2923.8     6532.0   0.448
## Crowded.Housing                                  -411.2     4136.7  -0.099
## Dependency                                       -500.3     5275.8  -0.095
## No.High.School.Diploma                          12807.4     7238.8   1.769
## Unemployment                                    -3576.9     4441.1  -0.805
## PERCENT.OF.HOUSING.CROWDED                       -338.2     3272.7  -0.103
## PERCENT.HOUSEHOLDS.BELOW.POVERTY                -2231.4     8046.9  -0.277
## PERCENT.AGED.16..UNEMPLOYED                      2000.1     4389.9   0.456
## PERCENT.AGED.25..WITHOUT.HIGH.SCHOOL.DIPLOMA   -13337.8     7984.8  -1.670
## PERCENT.AGED.UNDER.18.OR.OVER.64                -3060.9     5438.0  -0.563
## HARDSHIP.INDEX                                  -3551.0     7418.9  -0.479
##                                                Pr(>|t|)
## (Intercept)                                     < 2e-16 ***
## Birth.Rate                                     0.056275 .
## General.Fertility.Rate                         0.094497 .
## Low.Birth.Weight                               0.207953
## Prenatal.Care.Beginning.in.First.Trimester     0.002278 **
## Preterm.Births                                 0.026738 *
## Teen.Birth.Rate                                0.961064
## Assault..Homicide.                             0.841577
## Breast.cancer.in.females                       0.364878
## Cancer..All.Sites.                             0.782329
## Colorectal.Cancer                              0.772050
## Diabetes.related                               0.000426 ***
## Firearm.related                                0.846716
## Infant.Mortality.Rate                          0.553011
## Lung.Cancer                                    0.650175
## Prostate.Cancer.in.Males                       0.227608
## Stroke..Cerebrovascular.Disease.               0.075557 .
## Childhood.Blood.Lead.Level.Screening           0.690847
## Childhood.Lead.Poisoning                       0.493433
## Tuberculosis                                   0.002935 **
## Below.Poverty.Level                            0.656531
## Crowded.Housing                                0.921257
## Dependency                                     0.924865
## No.High.School.Diploma                         0.083476 .
## Unemployment                                   0.424734
## PERCENT.OF.HOUSING.CROWDED                     0.918143
## PERCENT.HOUSEHOLDS.BELOW.POVERTY               0.782795
## PERCENT.AGED.16..UNEMPLOYED                    0.650806
## PERCENT.AGED.25..WITHOUT.HIGH.SCHOOL.DIPLOMA   0.101631
## PERCENT.AGED.UNDER.18.OR.OVER.64               0.576259
## HARDSHIP.INDEX                                 0.634457
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 6234 on 46 degrees of freedom
## Multiple R-squared:  0.8948, Adjusted R-squared:  0.8262
## F-statistic: 13.04 on 30 and 46 DF,  p-value: 4.426e-14
```

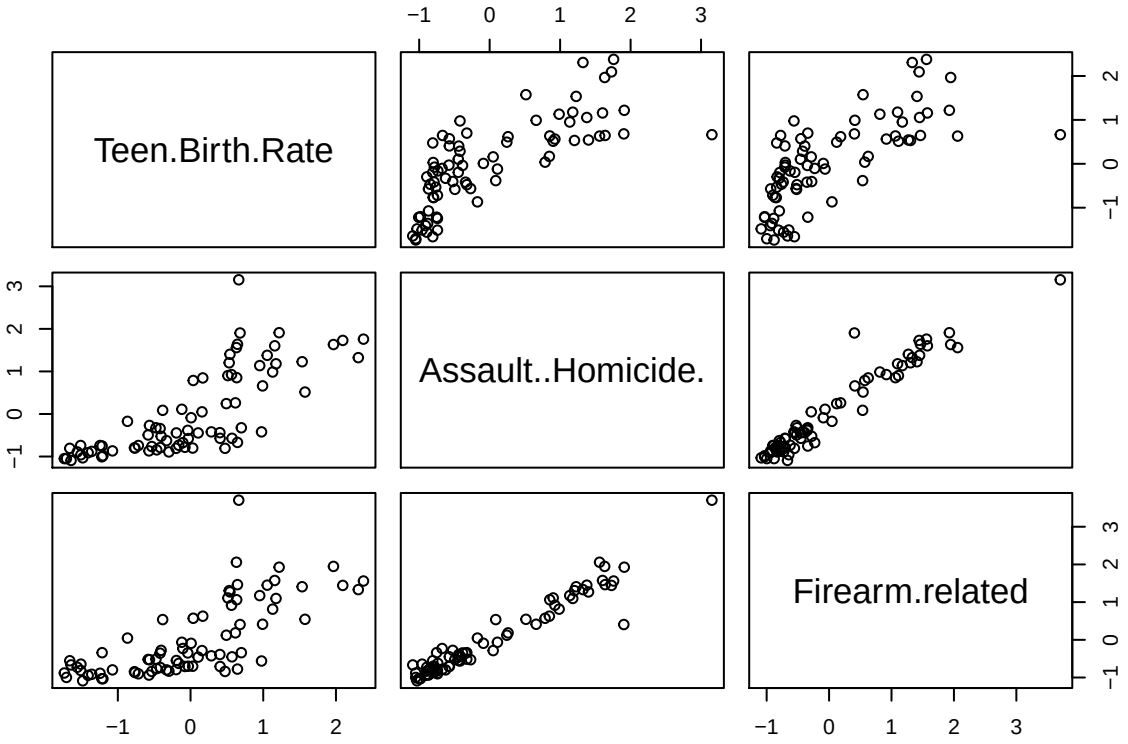
Observations from the Linear Model Summary: - 8 non-intercept parameters significant at alpha level of 0.1 or less - Multiple R-squared value of 0.8948 is higher than the adjusted R-squared value of 0.8262 which suggests we have unnecessary predictors in our model - The F-statistic of 13.04 is significant with a very low p-value, meaning that at least one non-intercept parameter is non-zero (we must reject our null hypothesis)

Part D: Selection of Predictors based on PCA

```
# separate output from inputs
Data.Output<-combined.data$Per.Capita.Income
Data.Input<-data.matrix(dataPredictors,rownames.force="automatic")
dim(Data.Input)
```

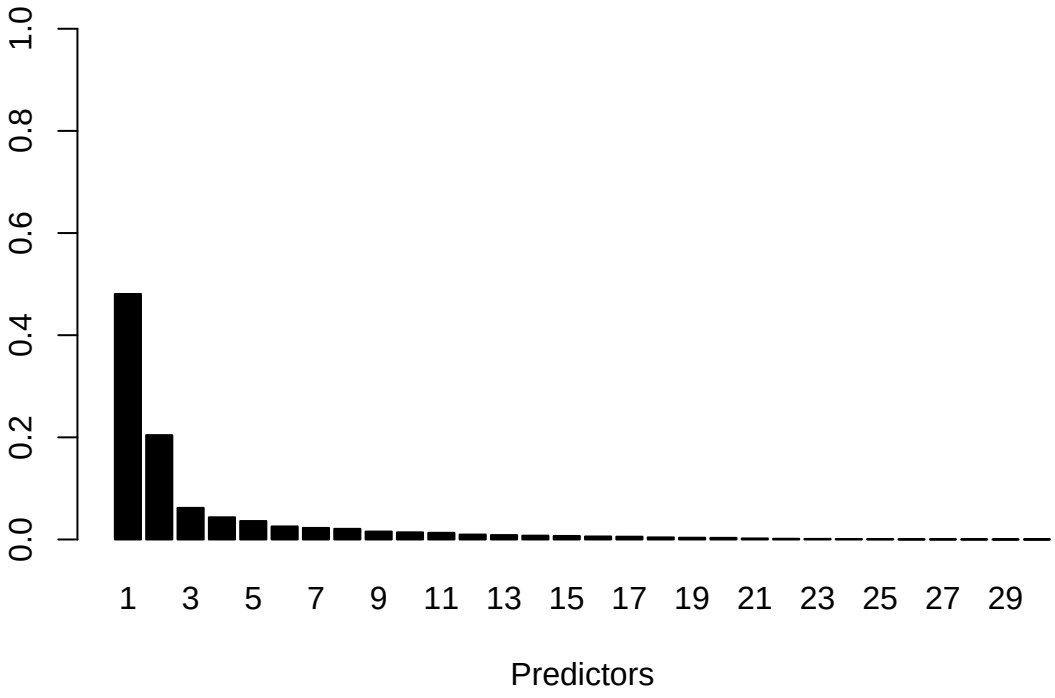
```
## [1] 77 30
```

