ST 405 – Multivariate Methods II

Mini Project

S/18/SP/608 – Akila Premathilake

Introduction

Factor analysis serves as a statistical method employed to explore the interrelationships among numerous observed variables, aiming to unveil a smaller set of latent variables, termed factors. My objective in this project lies in simplifying the intricacies within the dataset while revealing the underlying structure that elucidates the observed data.

Within predictive modeling, the reduction in variable count facilitated by factor analysis proves instrumental in mitigating challenges such as overfitting, multicollinearity, and the complexities associated with high dimensionality. This process encompasses both exploratory factor analysis (EFA) and confirmatory factor analysis (CFA), each serving distinct purposes.

Methodology

The dataset originates from the *National Institute of Diabetes and Digestive and Kidney Diseases*, intended for diagnosing diabetes based on specific diagnostic measurements. It focuses on predicting whether a patient has diabetes or not. All subjects included are female adults of Pima Indian ancestry, aged 21 and above.

The dataset encompasses various medical predictor variables alongside one target variable, Outcome. For this analysis, eight predictors are utilized:

- Pregnancies, Glucose, BloodPressure, SkinThickness, Insulin, BMI, DiabetesPedigreeFunction, and Age.

Exploratory Factor Analysis (EFA)

The primary objectives of an EFA are to determine

- The number of common factors influencing a set of measures
- The strength of the relationship between each factor and each observed measure
- The factor scores

some common uses of EFA are to

- To reduce a large number of variables to a smaller number of factors for modeling purposes, where a large number of variables precludes modeling all the measures individually.
- To select a subset of variables from a larger set, based on which original variables have the highest correlations with the principal component factors.
- To create a set of factors to be treated as uncorrelated variables as one approach to handling multicollinearity in such procedures as multiple regression

Confirmatory Factor Analysis (CFA)

Confirmatory factor analysis (CFA) can be used to study how well a hypothesized factor model fits a new sample from the same population or a sample from a different population. The CFA model is the same as the EFA model with the exception that restrictions can be placed on factor loadings, variances, covariances, and residual variances resulting in a more parsimonious model. Using CFA, one can

- Investigate if a factor model fits a new sample from the same population – the confirmatory aspect.
- Evaluate if a factor model fits a sample from a different population – measurement invariance
- Study the behavior of new measurement items embedded in a previously studied measurement instrument, and
- Estimate factor scores

