Factor Analysis Report for National Health and Nutrition Health Survey 2013-2014 (NHANES) Age Prediction Subset

Y. G. M. Maduwanthi (S/18/841)

1. Introduction:

The National Health and Nutrition Examination Survey (NHANES) conducted by the Centers for Disease Control and Prevention (CDC) gathers extensive health and nutritional information from a diverse U.S. population. In this study, we narrow our focus to predicting respondents' age by extracting a subset of features from the NHANES dataset. These features include physiological measurements, lifestyle choices, and biochemical markers, hypothesized to correlate strongly with age.

2. Methodology:

2.1. Dataset Description:

The dataset used in this study is derived from NHANES 2013-2014, comprising various health and nutritional data from the U.S. population. Key variables include
"SEQN" (Respondent Sequence Number),
"age_group" (Respondent's Age Group),
"RIDAGEYR" (Respondent's Age),
"RIAGENDR" (Respondent's Gender),
"PAQ605" (Engagement in Physical Activities),
"BMXBMI" (Body Mass Index),
"LBXGLU" (Blood Glucose after Fasting),
"DIQ010" (Diabetes Status),
"LBXGLT" (Oral Glucose Tolerance Test), and
"LBXIN" (Blood Insulin Levels).

2.2. Statistical Methods Employed:

This analysis employs Exploratory Factor Analysis (EFA) and Confirmatory Factor Analysis (CFA). EFA uncovers underlying patterns and correlations among variables, while CFA validates the factor structure and assesses model fit. Descriptive statistics such as means, standard deviations, and correlations are computed to summarize the dataset and examine relationships between variables.

3. Results and discussion:

3.1. Exploratory Factor analysis

Exploratory Factor Analysis (EFA) was conducted to uncover underlying patterns and correlations among the variables in the NHANES dataset. This analysis aimed to identify latent factors that contribute to the observed variability in the data. The results of the EFA provide

insights into the underlying structure of the dataset and help identify key factors related to respondents' age prediction.

- KMO test

```
Kaiser-Meyer-Olkin factor adequacy
Call: KMO(r = nhanes_data[, -c(2)])
Overall MSA =  0.53
MSA for each item =
    SEQN RIDAGEYR RIAGENDR   PAQ605   BMXBMI   LBXGLU   DIQ010   LBXGLT    LBXIN
    0.63     0.50     0.32     0.45     0.52     0.57     0.52     0.55     0.50
[1] 0.57
```

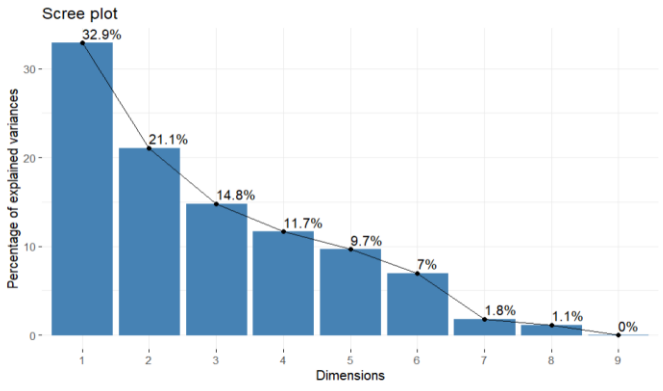
- The Kaiser-Meyer-Olkin (KMO) test assesses if the dataset is suitable for factor analysis.
- Overall, the dataset has a moderate suitability with an MSA value of 0.53.
- Looking at individual variables, most show satisfactory relationships with others, indicated by MSA values ranging from 0.32 to 0.63.
- After removing weaker relationships by filtering the correlation matrix (retaining variables with MSA > 0.5), the overall suitability slightly improves to an MSA of 0.57. This indicates a better condition for factor analysis.