```
# explore the dimensionality of a set of three input variables
Combined.data.1.2.3<-Data.Input[,c(6,7,12)]
pairs(Combined.data.1.2.3)
```

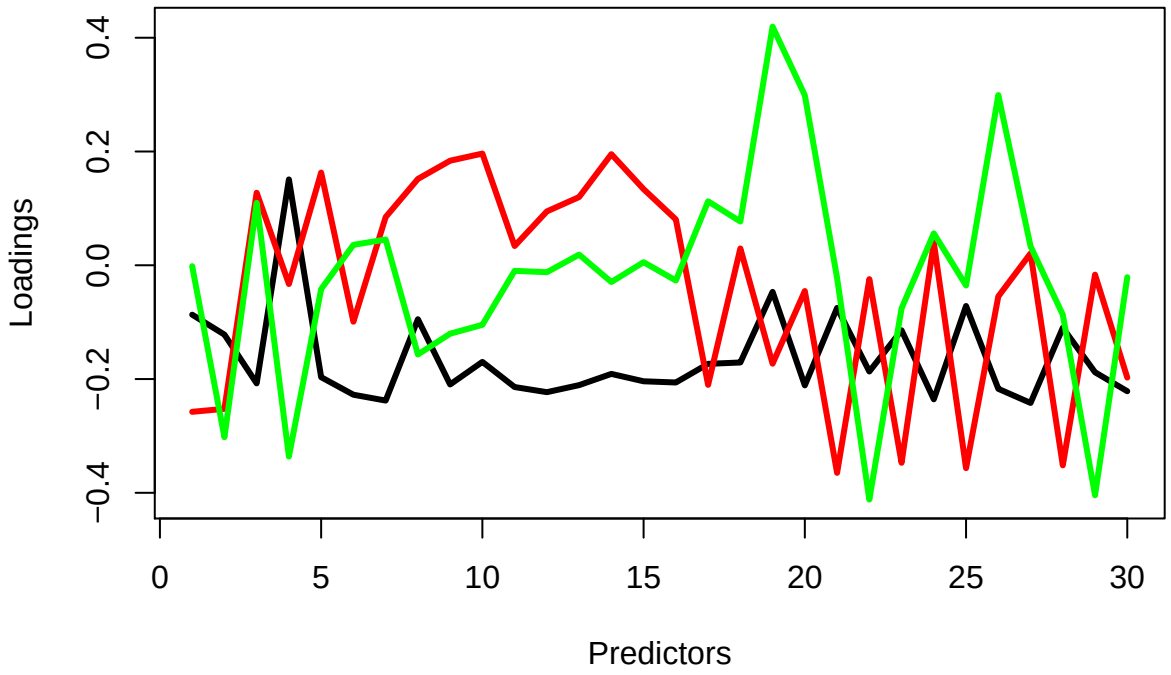



```
# we can see that the columns of Assualt.Homocide and Firearm.Related are highly correlated

# perform PCA by manually calculating factors, loadings and analyzing the importance of factors

# calculate 3 factor loadings using manual method based on eigen-decomposition
```

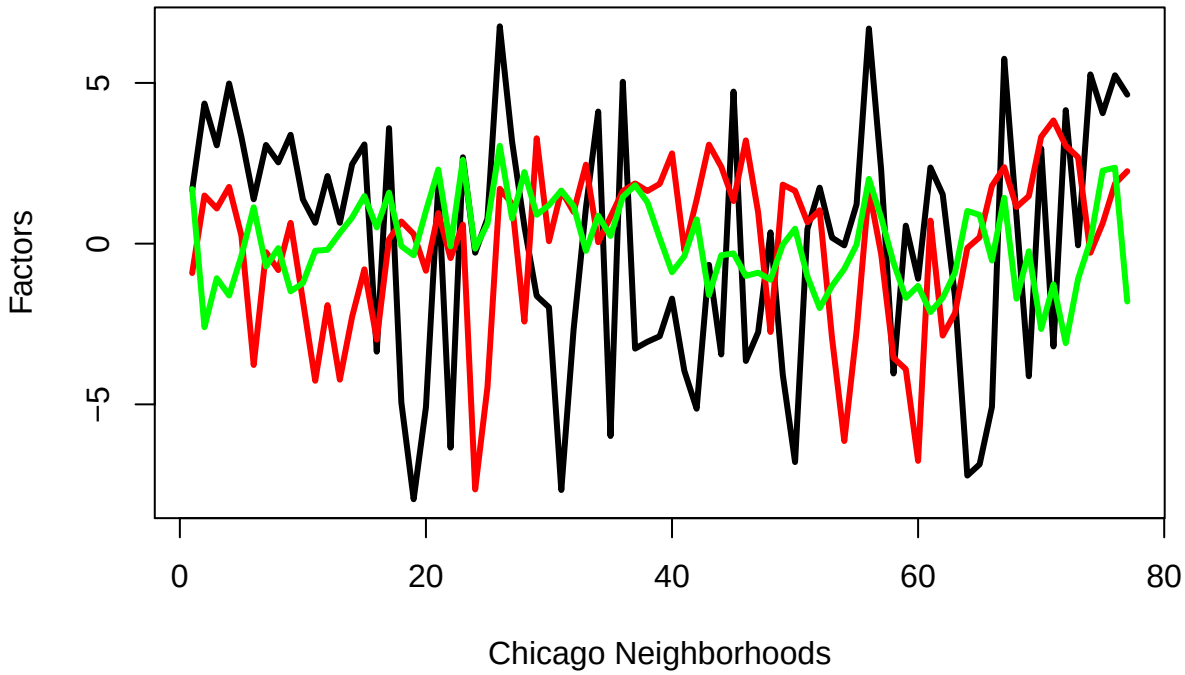
```r
# STEP 1: create a centered matrix
centered.matrix <- dataPredictors
col.means <- colMeans(dataPredictors)
for (i in 1:ncol(dataPredictors)){
  centered.matrix[,i] <- dataPredictors[,i] - col.means[i]
}

# STEP 2: calculate the covariance matrix
cov.matrix <- cov(centered.matrix)

# STEP 3: perform eigenvalue decomposition of the covariance matrix
eigens <- eigen(cov.matrix)
eigen.vecs <- eigens$vectors
eigen.vals <- eigens$values

# plot the normalized eigen values
barplot(eigen.vals/sum(eigen.vals),width=2,col = "black", ylim = c(0, 1), names.arg = rep(1:30), xlab='
```



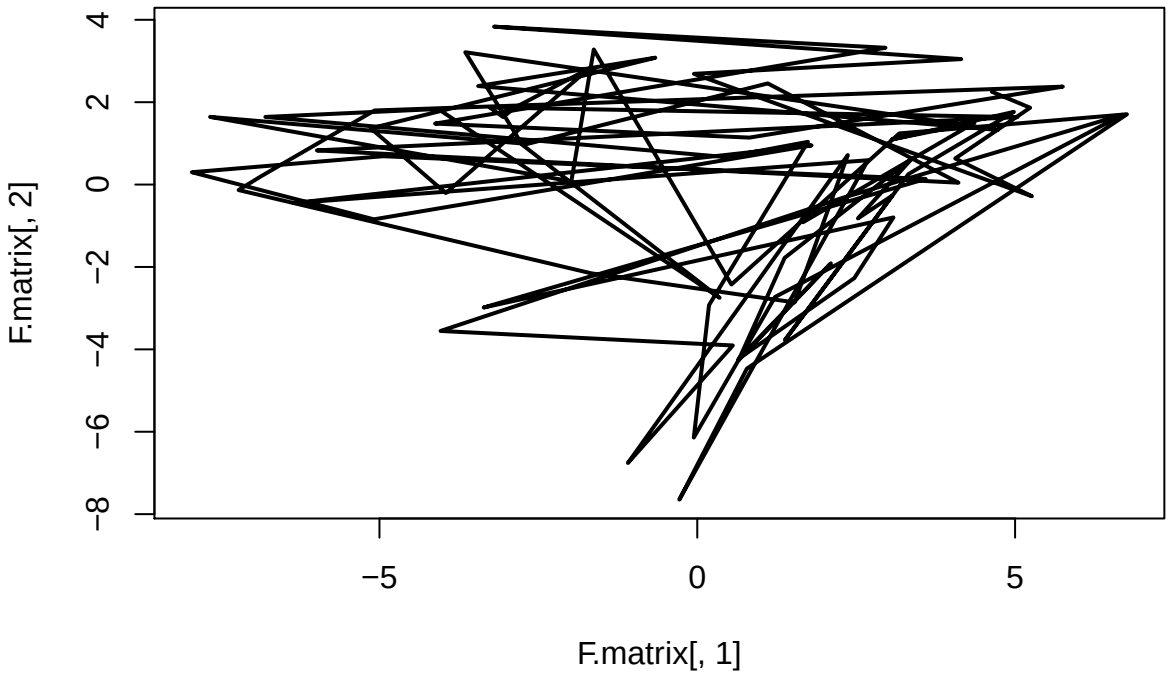