Results and Discussion

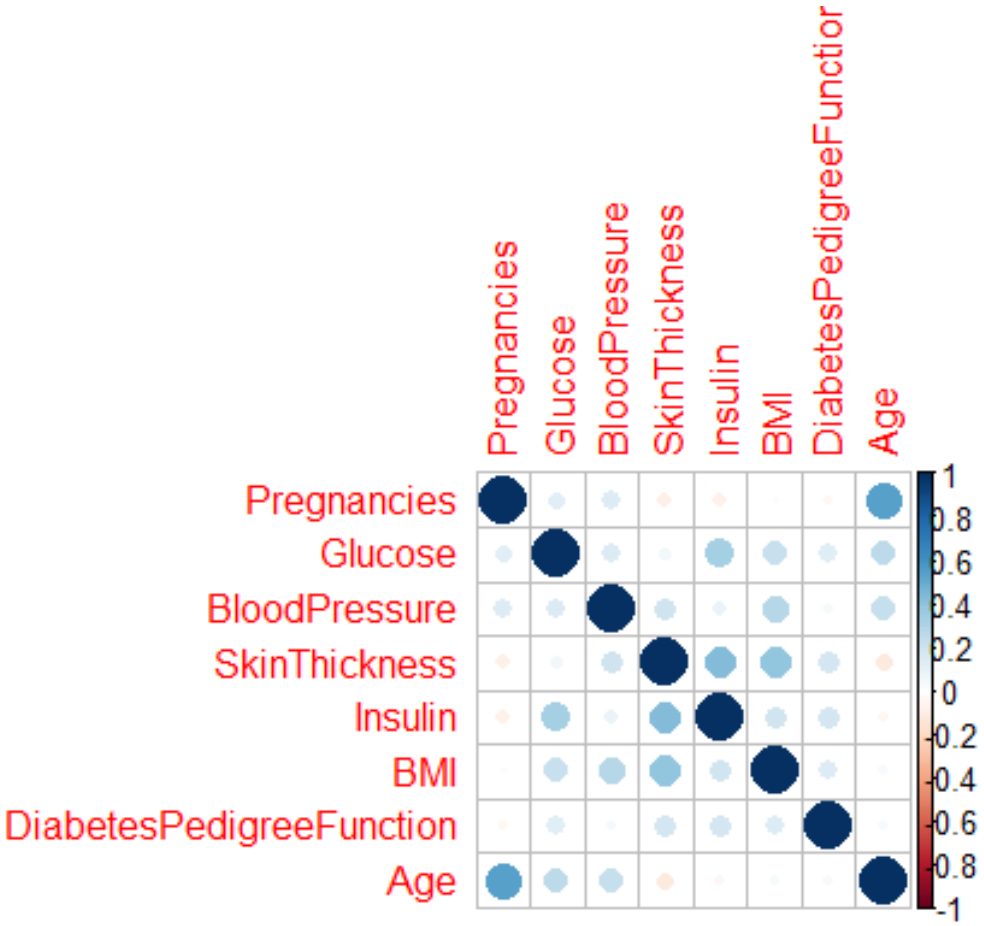


Figure 1: Correlation Plot

Bartlett's test of sphericity

Chi-square: 948.2262

P-value: 1.25755e-181

df:28

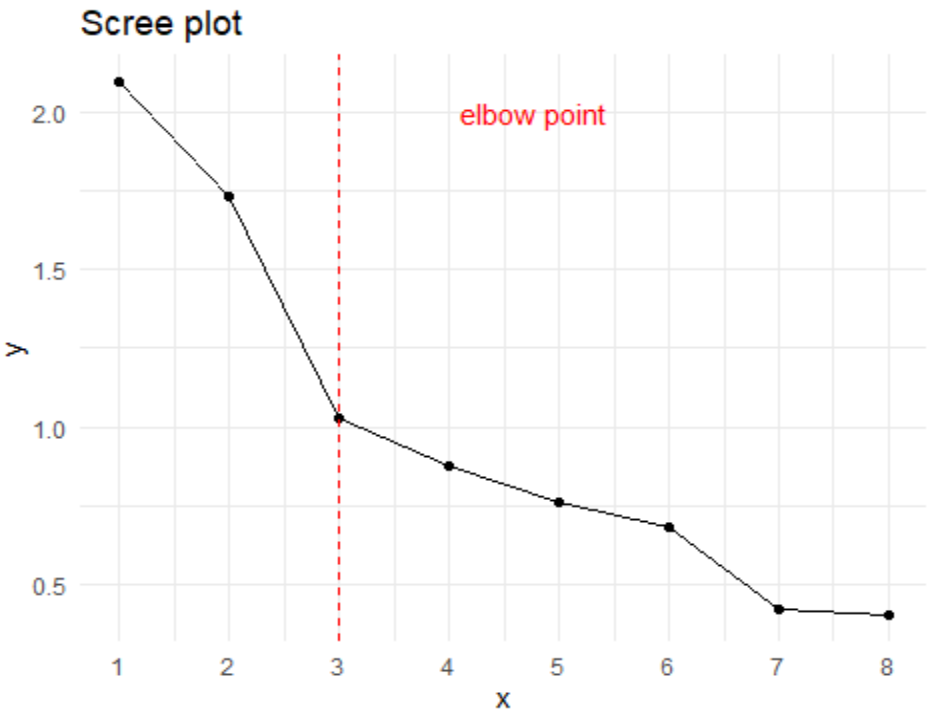


Figure 2: Scree Plot

Eigen Value	**Cumulative Proportion**
2. 0943799	0. 2617975
1. 7312101	0. 4781988
1. 0296299	0. 6069025
0. 8755290	0. 7163436
0. 7623444	0. 8116367
0. 6826284	0. 8969652
0. 4198162	0. 9494422
0. 4044620	1.0000000