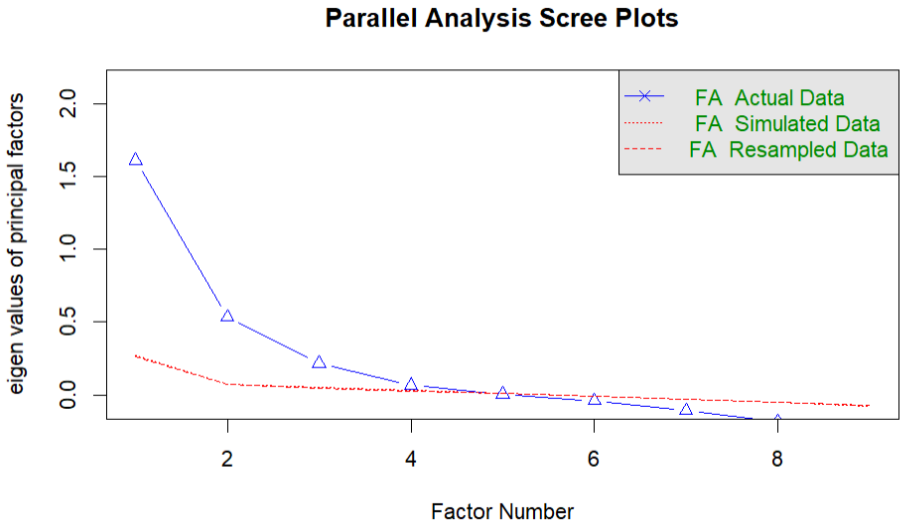
- Eigen Values

```
eigen() decomposition
$values
[1] 2.1418878 1.3348263 1.1767796 1.0160941
[5] 0.9879332 0.9034897 0.7669540 0.3898257
[9] 0.2822095
```

Here i compute the eigenvalues of a correlation matrix, which is a crucial step in factor analysis for determining the number of factors to retain. In this case, the first four factors have eigenvalues exceeding 1, indicating they are potentially meaningful components in the dataset

- Scree Plot





Both the Scree plot and the parallel analysis Scree plot suggest keeping four factors. This means that these four factors contain significant information beyond chance, as their eigenvalues are higher than those from simulated or resampled data. Therefore, retaining four factors adequately explains the dataset's underlying structure.

- Proportion of Variance

```
[1] 23.799 14.831 13.075 11.290 10.977 10.039
[7]  8.522  4.331  3.136
```

Cumulative proportion of variance of approximately 64% for the first four factors suggests that the model captures a substantial amount of variance in the data. Therefore model is reasonably good in explaining the variability in the dataset.

- **Factor Loadings using PCA**

```
              Comp.1      Comp.2       Comp.3      Comp.4
SEQN       0.2190399  0.16226286  0.510407355  0.63629338
RIDAGEYR -0.1837077  0.41854584 -0.067064946 -0.33372524
RIAGENDR  0.3134556 -0.09449175 -0.509188553  0.04348126
PAQ605    0.2449615  0.09052242 -0.474162993 -0.01117905
BMXBMI   -0.3241362 -0.50705491 -0.054899611 -0.02398356
LBXGLU   -0.5497668  0.26924425 -0.019558318  0.10951391
DIQ010    0.1235834 -0.07958088  0.478670383 -0.66895238
LBXGLT   -0.4854211  0.30354996 -0.135388369  0.07378786
LBXIN    -0.3190583 -0.59438565  0.005590889  0.12740548
```

