Code

ANLY_699_Assignment4

Subhash Pemmaraju

July 5, 2020

Description of the data including dependent variable

Hide

```
fa_data <- merged_data[, c(6,9:27)]
fa_data1 <- fa_data[complete.cases(fa_data),]
#dim(fa_data1)
str(fa_data1)
```

```
## 'data.frame':    2828 obs. of  20 variables:
##  $ park_access         : int  20 20 27 38 16 2 2 29 14 22 ...
##  $ yrs_plr             : int  8129 7354 10254 11978 11335 11680 14360 12079 11113 12350 ...
##  $ perc_fair_poor_health: int  21 18 30 19 22 31 28 23 24 21 ...
##  $ avg_phy_unh_days    : num  4.7 4.2 5.4 4.6 4.9 5.4 5.4 4.9 4.9 4.8 ...
##  $ avg_mental_unh_days : num  4.7 4.3 5.2 4.6 4.9 4.9 5.3 4.8 4.9 4.7 ...
##  $ perc_smokers        : int  18 17 22 19 19 23 22 21 19 17 ...
##  $ perc_obese          : int  33 31 42 38 34 37 43 39 40 35 ...
##  $ food_env_ind        : num  7.2 8 5.6 7.8 8.4 4.3 6.6 6.9 6.4 8.3 ...
##  $ perc_phy_inact      : int  35 27 24 34 30 25 40 32 30 31 ...
##  $ perc_excess_drink   : num  15 18 12.8 15.6 14.2 ...
##  $ perc_uninsured      : int  9 11 12 10 13 11 11 12 12 11 ...
##  $ perc_college        : int  62 67 35 44 53 35 42 59 48 52 ...
##  $ perc_unemp          : num  3.6 3.6 5.2 4 3.5 4.7 4.8 4.7 3.9 3.6 ...
##  $ perc_child_pov      : int  19 14 44 28 18 68 36 27 31 25 ...
##  $ perc_diabetes       : int  11 11 18 15 17 24 19 18 20 15 ...
##  $ median_income       : int  59338 57588 34382 46064 50412 29267 37365 45400 39917 42132
...
##  $ perc_65up           : num  15.6 20.4 19.4 16.5 18.2 16.4 20.3 17.7 19.5 23 ...
##  $ perc_black          : num  19.3 8.8 48 21.1 1.5 69.5 44.6 20.9 39.6 4.2 ...
##  $ perc_female         : num  51.4 51.5 47.2 46.8 50.7 45.5 53.4 51.9 52.1 50.5 ...
##  $ perc_18less         : num  23.7 21.6 20.9 20.5 23.2 21.1 22.2 21.6 20.8 19.2 ...
```

Hide

```
#names(fa_data1)
```

The dataset contains 20 variables.

The dimensionality of the data is 2828, 20

The datatypes are all numeric/integer as shown in the table.

The names of the variables are: park_access, yrs_plr, perc_fair_poor_health, avg_phy_unh_days, avg_mental_unh_days, perc_smokers, perc_obese, food_env_ind, perc_phy_inact, perc_excess_drink, perc_uninsured, perc_college, perc_unemp, perc_child_pov, perc_diabetes, median_income, perc_65up, perc_black, perc_female, perc_18less