Table 1: Eigen Values and Cumulative Proportions

```
Loadings:
                         Factor1 Factor2 Factor3
Pregnancies               -0.298   0.133   0.529
Glucose                           0.997
BloodPressure              0.125   0.155   0.339
SkinThickness              0.849           0.297
Insulin                    0.459   0.333
BMI                        0.379   0.223   0.199
DiabetesPedigreeFunction   0.187   0.138
Age                       -0.408   0.268   0.729

                Factor1 Factor2 Factor3
SS loadings       1.381   1.292   1.065
Proportion Var    0.173   0.161   0.133
Cumulative Var    0.173   0.334   0.467

Test of the hypothesis that 3 factors are sufficient.
The chi square statistic is 39 on 7 degrees of freedom.
The p-value is 1.95e-06
```

Figure 3: Factor Loadings

```
Loadings:
                         Factor1 Factor2 Factor3
Pregnancies                       0.619
Glucose                                   0.953
BloodPressure
SkinThickness              0.886
Insulin                    0.515
BMI                        0.467
DiabetesPedigreeFunction
Age                               0.869

                Factor1 Factor2 Factor3
SS loadings       1.455   1.263   1.020
Proportion Var    0.182   0.158   0.128
Cumulative Var    0.182   0.340   0.467

Test of the hypothesis that 3 factors are sufficient.
The chi square statistic is 39 on 7 degrees of freedom.
The p-value is 1.95e-06
```

Figure 4: Factor Rotations

Factor1	Factor2	Factor3
0.95683200	1.5192617	0.3677035
0.19649075	-0.2537621	-1.1804957
-1.24427322	-0.1154597	2.3615440
-0.08141387	-1.1273334	-0.8159980
1.27414283	-0.4216552	0.2993244
-1.43085723	-0.2048091	0.2277392

Figure 5: Factor Scores

Latent Variables:

	Estimate	Std.Err	z-value	P(>|z|)
metabolic_health =~				
SkinThickness	1.000			
Insulin	6.731	0.722	9.326	0.000
BMI	0.342	0.039	8.736	0.000
reproductive_health =~				
Pregnancies	1.000			
Age	6.280	1.141	5.505	0.000
glucose_metabolism =~				
Glucose	1.000			

Figure 6: Latent Variables

Covariances:

	Estimate	Std.Err	z-value	P(>|z|)
metabolic_health ~~				
reprodctv_hlth	-1.878	0.979	-1.918	0.055
glucose_mtblsm	106.269	16.910	6.284	0.000
reproductive_health ~~				
glucose_mtblsm	15.741	3.606	4.365	0.000

Figure 7: Covariances

Variances:

	Estimate	Std.Err	z-value	P(>|z|)
.SkinThickness	136.751	13.696	9.985	0.000
.Insulin	7945.596	674.837	11.774	0.000
.BMI	48.323	2.914	16.585	0.000
.Pregnancies	7.909	0.718	11.007	0.000
.Age	2.840	23.442	0.121	0.904
.Glucose	0.000			
metabolic_hlth	117.391	16.074	7.303	0.000
reprodctv_hlth	3.430	0.725	4.733	0.000
glucose_mtblsm	1020.917	52.098	19.596	0.000

Figure 8: Variances

EFA

I used three methods to select the optimal number of factors. According to the eigenvalues, selected factors such a way that eig. values >1. Also, I used cumulative performance of the eig values (60.69%). Thirdly, I used Scree plot to determine the number of factors. According to the three plot, I used three factors. (Table 1)