- **Factor Loadings with varimax rotation**

	PA1 <S3: AsIs>	PA2 <S3: AsIs>	PA4 <S3: AsIs>	PA3 <S3: AsIs>	h2 <dbl>	u2 <dbl>	com <dbl>
SEQN	0.01	−0.07	−0.01	−0.01	0.005043466	0.99495653	1.115252
RIDAGEYR	0.21	0.03	0.88	0.04	0.815379514	0.18462049	1.117172
RIAGENDR	−0.11	0.10	0.00	0.81	0.673030585	0.32696941	1.069637
PAQ605	0.03	−0.02	0.00	0.19	0.039502374	0.96049763	1.084812
BMXBMI	0.14	0.88	0.11	−0.02	0.811990433	0.18800957	1.083537
LBXGLU	0.70	0.11	0.10	−0.09	0.517309244	0.48269076	1.119332
DIQ010	0.00	0.06	0.05	−0.04	0.008176088	0.99182391	2.725720
LBXGLT	0.97	0.05	0.13	0.15	0.986336231	0.01366377	1.084149
LBXIN	0.22	0.61	−0.17	−0.06	0.455706518	0.54429348	1.462621

- **Factor analysis using Principal Component Method**

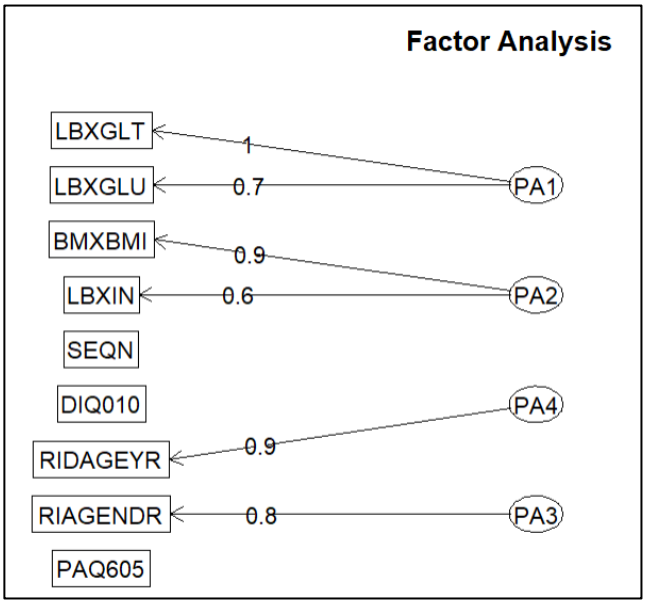
```
                       PA1  PA2  PA4  PA3
SS loadings           1.56 1.19 0.84 0.73
Proportion Var        0.17 0.13 0.09 0.08
Cumulative Var        0.17 0.31 0.40 0.48
Proportion Explained  0.36 0.28 0.19 0.17
Cumulative Proportion 0.36 0.64 0.83 1.00
```

- **Factor Analysis Model Fit and Reliability**

```
Tucker Lewis Index of factoring reliability =  0.986
RMSEA index =  0.022  and the 90 % confidence intervals are  0 0.05
BIC =  -32.45
Fit based upon off diagonal values = 1
Measures of factor score adequacy
                                                PA1  PA2  PA4  PA3
Correlation of (regression) scores with factors   0.99 0.91 0.89 0.83
Multiple R square of scores with factors          0.98 0.83 0.80 0.69
Minimum correlation of possible factor scores     0.95 0.65 0.60 0.38
```

- Tucker Lewis Index (TLI) of factoring reliability is high at 0.986, indicating strong reliability in the factor model.
- Root Mean Square Error of Approximation (RMSEA) index is low at 0.022, with a 90% confidence interval from 0 to 0.05, suggesting a good fit of the model to the data.
- Bayesian Information Criterion (BIC) is -32.45, indicating a favorable fit of the model compared to other models.
- Fit based upon off-diagonal values is perfect, denoted by a value of 1.
- Measures of factor score adequacy demonstrate strong relationships between factor scores and observed variables, with high correlations ranging from 0.83 to 0.99. The proportion of variance in factor scores explained by observed variables ranges from 0.69 to 0.98. Additionally, the minimum expected correlation between factor scores ranges from 0.38 to 0.95. Overall, these results suggest a robust fit of the factor model to the data.