```r
# Define the L matrix as having columns corresponding to the eigenvectors with the 3 largest eigenvalue
L.matrix <- eigen.vecs

# plot the loadings
matplot(L.matrix[,1:3],type="l",lty=1,col=c("black","red","green"),lwd=3,xlab='Predictors',ylab='Loadin
```



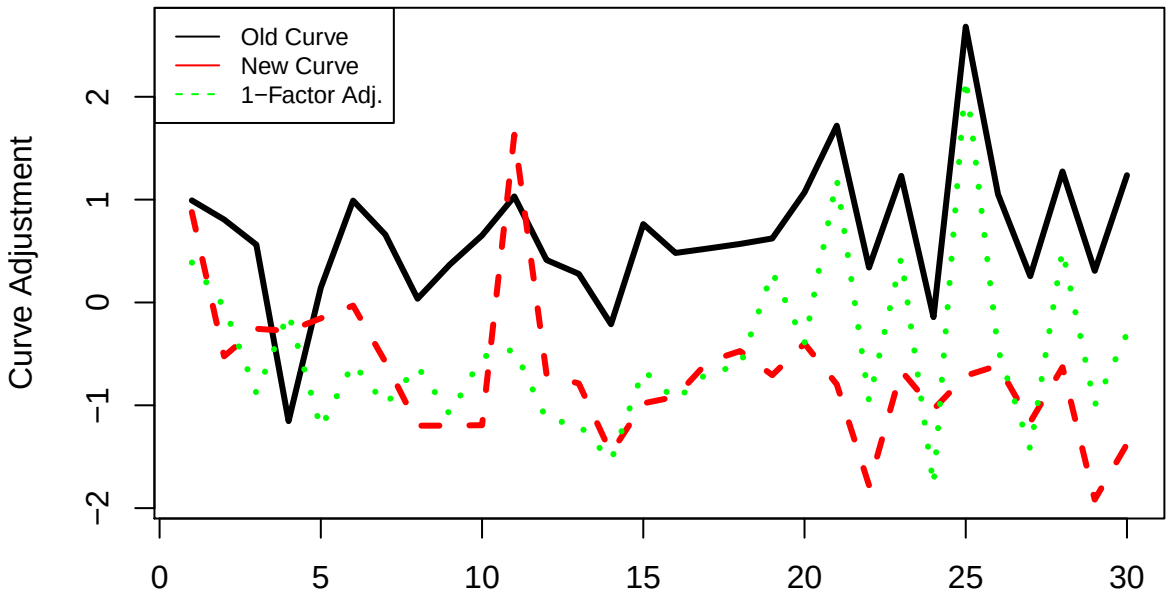

```
# Define the F matrix by multiplying the centered matrix by the L matrix
F.matrix <- as.matrix(centered.matrix) %*% L.matrix

# calculate and plot 3 selected factors
matplot(F.matrix[,1:3],type="l",col=c("black","red","green"),lty=1,lwd=3, xlab='Chicago Neighborhoods',
```



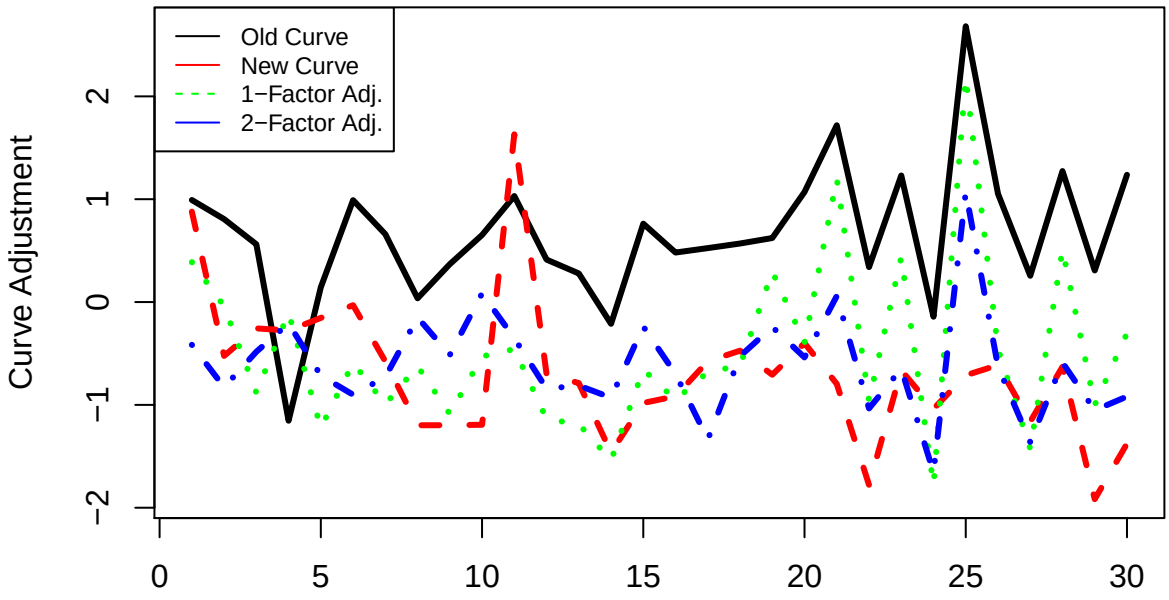

```
# compare factors
plot(F.matrix[,1],F.matrix[,2],type="l",lwd=2)
```



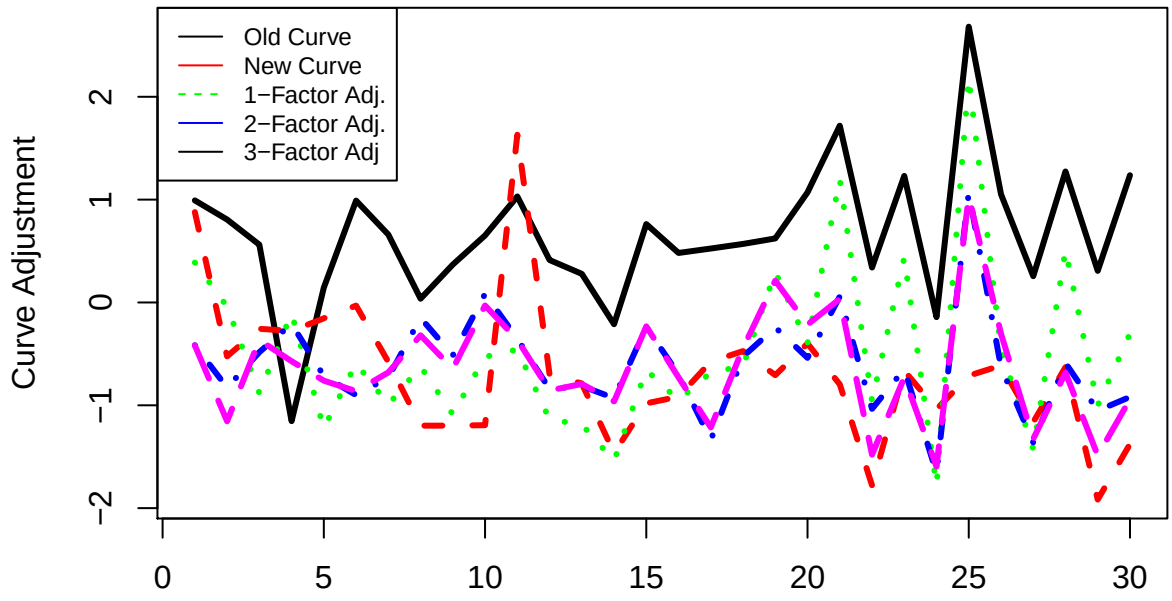

```
# analyze the adjustments that each factor makes to the curve (output variable).
# each of the factors makes an adjustment corresponding to the shape of its loading
# look at the shapes of the loadings and tell what mode of curve move corresponds to each factor
OldCurve<-Data.Input[16,]
NewCurve<-Data.Input[17,]
CurveChange<-NewCurve-OldCurve
FactorsChange<-F.matrix[17,]-F.matrix[16,]
ModelCurveAdjustment.1Factor<-OldCurve+t(L.matrix[,1])*FactorsChange[1]
ModelCurveAdjustment.2Factors<-OldCurve+t(L.matrix[,1])*FactorsChange[1]+t(L.matrix[,2])*FactorsChange[
ModelCurveAdjustment.3Factors<-OldCurve+t(L.matrix[,1])*FactorsChange[1]+t(L.matrix[,2])*FactorsChange[
  t(L.matrix[,3])*FactorsChange[3]