common factor part is based on the four factors, the uniqueness part is also called uniqueness factor, which is specific to each observed variable. Factor loadings are the coefficients of the latent factors on observed variables. ex: Pregnancies = -0. 298*Factor1 + 0. 133*Factor2 + 0. 529*Factor3 (Figure 3) SS Loadings is the sum squared loadings related to each other factor. It is the overall variance explained in all the 8 variables by each factor. Therefor, the 1st factor explains the total of 1.381 variance, that's about 1.381/8=17.26% (Figure 3). Proportion Var is the variance in the observed variables explained by each factor. Cumulative Var is the cumulative proportion of variance explained by all factors. According to the chi-square test we fail to reject the null hypothesis, i.e. the factor model have a good fit to the data. After applying the rotation method, we can identify patterns of the factors. For example, the variable SkinThickness has a large loading 0. 886 on Factor1. In this case, we might say that the variable SkinThickness is mainly influenced by Factor2. (Figure 4) Based on the rotated factor loadings, we can name the factors in the model. For example, the Factor1 is indicated by SkinThickness, Insulin, and BMI, all of are related to Metabolic health. Similarly, the second is called the reproductive health, the 3rd is called glucose metabolism.

CFA

CFA also test the significance of the factor loadings based on a z-test. For example, the factor loading for Insulin on factor 1 is 4.350 with the standard error 0.537 with p-value almost 0. Therefore, this factor loading is statistically significant from 0.

- The chi-square statistic is 118.142 with df 7 and p-value 0. Therefore, one would reject the hypothesis that the model fits the data simply based on it.

- CFI is 0.854, which is smaller than the cut-off value 0.95. it also suggests a bad fit
- The RMSEA = 0.144, which do not lie in the range of reasonable fit model
- The SRMR = 0.061, which indicate the model is not a good one.

Conclusion and recommendation

Overall, the model fit indices suggest that the specified model may not fully capture the relationships between the observed variables and the latent factors. It is advisable to consider model modifications, such as adding or removing paths, covariances, or allowing for correlated residuals, to improve the fit of the model to the data.

References

1. Gündoğdu, Y., Karabağlı, P., Alptekin, H., Şahin, M., & Kılıç, H. (2019, November). Comparison of performances of Principal Component Analysis (PCA) and Factor Analysis (FA) methods on the identification of cancerous and healthy colon tissues. International Journal of Mass Spectrometry, 445, 116204. doi:10.1016/j.ijms.2019.116204

2. Kustra. (2006). A factor analysis model for functional genomics. BMC Bioinformatics, 7. doi:10.1186/1471-2105-7-21

3. Yong, A., & Pearce, S. (2013, October 1). A Beginner's Guide to Factor Analysis: Focusing on Exploratory Factor Analysis. Tutorials in Quantitative Methods for Psychology, 9, 79-94. doi:10.20982/tqmp.09.2.p079

4. Zhang, Z., & Wang, L. (2017). In Advanced statistics using R. Granger. ISDSA Press. doi:10.35566/advstats

Appendices

Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age
6	148	72	35	0	33.6	0.627	50
1	85	66	29	0	26.6	0.351	31
8	183	64	0	0	23.3	0.672	32
1	89	66	23	94	28.1	0.167	21
0	137	40	35	168	43.1	2.288	33
5	116	74	0	0	25.6	0.201	30
3	78	50	32	88	31.0	0.248	26
10	115	0	0	0	35.3	0.134	29
2	197	70	45	543	30.5	0.158	53
8	125	96	0	0	0.0	0.232	54
4	110	92	0	0	37.6	0.191	30
10	168	74	0	0	38.0	0.537	34
10	139	80	0	0	27.1	1.441	57
1	189	60	23	846	30.1	0.398	59