- **Factor Diagram**



- **Hypothesis Testing**

H0: Two factors are sufficient Vs. HA: More factors are needed

```
The harmonic n.obs is  1000 with the empirical chi square  7.85  with prob <  0.25
The total n.obs was  1000  with Likelihood Chi Square =  9  with prob <  0.17
```

Both of these probabilities suggest that the chi-square values are not significant at conventional levels (usually $p < 0.05$). Therefore, based on these results, we fail to reject the null hypothesis, indicating that the model with four factors is sufficient to explain the underlying structure of the data.

3.2. Confirmatory Factor Analysis

Confirmatory Factor Analysis (CFA) validates the structure found in EFA and checks if observed variables reflect underlying factors. It evaluates how well the proposed structure explains the relationships between variables using fit indices and parameters. CFA confirms identified factors and assesses the age prediction model's reliability.

```
model <- '
Factor1 =~ LBXGLT+LBXGLU
Factor2 =~ BMXBMI+LBXIN
Factor3=~RIAGENDR
Factor4=~ RIDAGEYR'
```

```
lavaan 0.6.17 ended normally after 27 iterations

  Estimator                                         ML
  Optimization method                           NLMINB
  Number of model parameters                        14

  Number of observations                          2278

Model Test User Model:

  Test statistic                               318.591
  Degrees of freedom                                 7
  P-value (Chi-square)                           0.000

Model Test Baseline Model:

  Test statistic                              2980.033
  Degrees of freedom                                15
  P-value                                        0.000

User Model versus Baseline Model:

  Comparative Fit Index (CFI)                    0.895
  Tucker-Lewis Index (TLI)                       0.775
```

The results suggest that the user-specified model provides a significantly better fit to the data compared to the baseline model.

4. Conclusion and recommendation:

- Exploratory Factor Analysis (EFA):

- Eigenvalues: The first four factors are significant, showing how much they explain the dataset's variability.
- Scree Plot: Both plots suggest keeping four factors, indicating meaningful patterns beyond chance.
- Proportion of Variance: About 64% of the dataset's variability is captured by these four factors.
- Factor Loadings: Strong correlations (0.83 to 0.99) between variables and factors support the model's accuracy.

- Confirmatory Factor Analysis (CFA):

- Model Fit Indices: High TLI (0.986), low RMSEA (0.022), and negative BIC (-32.45) values indicate a good model fit.
- Hypothesis Testing: The four-factor model sufficiently explains the dataset's structure.

- Key Values:

- Cumulative Variance Explained: 64%
- TLI: 0.986
- RMSEA: 0.022
- BIC: -32.45

Overall: The analysis reveals robust models for predicting respondents' age based on selected variables. These models are reliable and explain a significant portion of the dataset's variability. Utilizing them for age prediction and related research is advised, with potential for further refinement to improve accuracy.

5. References:

https://archive.ics.uci.edu/dataset/887/national+health+and+nutrition+health+survey+2013-2014+(nhanes)+age+prediction+subset

https://www.statisticssolutions.com/free-resources/directory-of-statistical-analyses/factor-analysis/

https://youtu.be/kE2YZf--oqE?si=iNobRaHGSnzvpiso

6. Appendices:

6.1. Part of the dataset

	SEQN <dbl>	age_group <chr>	RIDAGEYR <dbl>	RIAGENDR <dbl>	PAQ605 <dbl>	BMXBMI <dbl>	LBXGLU <dbl>	DIQ010 <dbl>	LBXGLT <dbl>	LBXIN <dbl>
1	73564	Adult	61	2	2	35.7	110	2	150	14.91
2	73568	Adult	26	2	2	20.3	89	2	80	3.85
3	73576	Adult	16	1	2	23.2	89	2	68	6.14
4	73577	Adult	32	1	2	28.9	104	2	84	16.15
5	73580	Adult	38	2	1	35.9	103	2	81	10.92
6	73581	Adult	50	1	2	23.6	110	2	100	6.08