# 1 factor adjustment
matplot(t(rbind(OldCurve,NewCurve,ModelCurveAdjustment.1Factor)),type="l",col=c("black","red","green"),
legend(x="topleft",c("Old Curve","New Curve","1-Factor Adj."),lty=c(1,1,2),lwd=1,col=c("black","red","g
```




```
# 2 factor adjustment
matplot(t(rbind(OldCurve,NewCurve,ModelCurveAdjustment.1Factor,ModelCurveAdjustment.2Factors)),type="l"
legend(x="topleft",c("Old Curve","New Curve","1-Factor Adj.","2-Factor Adj."),lty=c(1,1,2),lwd=1,col=c(
```




```
# 3 factor adjustment
matplot(t(rbind(OldCurve,NewCurve,ModelCurveAdjustment.1Factor,ModelCurveAdjustment.2Factors,
                ModelCurveAdjustment.3Factors)),type="l",col=c("black","red","green","blue","magenta"),l
legend(x="topleft",c("Old Curve","New Curve","1-Factor Adj.","2-Factor Adj.","3-Factor Adj"),lty=c(1,1,1
```

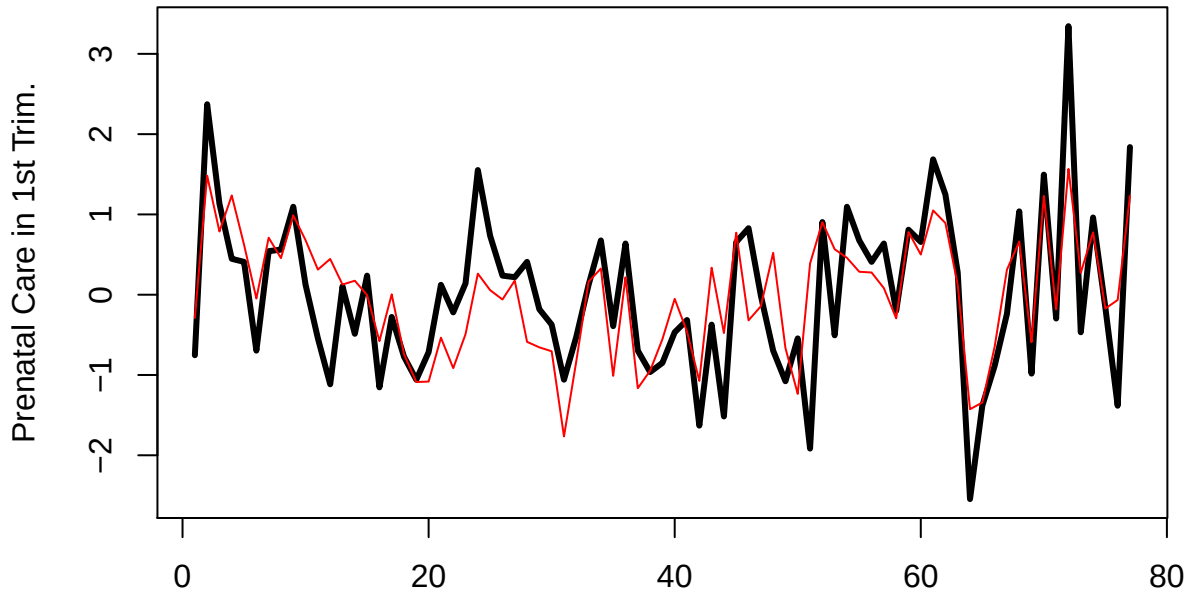



```
# check how well the curve change was estimated by 3 factors
rbind(CurveChange,ModelCurveAdjustment.3Factors-OldCurve)
```

```
##             Birth.Rate General.Fertility.Rate Low.Birth.Weight
## CurveChange -0.1133551              -1.330496       -0.8175935
##             -1.4096925              -1.961907       -0.9266011
##             Prenatal.Care.Beginning.in.First.Trimester Preterm.Births
## CurveChange                                  0.8769064     -0.2983179
##                                              0.5828335     -0.9045186
##             Teen.Birth.Rate Assault..Homicide. Breast.cancer.in.females
## CurveChange       -1.021432          -1.237842               -1.2346208
##                   -1.853100          -1.340275               -0.3556557
##             Cancer..All.Sites. Colorectal.Cancer Diabetes.related
## CurveChange          -1.566194        -1.8452739        0.6000558
##                      -1.012976        -0.6811636       -1.3934957
##             Firearm.related Infant.Mortality.Rate Lung.Cancer
## CurveChange       -1.113245             -1.066058  -1.2530468
##                   -1.268534             -1.070475  -0.7500482
##             Prostate.Cancer.in.Males Stroke..Cerebrovascular.Disease.
## CurveChange               -1.7431255                       -1.398062
##                           -0.9943441                       -1.210237
##             Childhood.Blood.Lead.Level.Screening Childhood.Lead.Poisoning
## CurveChange                           -1.100947                -1.042011
##                                       -1.739620                -1.012673
##             Tuberculosis Below.Poverty.Level Crowded.Housing Dependency
## CurveChange   -1.3295681           -1.469950       -2.515489  -2.118352
##               -0.4110494           -1.285982       -1.684442  -1.819852
##             No.High.School.Diploma Unemployment PERCENT.OF.HOUSING.CROWDED
## CurveChange              -1.893971   -0.8959088                  -3.394076
##                          -1.960618   -1.4528675                  -1.649291
##             PERCENT.HOUSEHOLDS.BELOW.POVERTY PERCENT.AGED.16..UNEMPLOYED
## CurveChange                        -1.665227                   -1.418456
##                                    -1.357239                   -1.581612
##             PERCENT.AGED.25..WITHOUT.HIGH.SCHOOL.DIPLOMA
```

```
## CurveChange                                      -1.903046
##                                                  -1.956046
##             PERCENT.AGED.UNDER.18.OR.OVER.64 HARDSHIP.INDEX
## CurveChange                        -2.224431      -2.614101
##                                    -1.796531      -2.178753
```

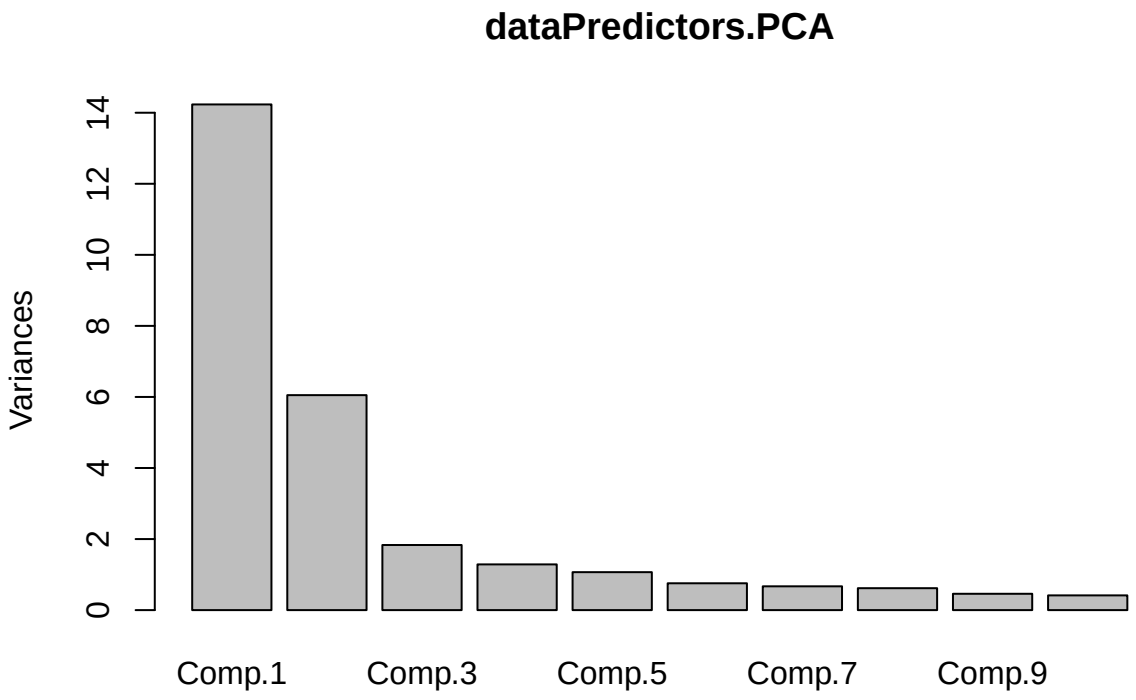
```
# estimate all the values for the 4th column (prenatal) using three terms of factors and loadings
Model.Prenatal<-col.means[4]+L.matrix[4,1]*F.matrix[,1]+L.matrix[4,2]*F.matrix[,2]+L.matrix[4,3]*F.matr
matplot(cbind(Data.Input[,4],Model.Prenatal),type="l",lty=1,lwd=c(3,1),col=c("black","red"),ylab="Prena
```