Figure 9: Dataset Preview

R Markdown

ST405 - Mini Project

S/18/SP/608 | S.D.A.V.S.Preamthilake

2024-04-07

```
data <- read.csv("D:/University/4th yr/ST405/mini project/factor analysis/diabetes.csv")
head(data)
```

```
##   Pregnancies Glucose BloodPressure SkinThickness Insulin  BMI
## 1           6     148            72            35       0 33.6
## 2           1      85            66            29       0 26.6
## 3           8     183            64             0       0 23.3
## 4           1      89            66            23      94 28.1
## 5           0     137            40            35     168 43.1
## 6           5     116            74             0       0 25.6
##   DiabetesPedigreeFunction Age Outcome
## 1                    0.627  50       1
## 2                    0.351  31       0
## 3                    0.672  32       1
```

```
## 4                    0.167  21       0
## 5                    2.288  33       1
## 6                    0.201  30       0
```

```
str(data)
```

```
## 'data.frame':    768 obs. of  9 variables:
##  $ Pregnancies             : int  6 1 8 1 0 5 3 10 2 8 ...
##  $ Glucose                 : int  148 85 183 89 137 116 78 115 197 125 ...
##  $ BloodPressure           : int  72 66 64 66 40 74 50 0 70 96 ...
##  $ SkinThickness           : int  35 29 0 23 35 0 32 0 45 0 ...
##  $ Insulin                 : int  0 0 0 94 168 0 88 0 543 0 ...
##  $ BMI                     : num  33.6 26.6 23.3 28.1 43.1 25.6 31 35.3 30
.5 0 ...
##  $ DiabetesPedigreeFunction: num  0.627 0.351 0.672 0.167 2.288 ...
##  $ Age                     : int  50 31 32 21 33 30 26 29 53 54 ...
##  $ Outcome                 : int  1 0 1 0 1 0 1 0 1 1 ...
```

```
data = data[,-9]
head(data)
```

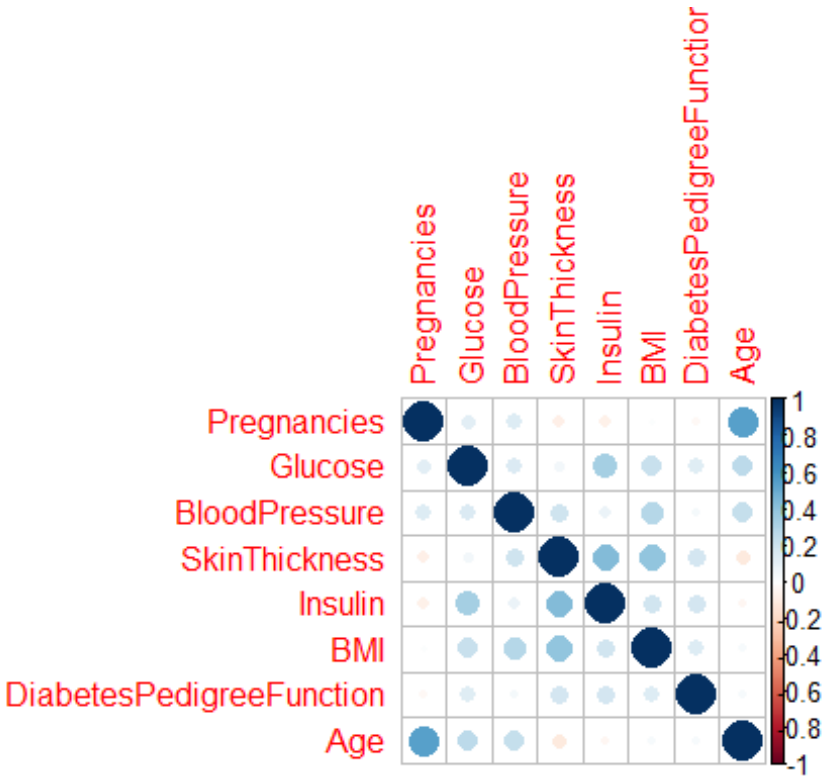
```
##   Pregnancies Glucose BloodPressure SkinThickness Insulin  BMI
## 1           6     148            72            35       0 33.6
## 2           1      85            66            29       0 26.6
## 3           8     183            64             0       0 23.3
## 4           1      89            66            23      94 28.1
## 5           0     137            40            35     168 43.1
## 6           5     116            74             0       0 25.6
##   DiabetesPedigreeFunction Age
## 1                    0.627  50
## 2                    0.351  31
## 3                    0.672  32
## 4                    0.167  21
## 5                    2.288  33
## 6                    0.201  30
```

```
library(corrplot)
```

```
## corrplot 0.92 loaded
```

```
corr_mat =cor(data)
```

```
corrplot(corr_mat)
```




```
library(psych)
```

```
## Warning: package 'psych' was built under R version 4.3.3
```

```
cortest.bartlett(R=corr_mat, n=768)
```

```
## $chisq
## [1] 948.2262
##
## $p.value
## [1] 1.25755e-181
##
## $df
## [1] 28
```

```
eigen_vals = eigen(corr_mat)$values
eigen_vals
```

```
## [1] 2.0943799 1.7312101 1.0296299 0.8755290 0.7623444 0.6826284 0.4198162
## [8] 0.4044620
```

```
 print(sum(eigen_vals))
```

```
## [1] 8
```

```
print(cumsum(eigen_vals))
```

```
## [1] 2.094380 3.825590 4.855220 5.730749 6.493093 7.175722 7.595538 8.00000
0
```