Correlation Matrix of all the variables

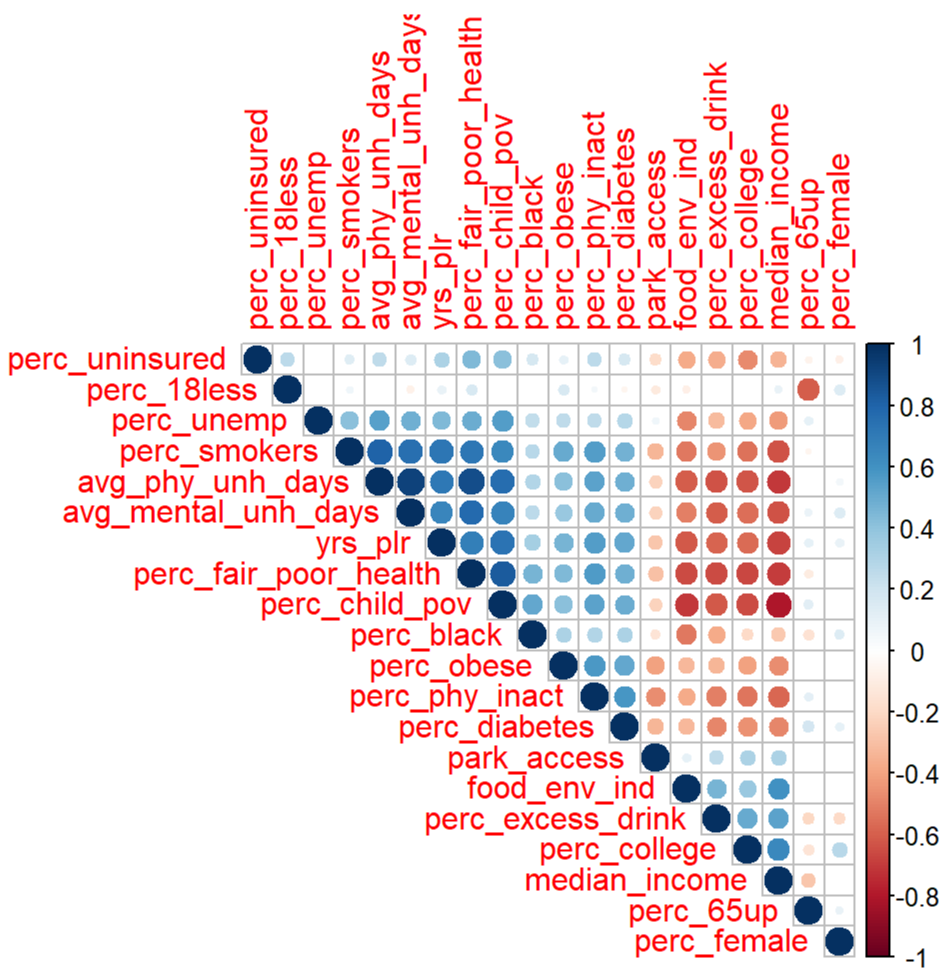
Hide

```
corr_matrix <- cor(fa_data1)
#corr_matrix
#corrplot(corr_matrix, order="hclust", type="upper", tl.srt = 45)
```

Correlation Plot

Hide

```
#Plot statistically significant correlations
res2 <- rcorr(as.matrix(fa_data1), type="pearson")
# Extract the correlation coefficients
#res2$r
# Extract p-values
#res2$P
# Insignificant correlations are leaved blank
corrplot(res2$r, type="upper", order="hclust",
         p.mat = res2$P, sig.level = 0.01, insig = "blank")
```



As can be seen from the correlation plot, many of the variables are highly correlated with each other. This leads to multicollinearity. We can confirm this fact by calculating the Variance Inflation factor for a regression of the dependent variable against all the independent variables

Hide

```
model <- lm(yrs_plr ~., data = fa_data1)
vif(model)
```

```
##          park_access perc_fair_poor_health      avg_phy_unh_days
##             1.541772             11.011236             20.589187
##   avg_mental_unh_days          perc_smokers            perc_obese
##            10.170958              3.856079              2.001760
##         food_env_ind         perc_phy_inact     perc_excess_drink
##             2.946384              2.352652              2.539704
##       perc_uninsured          perc_college            perc_unemp
##             2.081942              3.538793              1.857192
##       perc_child_pov         perc_diabetes         median_income
##             7.223652              1.912462              4.498737
##            perc_65up            perc_black           perc_female
##             3.029743              2.650452              1.902329
##          perc_18less
##             2.804652
```

The VIF has several variables with VIF being high >> 2.5. Therefore multicollinearity is a big problem.