```
# run PCA on predictors
dataPredictors.PCA <- princomp(dataPredictors)

# explore the PCA object
names(dataPredictors.PCA)
```

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

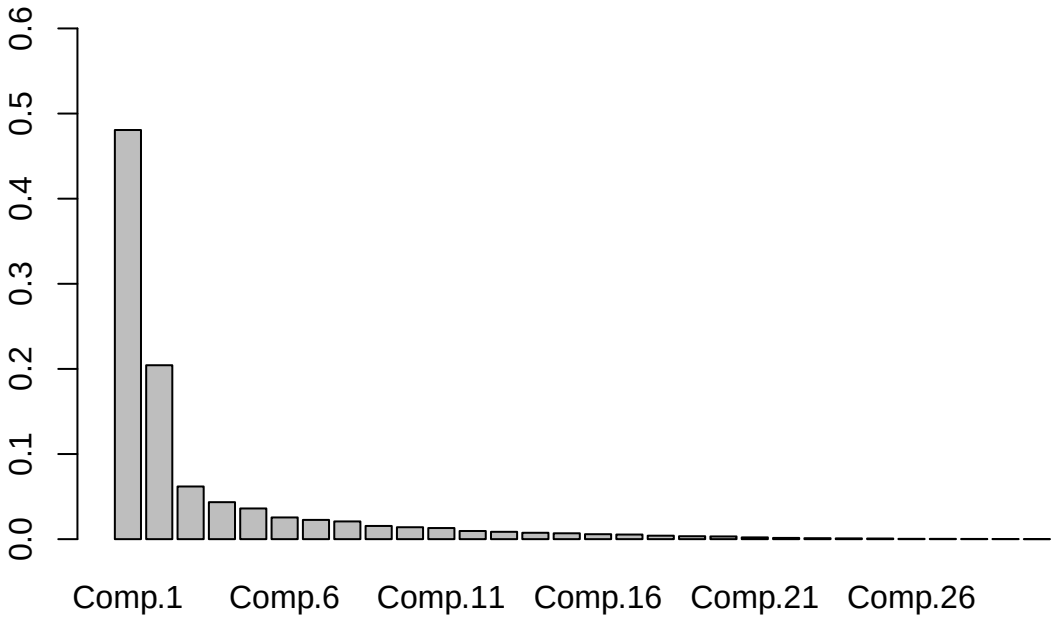
```
# plot the principle components
plot(dataPredictors.PCA)
```




```
# look at the variance of the predictors explained by each priniciple component
dataPredictors.PCA$sdev^2
```

```
##       Comp.1       Comp.2       Comp.3       Comp.4       Comp.5
## 14.233153824  6.050260891  1.832196336  1.286294321  1.067750539
##       Comp.6       Comp.7       Comp.8       Comp.9      Comp.10
##  0.755380838  0.669701139  0.616836975  0.459552874  0.414610600
##      Comp.11      Comp.12      Comp.13      Comp.14      Comp.15
##  0.387005582  0.283934771  0.254869690  0.221581677  0.204373476
##      Comp.16      Comp.17      Comp.18      Comp.19      Comp.20
##  0.172440183  0.160755113  0.120508215  0.102183251  0.094266587
##      Comp.21      Comp.22      Comp.23      Comp.24      Comp.25
##  0.060796816  0.042271426  0.033489946  0.027011876  0.022090165
##      Comp.26      Comp.27      Comp.28      Comp.29      Comp.30
##  0.011300247  0.010235465  0.007309820  0.005340049  0.002886918
```

```
# plot the normalized variances explained by each component
barplot(dataPredictors.PCA$sdev^2/sum(dataPredictors.PCA$sdev^2),ylim=c(0,0.6))
```




```
# evaluate the cumulative variance explained by all components
cumsum(dataPredictors.PCA$sdev^2/sum(dataPredictors.PCA$sdev^2))
```

```
##    Comp.1    Comp.2    Comp.3    Comp.4    Comp.5    Comp.6    Comp.7
## 0.4806811 0.6850101 0.7468869 0.7903275 0.8263875 0.8518982 0.8745153
##    Comp.8    Comp.9   Comp.10   Comp.11   Comp.12   Comp.13   Comp.14
## 0.8953470 0.9108670 0.9248692 0.9379392 0.9475282 0.9561356 0.9636189
##   Comp.15   Comp.16   Comp.17   Comp.18   Comp.19   Comp.20   Comp.21
## 0.9705210 0.9763446 0.9817736 0.9858434 0.9892943 0.9924779 0.9945311
##   Comp.22   Comp.23   Comp.24   Comp.25   Comp.26   Comp.27   Comp.28
## 0.9959587 0.9970897 0.9980020 0.9987480 0.9991296 0.9994753 0.9997222
##   Comp.29   Comp.30
## 0.9999025 1.0000000
```

```
# we must make a decision of how many components to use based on our desired r^2 value