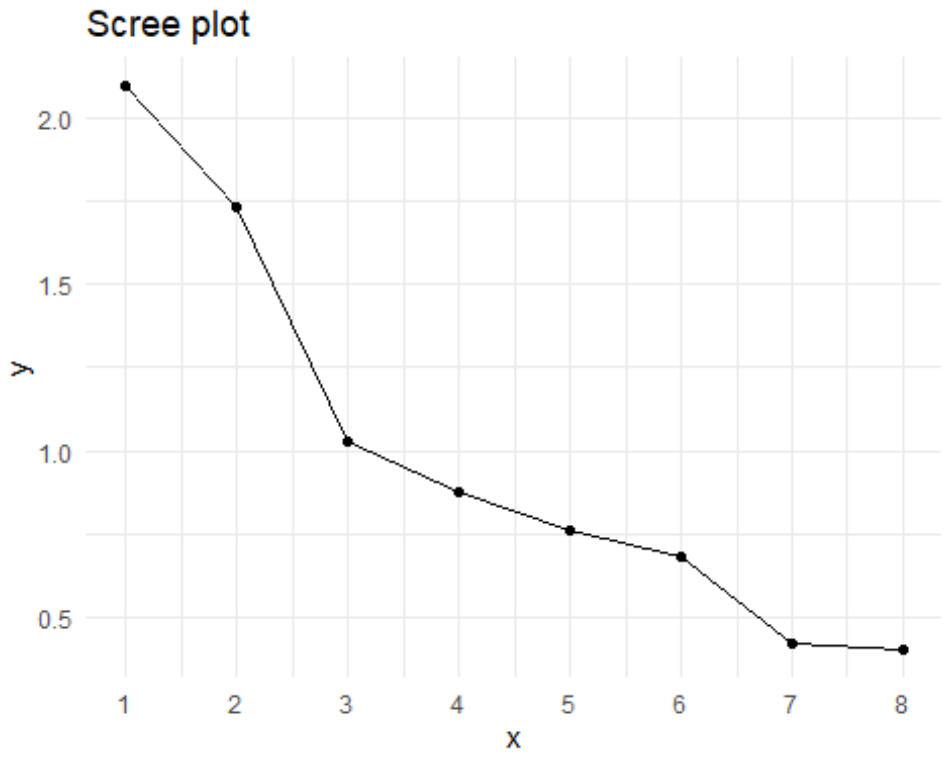
```r
print(cumsum(eigen_vals)/8)
```

```
## [1] 0.2617975 0.4781988 0.6069025 0.7163436 0.8116367 0.8969652 0.9494422
## [8] 1.0000000
```