KMO Analysis

we now conduct the KMO test to check whether factor analysis is the right approach for this.

Hide

```
data_fa <- fa_data1[,-1:-2]
datamatrix <- cor(data_fa)
KMO(r=datamatrix)
```

```
## Kaiser-Meyer-Olkin factor adequacy
## Call: KMO(r = datamatrix)
## Overall MSA =  0.85
## MSA for each item =
## perc_fair_poor_health      avg_phy_unh_days   avg_mental_unh_days
##                  0.90                  0.85                  0.86
##          perc_smokers            perc_obese          food_env_ind
##                  0.95                  0.89                  0.87
##        perc_phy_inact     perc_excess_drink        perc_uninsured
##                  0.94                  0.93                  0.75
##          perc_college            perc_unemp        perc_child_pov
##                  0.85                  0.89                  0.91
##         perc_diabetes         median_income             perc_65up
##                  0.95                  0.91                  0.41
##            perc_black           perc_female           perc_18less
##                  0.65                  0.21                  0.35
```

KMO Output

From the KMO test result, we can see that overall MSA > 0.5 and therefore, factor analysis is appropriate here. Notice that we dropped the dependent variable of years of potential lives lost as well as the core independent variable (% of population within half a mile of a park).

Eigen Values and Optimal number of Factors

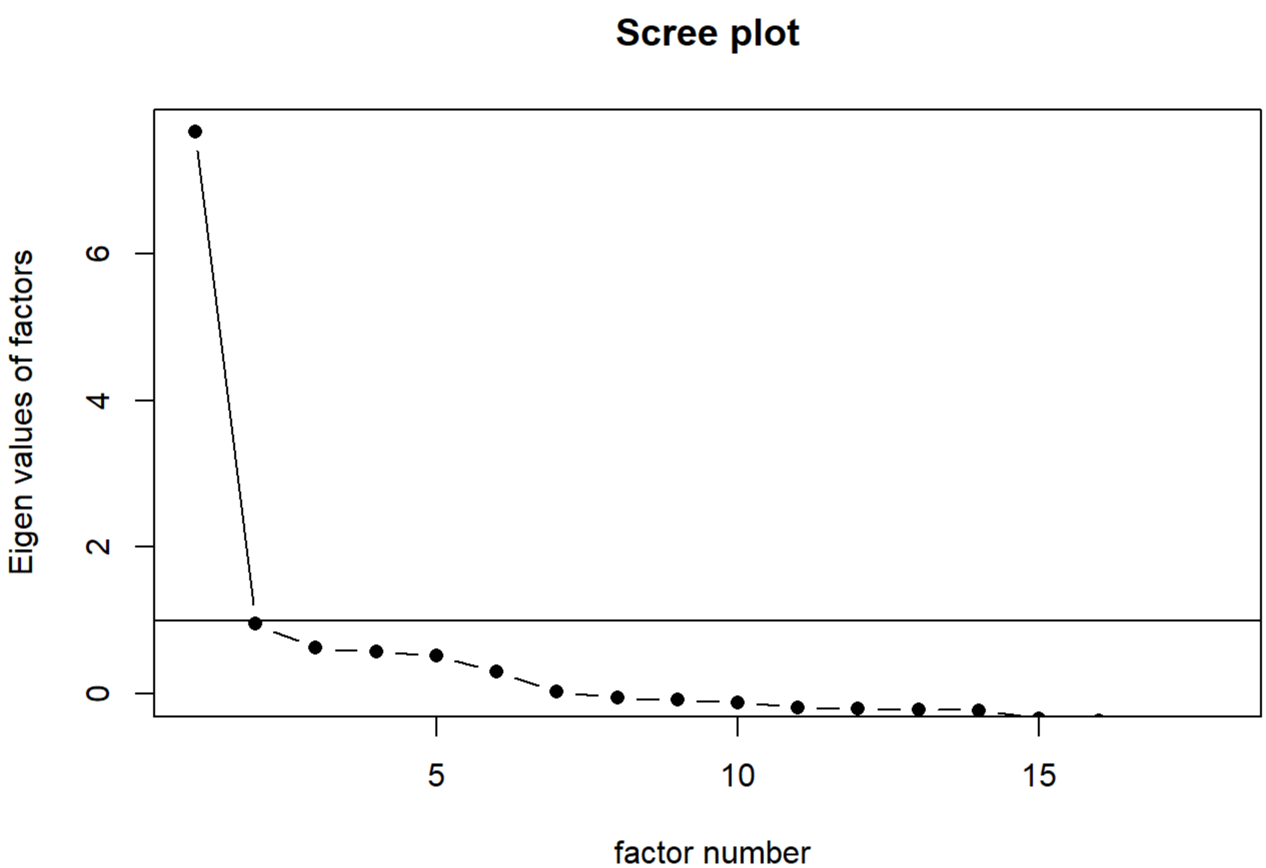
Hide

```
ev <- eigen(cor(data_fa))
ev$values
```

```
##  [1] 8.05877320 1.84448993 1.35045457 1.17455891 1.13921951 0.97511632
##  [7] 0.66543453 0.54698421 0.41587549 0.36743725 0.35775446 0.30494408
## [13] 0.22821449 0.19301932 0.15015195 0.10759378 0.08698198 0.03299602
```