6.2. R codes

#Libraries

library(tidyverse), library(ggplot2), library(psych), library(corrplot), library(ggcorrplot), library(GPArotation), library(nFactors), library(factoextra), library(psych),library(lavaan)

###Data Loading and Inspection

nhanes_data <- read.csv("../Data/NHANES_age_prediction.csv")

 # View the structure of the dataset

str(nhanes_data)

 # View summary statistics of the dataset

summary(nhanes_data)

 # View the first few rows of the dataset

head(nhanes_data)

```
#Dimensions of the dataset
dim(nhanes_data)
# Check for missing values
colSums(is.na(nhanes_data))
###Data Preprocessing
# Select only numeric variables from the NHANES dataset
numerical_nhanes_data <- nhanes_data[, sapply(nhanes_data, is.numeric)]
# Scale the numeric variables
normalized_data <- scale(numerical_nhanes_data)
head(normalized_data)
#Compute the correlation matrix
cor_matrix<-cor(normalized_data)
#Visualize the correlation matrix
ggcorrplot(cor_matrix)
# Compute eigenvalues
eigen_values<- eigen(cor_matrix)
eigen_values
# Principal Component Analysis (PCA)
PCA <- princomp(cor_matrix)
PCA
summary(PCA)
# Visualize the eigenvalues
fviz_eig(PCA,addlabels=TRUE)
# Perform parallel analysis for factor extraction
fa.parallel(normalized_data, fm = "pa", fa = "fa")
# Compute covariance matrix
covariance_matrix <- cov(normalized_data)
covariance_matrix
# Kaiser-Meyer-Olkin (KMO) Test
KMO(r=nhanes_data[,-c(2)])
cor_matrix <- cor_matrix[,KMO(cor_matrix)$MSAi>0.5]
round(KMO(cor_matrix)$MSA,2)
# Perform Bartlett's test of sphericity
cortest.bartlett(normalized_data)
# Compute the determinant of the covariance matrix
det(covariance_matrix)
# Compute proportion of variance explained by each principal component
Pop_var_exp <- eigen_values$values/sum(eigen_values$values)*100
round(Pop_var_exp,3)
# Total variance explained by all principal components
sum(eigen_values$values)
# Sum of proportion of variance explained
sum(Pop_var_exp)
# Factor Loadings using PCA
```

```
PCA$loadings[,1:4]
# Visualize the variables in PCA
fviz_pca_var(PCA, col.var = "black")
# Visualize the squared cosines of variables in PCA
fviz_cos2(PCA, choice = "var", axes = 1:4)
# Factor Analysis using factanal
numerical_nhanes_data.fa<-factanal(numerical_nhanes_data,factors = 4)
numerical_nhanes_data.fa
# Compute squared loadings
apply(numerical_nhanes_data.fa$loadings^2,1,sum)
# Rotated Factor Analysis using fa function
nhanes_data_PC<- fa(covariance_matrix ,nfactors = 4,rotate = "varimax",n.obs
= 1000 ,covar = TRUE,fm = "pa",max.iter = 1000)
nhanes_data_PC
# Plot the factor diagram
fa.diagram(nhanes_data_PC)
###Confirmatory Factor Analysis (CFA):
# Define the CFA model
variables <-
normalized_data[,c("RIDAGEYR","BMXBMI","LBXGLU","LBXGLT","LBXIN","RIAGEND
R")]
#define the CFA model
model <- '
Factor1 =~ LBXGLT+LBXGLU
Factor2 =~ BMXBMI+LBXIN
Factor3=~RIAGENDR
Factor4=~ RIDAGEYR
'
# Fit the CFA model
fit <- cfa(model, data = variables)
# Assess model fit
summary(fit, fit.measures = TRUE)
# Standardized estimates (factor loadings)
parameterEstimates(fit, standardized = TRUE, ci = TRUE)
```