# interpret factor loadings
dataPredictors.Loadings<-dataPredictors.PCA$loadings
dataPredictors.Loadings[,1:5]
```

```
##                                              Comp.1      Comp.2
## Birth.Rate                                -0.08684225 -0.25763708
## General.Fertility.Rate                    -0.12186135 -0.25235423
## Low.Birth.Weight                          -0.20759877  0.12754150
## Prenatal.Care.Beginning.in.First.Trimester  0.15094044 -0.03298429
## Preterm.Births                            -0.19673694  0.16295929
## Teen.Birth.Rate                           -0.22757449 -0.09920135
## Assault..Homicide.                        -0.23782487  0.08469800
## Breast.cancer.in.females                  -0.09485616  0.15169902
## Cancer..All.Sites.                        -0.20951060  0.18383539
## Colorectal.Cancer                         -0.16981598  0.19633206
## Diabetes.related                          -0.21408983  0.03384696
## Firearm.related                           -0.22305544  0.09469853
```

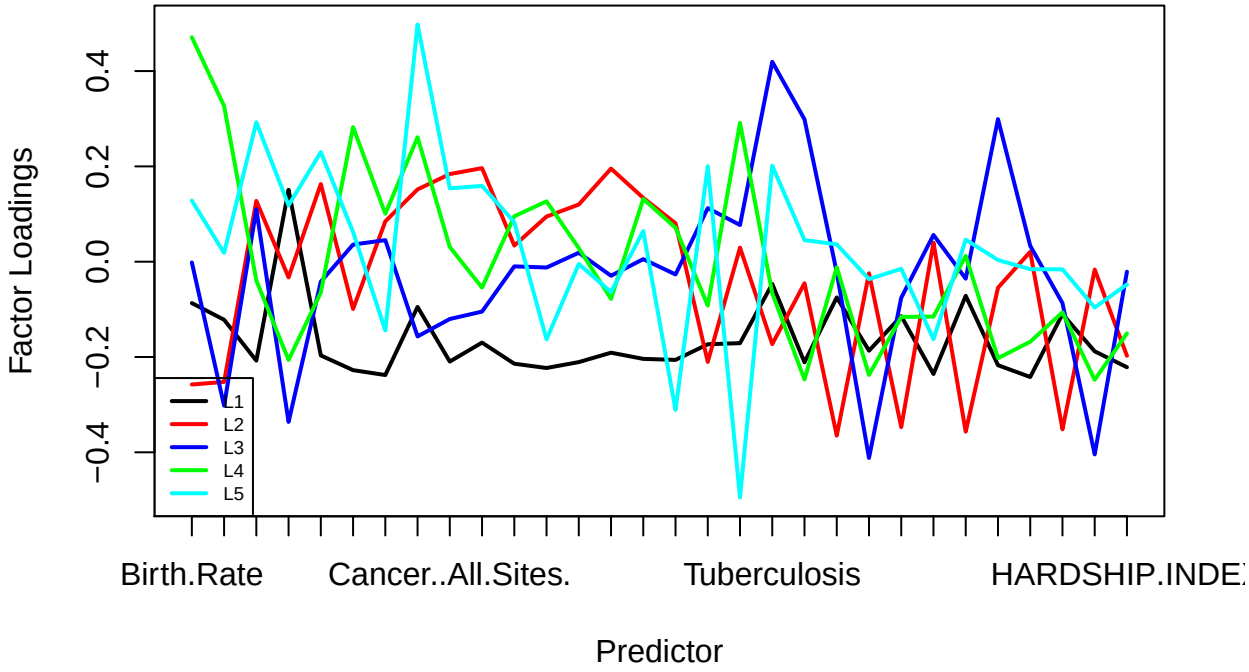
```
## Infant.Mortality.Rate                          -0.21071561  0.11993998
## Lung.Cancer                                    -0.19097114  0.19531884
## Prostate.Cancer.in.Males                       -0.20394201  0.13384539
## Stroke..Cerebrovascular.Disease.               -0.20604988  0.08053919
## Childhood.Blood.Lead.Level.Screening           -0.17328532 -0.21029553
## Childhood.Lead.Poisoning                       -0.17089011  0.02957461
## Tuberculosis                                   -0.04670370 -0.17299660
## Below.Poverty.Level                            -0.21112075 -0.04524534
## Crowded.Housing                                -0.07495312 -0.36492949
## Dependency                                     -0.18673340 -0.02435277
## No.High.School.Diploma                         -0.11432212 -0.34724076
## Unemployment                                   -0.23564897  0.04001850
## PERCENT.OF.HOUSING.CROWDED                     -0.07156888 -0.35665665
## PERCENT.HOUSEHOLDS.BELOW.POVERTY               -0.21720236 -0.05469354
## PERCENT.AGED.16..UNEMPLOYED                    -0.24200361  0.02082010
## PERCENT.AGED.25..WITHOUT.HIGH.SCHOOL.DIPLOMA -0.10991846 -0.35172059
## PERCENT.AGED.UNDER.18.OR.OVER.64               -0.18807809 -0.01647772
## HARDSHIP.INDEX                                 -0.22145392 -0.19744768
##                                                      Comp.3      Comp.4
## Birth.Rate                                     -0.001590648  0.47091967
## General.Fertility.Rate                         -0.302267762  0.32676026
## Low.Birth.Weight                                0.109758979 -0.03932341
## Prenatal.Care.Beginning.in.First.Trimester     -0.336242594 -0.20628335
## Preterm.Births                                 -0.041854797 -0.06472137
## Teen.Birth.Rate                                 0.035912561  0.28238360
## Assault..Homicide.                              0.045238308  0.10066721
## Breast.cancer.in.females                       -0.156828464  0.26114302
## Cancer..All.Sites.                             -0.120259481  0.03087567
## Colorectal.Cancer                              -0.104714822 -0.05439721
## Diabetes.related                               -0.009795640  0.09526213
## Firearm.related                                -0.012234811  0.12667967
## Infant.Mortality.Rate                           0.018713859  0.02843778
## Lung.Cancer                                    -0.029487186 -0.07800621
## Prostate.Cancer.in.Males                        0.005429968  0.13171452
## Stroke..Cerebrovascular.Disease.               -0.026791348  0.07111400
## Childhood.Blood.Lead.Level.Screening            0.112438075 -0.09209481
## Childhood.Lead.Poisoning                        0.076920983  0.29132065
## Tuberculosis                                    0.419419425 -0.06636147
## Below.Poverty.Level                             0.298761018 -0.24699154
## Crowded.Housing                                -0.022411370 -0.01166089
## Dependency                                     -0.411879896 -0.23767532
## No.High.School.Diploma                         -0.075719858 -0.11591666
## Unemployment                                    0.056090210 -0.11521350
## PERCENT.OF.HOUSING.CROWDED                     -0.035541085  0.01228458
## PERCENT.HOUSEHOLDS.BELOW.POVERTY                0.299240196 -0.20243873
## PERCENT.AGED.16..UNEMPLOYED                     0.033312544 -0.16797232
## PERCENT.AGED.25..WITHOUT.HIGH.SCHOOL.DIPLOMA -0.086872186 -0.10620009
## PERCENT.AGED.UNDER.18.OR.OVER.64               -0.404401036 -0.24779119
## HARDSHIP.INDEX                                 -0.020994095 -0.15049778
##                                                      Comp.5
## Birth.Rate                                      0.128309598
## General.Fertility.Rate                          0.018451752
## Low.Birth.Weight                                0.292555486
## Prenatal.Care.Beginning.in.First.Trimester      0.118227435
```

```
## Preterm.Births                                0.230042943
## Teen.Birth.Rate                               0.062092526
## Assault..Homicide.                           -0.144309974
## Breast.cancer.in.females                      0.497767199
## Cancer..All.Sites.                            0.154038096
## Colorectal.Cancer                             0.159024986
## Diabetes.related                              0.082842190
## Firearm.related                              -0.162913734
## Infant.Mortality.Rate                        -0.004887082
## Lung.Cancer                                  -0.063553444
## Prostate.Cancer.in.Males                      0.064165974
## Stroke..Cerebrovascular.Disease.             -0.311083701
## Childhood.Blood.Lead.Level.Screening          0.200419304
## Childhood.Lead.Poisoning                     -0.494558082
## Tuberculosis                                  0.201457999
## Below.Poverty.Level                           0.045325511
## Crowded.Housing                               0.036379056
## Dependency                                   -0.036572376
## No.High.School.Diploma                       -0.014911659
## Unemployment                                 -0.162571370
## PERCENT.OF.HOUSING.CROWDED                    0.046286282
## PERCENT.HOUSEHOLDS.BELOW.POVERTY              0.003554838
## PERCENT.AGED.16..UNEMPLOYED                  -0.015744043
## PERCENT.AGED.25..WITHOUT.HIGH.SCHOOL.DIPLOMA -0.016052604
## PERCENT.AGED.UNDER.18.OR.OVER.64             -0.096106406
## HARDSHIP.INDEX                               -0.047729744
```

```
# compare the eigen vectors with the loadings obtained from PCA
Project.Data.PCA.Eigen.Loadings <- cbind(eigen.vecs[,1:3], dataPredictors.PCA$loadings[,1:3])
colnames(Project.Data.PCA.Eigen.Loadings) <- c('L1.eigen', 'L2.eigen', 'L3.eigen', 'L1.PCA', 'L2.PCA',

head(Project.Data.PCA.Eigen.Loadings)
```

```
##                                           L1.eigen    L2.eigen
## Birth.Rate                               -0.08684225 -0.25763708
## General.Fertility.Rate                   -0.12186135 -0.25235423
## Low.Birth.Weight                         -0.20759877  0.12754150
## Prenatal.Care.Beginning.in.First.Trimester  0.15094044 -0.03298429
## Preterm.Births                           -0.19673694  0.16295929
## Teen.Birth.Rate                          -0.22757449 -0.09920135
##                                              L3.eigen      L1.PCA
## Birth.Rate                               -0.001590648 -0.08684225
## General.Fertility.Rate                   -0.302267762 -0.12186135
## Low.Birth.Weight                          0.109758979 -0.20759877
## Prenatal.Care.Beginning.in.First.Trimester -0.336242594  0.15094044
## Preterm.Births                           -0.041854797 -0.19673694
## Teen.Birth.Rate                           0.035912561 -0.22757449
##                                              L2.PCA       L3.PCA
## Birth.Rate                               -0.25763708 -0.001590648
## General.Fertility.Rate                   -0.25235423 -0.302267762
## Low.Birth.Weight                          0.12754150  0.109758979
## Prenatal.Care.Beginning.in.First.Trimester -0.03298429 -0.336242594
## Preterm.Births                            0.16295929 -0.041854797
## Teen.Birth.Rate                          -0.09920135  0.035912561
```

```
# plot loadings
matplot(1:30,dataPredictors.PCA$loadings[,1:5],type="l",lty=1,lwd=2,xaxt="n",xlab="Predictor",ylab="Fac
axis(1, 1:30,labels=colnames(dataPredictors))
legend("bottomleft",legend=c("L1","L2","L3","L4","L5"),lty=1,lwd=2,cex=.6,col=c("black","red","blue","g
```




```
# create a new data frame with principal components as predictors
dataPCAFactors<-dataPredictors.PCA$scores
dataRotated<-as.data.frame(cbind(Output=combined.data$Per.Capita.Income,dataPCAFactors))

# look at the factors (scores)
matplot(dataPredictors.PCA$scores[,1:3],type="l",lty=1,lwd=2)

# compare the F.matrix with the factors/scores obtained from PCA
F.matrix.PCA <- dataPredictors.PCA$scores[,1:3]
Project.Data.PCA.Eigen.Factors <- cbind(F.matrix[,1:3], F.matrix.PCA)
colnames(Project.Data.PCA.Eigen.Factors) <- c('F.1', 'F.2', 'F.3', 'F1.PCA', 'F2.PCA', 'F3.PCA')

head(Project.Data.PCA.Eigen.Factors)
```

```
##           F.1        F.2        F.3   F1.PCA     F2.PCA     F3.PCA
## [1,] 1.663945 -0.9139935  1.6973315 1.663945 -0.9139935  1.6973315
## [2,] 4.358773  1.4971463 -2.5993422 4.358773  1.4971463 -2.5993422
## [3,] 3.057494  1.0934682 -1.0762576 3.057494  1.0934682 -1.0762576
## [4,] 4.981442  1.7645671 -1.6126368 4.981442  1.7645671 -1.6126368
## [5,] 3.321844  0.2533361 -0.3832937 3.321844  0.2533361 -0.3832937
## [6,] 1.379687 -3.7744107  1.1307452 1.379687 -3.7744107  1.1307452
```