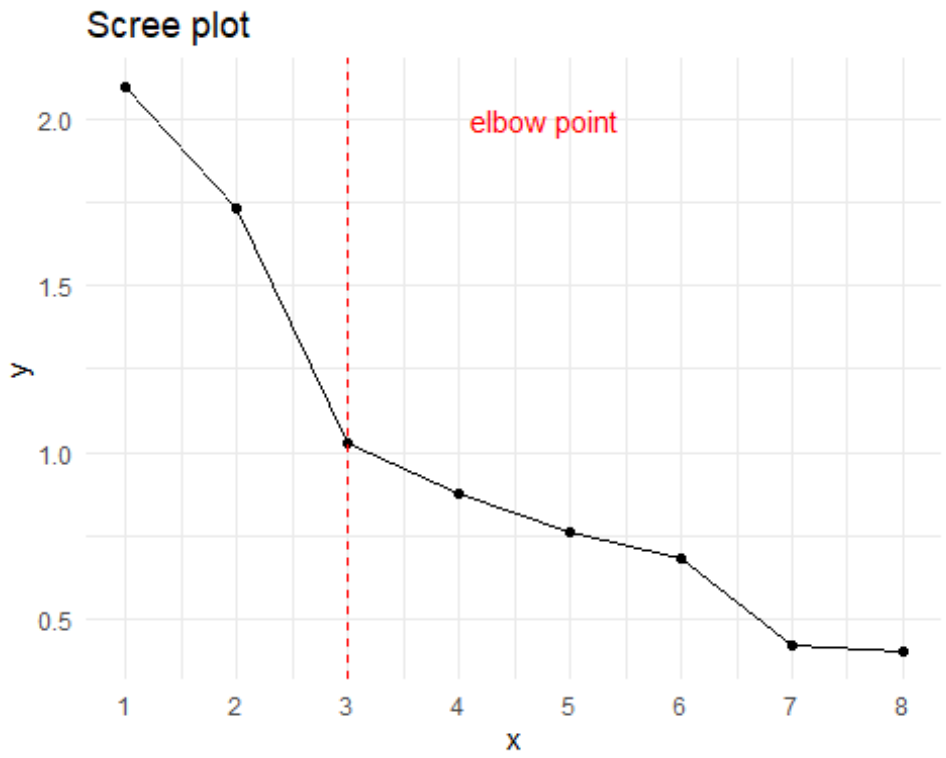
```r
library(tidyverse)
```

```
## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.
0.0 ──
## ✔ dplyr     1.1.2     ✔ readr     2.1.4
## ✔ forcats   1.0.0     ✔ stringr   1.5.0
## ✔ ggplot2   3.4.2     ✔ tibble    3.2.1
## ✔ lubridate 1.9.2     ✔ tidyr     1.3.0
## ✔ purrr     1.0.1
## ── Conflicts ────────────────────────────────────────── tidyverse_conflict
s() ──
## ✖ ggplot2::%+%()   masks psych::%+%()
## ✖ ggplot2::alpha() masks psych::alpha()
## ✖ dplyr::filter()  masks stats::filter()
## ✖ dplyr::lag()     masks stats::lag()
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force a
ll conflicts to become errors
```

```r
library(ggplot2)

n <- dim(corr_mat)[1]

scree_tbl <- tibble(x = 1:n, y = sort(eigen(corr_mat)$value, decreasing = TRU
E))

scree_plot <- scree_tbl |>
  ggplot(aes(x, y)) +
  geom_point() +
  geom_line() +
  theme_minimal() +
  scale_x_continuous(breaks = 1:n) +
  ggtitle("Scree plot")

scree_plot
```




```
scree_plot + geom_vline(xintercept = 3, color = "red", linetype = "dashed") +
annotate("text", 4.1, 2, label = "elbow point", color = "red", hjust = 0)
```



```
fa.res<-factanal(x=data, factors=3, rotation='none', method = "pc", n.obs = 3
29)

fa.res

##
## Call:
## factanal(x = data, factors = 3, n.obs = 329, rotation = "none",     method
= "pc")
##
## Uniquenesses:
##              Pregnancies                  Glucose            BloodPressure
##                    0.614                    0.005                    0.845
##            SkinThickness                  Insulin                      BMI
##                    0.188                    0.674                    0.767
## DiabetesPedigreeFunction                      Age
##                    0.939                    0.230
##
## Loadings:
##                          Factor1 Factor2 Factor3
## Pregnancies              -0.298   0.133   0.529
## Glucose                           0.997
## BloodPressure             0.125   0.155   0.339
## SkinThickness             0.849           0.297
## Insulin                   0.459   0.333
## BMI                       0.379   0.223   0.199
## DiabetesPedigreeFunction  0.187   0.138
## Age                      -0.408   0.268   0.729
##
##                Factor1 Factor2 Factor3
## SS loadings      1.381   1.292   1.065
## Proportion Var   0.173   0.161   0.133
## Cumulative Var   0.173   0.334   0.467
##
## Test of the hypothesis that 3 factors are sufficient.
## The chi square statistic is 39 on 7 degrees of freedom.
## The p-value is 1.95e-06

fa.res<-factanal(x=data, factors=3, rotation='varimax')

print(fa.res, cut=0.4)

##
## Call:
## factanal(x = data, factors = 3, rotation = "varimax")
##
## Uniquenesses:
##              Pregnancies                  Glucose            BloodPressure
##                    0.614                    0.005                    0.845
##            SkinThickness                  Insulin                      BMI
```

```