Hide

```
scree(data_fa, factors=TRUE, pc=FALSE)
```



From the Scree Plot shown above, we can conclude that the optimal number of factors is 2.

Factor Analysis

Hide

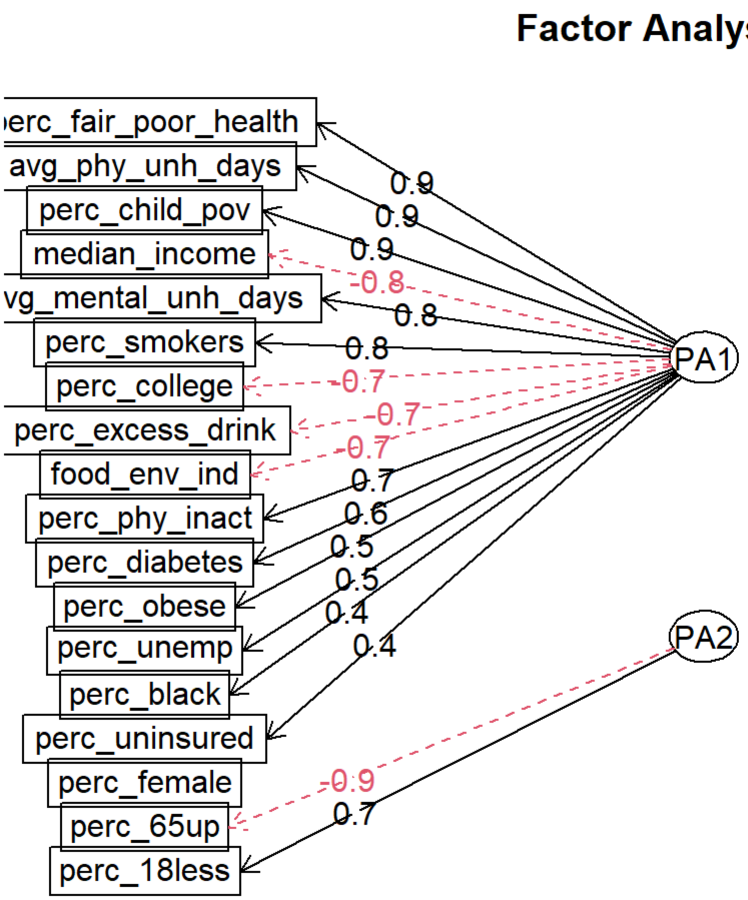
```
nfactors <- 2
fit1 <-factanal(data_fa,nfactors,scores = c("regression"),rotation = "varimax")
print(fit1)
```

```
##
## Call:
## factanal(x = data_fa, factors = nfactors, scores = c("regression"),     rotation = "varimax")
##
## Uniquenesses:
## perc_fair_poor_health      avg_phy_unh_days   avg_mental_unh_days
##                 0.128                 0.005                 0.124
##          perc_smokers            perc_obese          food_env_ind
##                 0.339                 0.746                 0.463
##        perc_phy_inact     perc_excess_drink        perc_uninsured
##                 0.600                 0.520                 0.717
##          perc_college            perc_unemp        perc_child_pov
##                 0.482                 0.681                 0.127
##         perc_diabetes         median_income            perc_65up
##                 0.678                 0.321                 0.991
##            perc_black           perc_female           perc_18less
##                 0.685                 0.990                 0.993
##
## Loadings:
##                       Factor1 Factor2
## perc_fair_poor_health  0.863   0.356
## avg_phy_unh_days       0.995
## avg_mental_unh_days    0.936
## perc_smokers           0.803   0.128
## perc_obese             0.394   0.315
## food_env_ind          -0.579  -0.450
## perc_phy_inact         0.509   0.375
## perc_excess_drink     -0.617  -0.316
## perc_uninsured         0.221   0.484
## perc_college          -0.595  -0.405
## perc_unemp             0.530   0.196
## perc_child_pov         0.725   0.589
## perc_diabetes          0.470   0.318
## median_income         -0.674  -0.474
## perc_65up
## perc_black             0.259   0.498
## perc_female
## perc_18less
##
##                Factor1 Factor2
## SS loadings      6.341   2.071
## Proportion Var   0.352   0.115
## Cumulative Var   0.352   0.467
##
## Test of the hypothesis that 2 factors are sufficient.
## The chi square statistic is 11201.96 on 118 degrees of freedom.
## The p-value is 0
```