```
# compare coefficients with factor loadings

# look at the intercepts and slopes for each predictor on the output (per.capita.income)
coeff.check <- t(apply(Data.Input, 2, function(Data.Input.col) lm(Data.Input.col~Data.Output)$coef))
```

```r
# look at the relationships of factors and the column response. This looks at the correlation (r^2) of
(rSqrCorrelations<-apply(dataPredictors.PCA$scores,2,cor,combined.data$Per.Capita.Income)^2)
```

```
##       Comp.1       Comp.2       Comp.3       Comp.4       Comp.5
## 5.179087e-01 1.624835e-01 7.089616e-02 2.922804e-02 1.869183e-03
##       Comp.6       Comp.7       Comp.8       Comp.9      Comp.10
## 5.224372e-03 1.168425e-03 6.865645e-03 1.754651e-02 4.670547e-03
##      Comp.11      Comp.12      Comp.13      Comp.14      Comp.15
## 3.391851e-04 2.537088e-02 2.220239e-05 1.359729e-03 1.947948e-03
##      Comp.16      Comp.17      Comp.18      Comp.19      Comp.20
## 7.153041e-05 5.788106e-03 4.392331e-04 8.844222e-03 1.358613e-02
##      Comp.21      Comp.22      Comp.23      Comp.24      Comp.25
## 2.904301e-04 3.946211e-03 1.692742e-03 7.347551e-04 3.958284e-03
##      Comp.26      Comp.27      Comp.28      Comp.29      Comp.30
## 3.866929e-05 1.658575e-04 1.830331e-03 5.336677e-03 1.186658e-03
```

```r
sum(rSqrCorrelations)
```

```
## [1] 0.8948109
```

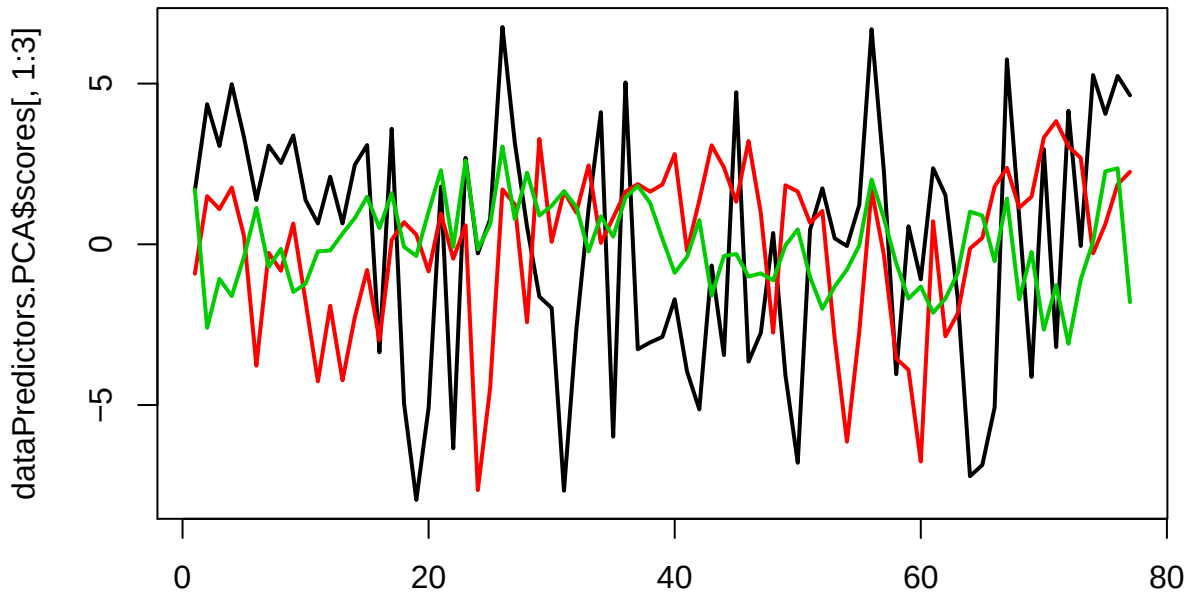
```r
# this is the same r^2 as in the summary of the linear model

# fit a linear model with the PCA factors as predictors
linModPCA <- lm(Output ~ ., data =dataRotated)
summary(linModPCA)
```

```
##
## Call:
## lm(formula = Output ~ ., data = dataRotated)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10929.6  -2409.1   -685.4   2266.3  23615.1
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  25106.7      710.4  35.343  < 2e-16 ***
## Comp.1        2833.7      188.3  15.049  < 2e-16 ***
## Comp.2        2434.4      288.8   8.429 6.94e-11 ***
## Comp.3        2922.2      524.8   5.568 1.28e-06 ***
## Comp.4        2239.3      626.3   3.575 0.000836 ***
## Comp.5         621.5      687.5   0.904 0.370650
## Comp.6        1235.4      817.3   1.512 0.137499
## Comp.7        -620.5      868.1  -0.715 0.478335
## Comp.8        1567.2      904.5   1.733 0.089842 .
## Comp.9        2902.7     1047.9   2.770 0.008057 **
## Comp.10      -1576.7     1103.2  -1.429 0.159718
## Comp.11        439.8     1141.9   0.385 0.701913
## Comp.12       4440.6     1333.1   3.331 0.001713 **
## Comp.13        138.7     1407.1   0.099 0.921935
## Comp.14      -1163.7     1509.1  -0.771 0.444581
```

```
## Comp.15       1450.3     1571.4   0.923 0.360845
## Comp.16       -302.6     1710.7  -0.177 0.860392
## Comp.17       2818.8     1771.8   1.591 0.118466
## Comp.18       -896.8     2046.3  -0.438 0.663242
## Comp.19       4370.4     2222.3   1.967 0.055275 .
## Comp.20      -5639.6     2313.7  -2.437 0.018715 *
## Comp.21       1026.7     2881.0   0.356 0.723185
## Comp.22       4538.9     3455.1   1.314 0.195474
## Comp.23       3339.8     3881.8   0.860 0.394043
## Comp.24      -2450.0     4322.2  -0.567 0.573575
## Comp.25      -6288.3     4779.6  -1.316 0.194805
## Comp.26       -869.0     6682.6  -0.130 0.897102
## Comp.27      -1891.0     7021.5  -0.269 0.788891
## Comp.28       7433.5     8308.7   0.895 0.375626
## Comp.29     -14850.6     9721.1  -1.528 0.133443
## Comp.30       9524.1    13221.2   0.720 0.474941
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6234 on 46 degrees of freedom
## Multiple R-squared:  0.8948, Adjusted R-squared:  0.8262
## F-statistic: 13.04 on 30 and 46 DF,  p-value: 4.426e-14
```

```
# calculate relative importance measures for the PCA factors
suppressMessages(library(relaimpo))
```

```
## Warning: package 'survey' was built under R version 3.3.2
```




```
metrics.data.pca <- calc.relimp(linModPCA, type = c("first", "last"))
metrics.data.pca
```

```
## Response variable: Output
## Total response variance: 223582409
```

```
## Analysis based on 77 observations
##
## 30 Regressors:
## Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7 Comp.8 Comp.9 Comp.10 Comp.11 Comp.12 Comp.13 Comp.
## Proportion of variance explained by model: 89.48%
## Metrics are not normalized (rela=FALSE).
##
## Relative importance metrics:
##
##                last        first
## Comp.1  5.179087e-01 5.179087e-01
## Comp.2  1.624835e-01 1.624835e-01
## Comp.3  7.089616e-02 7.089616e-02
## Comp.4  2.922804e-02 2.922804e-02
## Comp.5  1.869183e-03 1.869183e-03
## Comp.6  5.224372e-03 5.224372e-03
## Comp.7  1.168425e-03 1.168425e-03
## Comp.8  6.865645e-03 6.865645e-03
## Comp.9  1.754651e-02 1.754651e-02
## Comp.10 4.670547e-03 4.670547e-03
## Comp.11 3.391851e-04 3.391851e-04
## Comp.12 2.537088e-02 2.537088e-02
## Comp.13 2.220239e-05 2.220239e-05
## Comp.14 1.359729e-03 1.359729e-03
## Comp.15 1.947948e-03 1.947948e-03
## Comp.16 7.153041e-05 7.153041e-05
## Comp.17 5.788106e-03 5.788106e-03
## Comp.18 4.392331e-04 4.392331e-04
## Comp.19 8.844222e-03 8.844222e-03
## Comp.20 1.358613e-02 1.358613e-02
## Comp.21 2.904301e-04 2.904301e-04
## Comp.22 3.946211e-03 3.946211e-03
## Comp.23 1.692742e-03 1.692742e-03
## Comp.24 7.347551e-04 7.347551e-04
## Comp.25 3.958284e-03 3.958284e-03
## Comp.26 3.866929e-05 3.866929e-05
## Comp.27 1.658575e-04 1.658575e-04
## Comp.28 1.830331e-03 1.830331e-03
## Comp.29 5.336677e-03 5.336677e-03
## Comp.30 1.186658e-03 1.186658e-03
```

```
# sum the variances explained by each component to get the total variance explained by the model
sum(metrics.data.pca@first)
```

```
## [1] 0.8948109
```

```
metrics.data.pca@first.rank
```

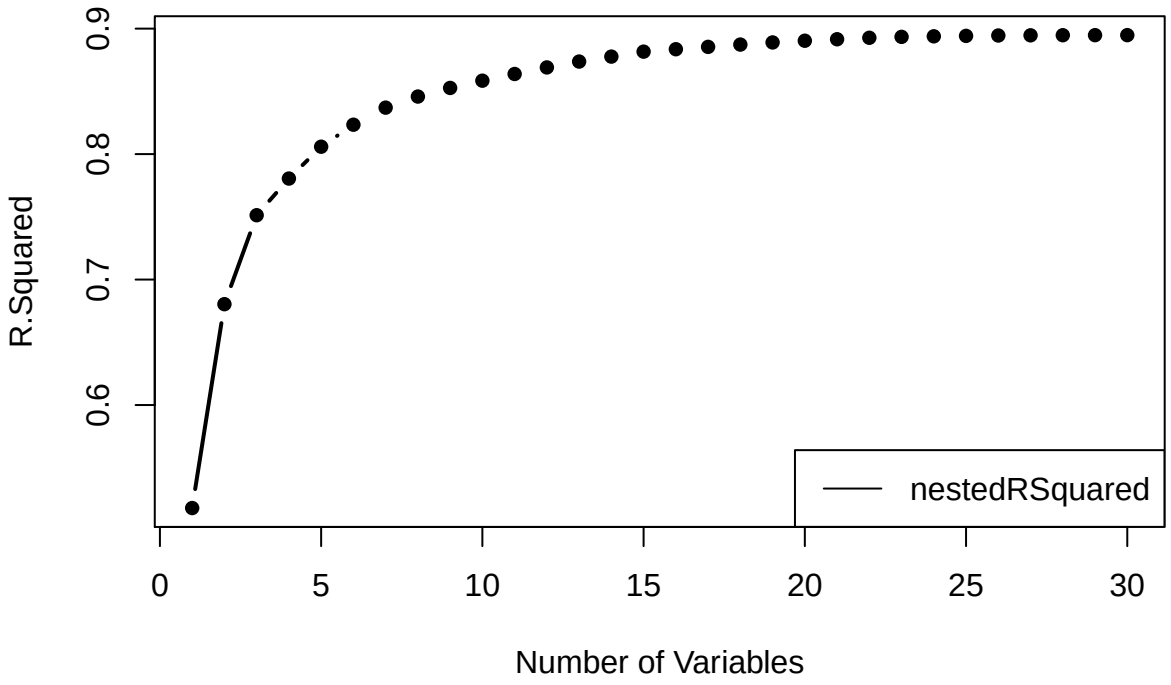
```
##  Comp.1  Comp.2  Comp.3  Comp.4  Comp.5  Comp.6  Comp.7  Comp.8  Comp.9
##       1       2       3       4      17      12      22       9       6
## Comp.10 Comp.11 Comp.12 Comp.13 Comp.14 Comp.15 Comp.16 Comp.17 Comp.18
##      13      25       5      30      20      16      28      10      24
## Comp.19 Comp.20 Comp.21 Comp.22 Comp.23 Comp.24 Comp.25 Comp.26 Comp.27
```

```
##       8       7      26      15      19      23      14      29      27
## Comp.28 Comp.29 Comp.30
##      18      11      21
```

```
# re-order the components from high importance to low importance
orderComponents <- order(metrics.data.pca@first.rank)

# fit the sequence of linear models
dataRotatedReordered<-dataRotated[,c(1,orderComponents+1)]
(nestedRSquared<-sapply(2:31,function(z) summary(lm(Output~.,data=dataRotatedReordered[,1:z]))$r.square
```

```
##  [1] 0.5179087 0.6803923 0.7512884 0.7805165 0.8058874 0.8234339 0.8370200
##  [8] 0.8458642 0.8527299 0.8585180 0.8638546 0.8690790 0.8737496 0.8777078
## [15] 0.8816541 0.8836020 0.8854712 0.8873015 0.8889943 0.8903540 0.8915406
## [22] 0.8927091 0.8934438 0.8938831 0.8942222 0.8945127 0.8946785 0.8947501
## [29] 0.8947887 0.8948109
```

```
# plot the r^2 values
matplot(1:30, nestedRSquared,type="b",xlab="Number of Variables",ylab="R.Squared",lty=1,lwd=2,pch=16)
legend("bottomright",legend="nestedRSquared",lty=1,lwd=1,col="black")
```



Part E: Restoring slopes for Original Predictors

```
# retrieve the order that the principle components should be in based on the relative importance measur
(PCA.rank<-metrics.data.pca@last.rank)
```

```
##  Comp.1  Comp.2  Comp.3  Comp.4  Comp.5  Comp.6  Comp.7  Comp.8  Comp.9
##       1       2       3       4      17      12      22       9       6
## Comp.10 Comp.11 Comp.12 Comp.13 Comp.14 Comp.15 Comp.16 Comp.17 Comp.18
##      13      25       5      30      20      16      28      10      24
```

```
## Comp.19 Comp.20 Comp.21 Comp.22 Comp.23 Comp.24 Comp.25 Comp.26 Comp.27
##       8       7      26      15      19      23      14      29      27
## Comp.28 Comp.29 Comp.30
##      18      11      21
```

```
# re-order the loadings according to this order
orderedLoadings<-dataPredictors.Loadings[,order(PCA.rank)]
# re-order the coefficients from the linear model with factors as predictors according to this order
orderedCoefficientsPCA<-linModPCA$coefficients[-1][order(PCA.rank)]

# multiply the ordered loading matrix by the vector of order coefficients to get the slopes of the orig
restoredCoefficients<-orderedLoadings%*%orderedCoefficientsPCA
cbind(restoredCoefficients,linMod$coefficients[-1])
```

```
##                                                [,1]        [,2]
## Birth.Rate                                  5203.8053   5203.8053
## General.Fertility.Rate                     -6450.7441  -6450.7441
## Low.Birth.Weight                           -2672.2227  -2672.2227
## Prenatal.Care.Beginning.in.First.Trimester -4008.0901  -4008.0901
## Preterm.Births                              4276.8285   4276.8285
## Teen.Birth.Rate                              127.1846    127.1846
## Assault..Homicide.                         -1103.4362  -1103.4362
## Breast.cancer.in.females                    1227.5568   1227.5568
## Cancer..All.Sites.                         -1224.4509  -1224.4509
## Colorectal.Cancer                            435.7621    435.7621
## Diabetes.related                           -5908.9593  -5908.9593
## Firearm.related                             -948.3048   -948.3048
## Infant.Mortality.Rate                      -1085.7303  -1085.7303
## Lung.Cancer                                -1399.7540  -1399.7540
## Prostate.Cancer.in.Males                    2037.9505   2037.9505
## Stroke..Cerebrovascular.Disease.            3182.4761   3182.4761
## Childhood.Blood.Lead.Level.Screening        -759.9257   -759.9257
## Childhood.Lead.Poisoning                    -949.3696   -949.3696
## Tuberculosis                               -3377.7554  -3377.7554
## Below.Poverty.Level                         2923.8253   2923.8253
## Crowded.Housing                             -411.1604   -411.1604
## Dependency                                  -500.2851   -500.2851
## No.High.School.Diploma                     12807.4494  12807.4494
## Unemployment                               -3576.8536  -3576.8536
## PERCENT.OF.HOUSING.CROWDED                  -338.1916   -338.1916
## PERCENT.HOUSEHOLDS.BELOW.POVERTY           -2231.3610  -2231.3610
## PERCENT.AGED.16..UNEMPLOYED                 2000.1349   2000.1349
## PERCENT.AGED.25..WITHOUT.HIGH.SCHOOL.DIPLOMA -13337.7899 -13337.7899
## PERCENT.AGED.UNDER.18.OR.OVER.64           -3060.8656  -3060.8656
## HARDSHIP.INDEX                             -3551.0405  -3551.0405
```