Factor analysis results are shown above. As can be seen, Factor 1 explains 35.2% of the variance and Factor 2 explains 11.5% of the variance. Together 46.7% of the variance is explained. We plot the factor analysis diagram as shown below. As we can see Factor2 has only age variables - % of population over 65 and % of population less than 18. Factor1 has variables on demographics and pre-existing health conditions.

Hide

```
fa_var <-  fa(r=data_fa, nfactors = 2, rotate = "varimax", fm="pa")
fa.diagram(fa_var)
```



Hide

```
regdata <- cbind(fa_data1[1],fa_data1[2], fa_var$scores)
#Labeling the data

names(regdata) <- c("park_access", "yrs_plr", "health_demo",
                    "age_dist")
```

Regression Analysis

Hide

```
#Regression Model using train data

model1 = lm(yrs_plr~., regdata)
summary(model1)
```

```
##
## Call:
## lm(formula = yrs_plr ~ ., data = regdata)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7731.1  -891.6   -78.8   766.7 14872.3
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 8606.694     46.636  184.55  < 2e-16 ***
## park_access   -1.384      1.384   -1.00  0.31722
## health_demo 2050.148     32.000   64.07  < 2e-16 ***
## age_dist     -79.609     30.615   -2.60  0.00936 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1577 on 2824 degrees of freedom
## Multiple R-squared:  0.6236, Adjusted R-squared:  0.6232
## F-statistic:  1560 on 3 and 2824 DF,  p-value: < 2.2e-16
```

We now carry out regression analysis with the factors as shown above. The dependent variable (yrs_plr) is regressed against the primary independent variable (park_access) and the two factor variables (health_demo and age_dist). As can be seen from the regression results, R^2 is high at 62% and the coefficient of park_access is not statistically significant. This suggests that pre-existing health conditions and demographics and age explain all of the impact on years of potential lives lost. We also check the regression for multicollinearity. Since we used factor analysis, multi-collinearity shouldn't be a problem. As we can see, the VIF < 2.5.

Hide

```
vif(model1)
```

```
## park_access health_demo    age_dist
##    1.126658    1.126561    1.000113
```