##                     0.188                     0.674                     0.767
## DiabetesPedigreeFunction                       Age
##                     0.939                     0.230
##
## Loadings:
##                           Factor1 Factor2 Factor3
## Pregnancies                        0.619
## Glucose                                    0.953
## BloodPressure
## SkinThickness              0.886
## Insulin                    0.515
## BMI                        0.467
## DiabetesPedigreeFunction
## Age                                0.869
##
##                Factor1 Factor2 Factor3
## SS loadings      1.455   1.263   1.020
## Proportion Var   0.182   0.158   0.128
## Cumulative Var   0.182   0.340   0.467
##
## Test of the hypothesis that 3 factors are sufficient.
## The chi square statistic is 39 on 7 degrees of freedom.
## The p-value is 1.95e-06
```

```
fa.res<-factanal(x=data, factors=3, rotation='varimax', scores='Bartlett')

head(fa.res$scores)
```

```
##          Factor1    Factor2    Factor3
## [1,]  0.95683200  1.5192617  0.3677035
## [2,]  0.19649075 -0.2537621 -1.1804957
## [3,] -1.24427322 -0.1154597  2.3615440
## [4,] -0.08141387 -1.1273334 -0.8159980
## [5,]  1.27414283 -0.4216552  0.2993244
## [6,] -1.43085723 -0.2048091  0.2277392
```

```
library(lavaan)

## Warning: package 'lavaan' was built under R version 4.3.3

## This is lavaan 0.6-17
## lavaan is FREE software! Please report any bugs.

##
## Attaching package: 'lavaan'

## The following object is masked from 'package:psych':
##
##     cor2cov

# Define CFA model
cfa.model <- '
```

```
  metabolic_health =~ SkinThickness + Insulin + BMI
  reproductive_health =~ Pregnancies + Age
  glucose_metabolism =~ Glucose
'

# Fit the CFA model
data <- scale(data)
cfa.est <- cfa(cfa.model, data = data)

summary(cfa.est, fit = TRUE)

## lavaan 0.6.17 ended normally after 32 iterations
##
##   Estimator                                         ML
##   Optimization method                           NLMINB
##   Number of model parameters                        14
##
##   Number of observations                           768
##
## Model Test User Model:
##
##   Test statistic                               118.142
##   Degrees of freedom                                 7
##   P-value (Chi-square)                           0.000
##
## Model Test Baseline Model:
##
##   Test statistic                               776.259
##   Degrees of freedom                                15
##   P-value                                        0.000
##
## User Model versus Baseline Model:
##
##   Comparative Fit Index (CFI)                    0.854
##   Tucker-Lewis Index (TLI)                       0.687
##
## Loglikelihood and Information Criteria:
##
##   Loglikelihood user model (H0)              -6206.409
##   Loglikelihood unrestricted model (H1)      -6147.337
##
##   Akaike (AIC)                               12440.817
##   Bayesian (BIC)                             12505.830
##   Sample-size adjusted Bayesian (SABIC)      12461.374
##
## Root Mean Square Error of Approximation:
##
##   RMSEA                                          0.144
##   90 Percent confidence interval - lower         0.122
##   90 Percent confidence interval - upper         0.167
```

```
##   P-value H_0: RMSEA <= 0.050                    0.000
##   P-value H_0: RMSEA >= 0.080                    1.000
##
## Standardized Root Mean Square Residual:
##
##   SRMR                                           0.061
##
## Parameter Estimates:
##
##   Standard errors                             Standard
##   Information                                 Expected
##   Information saturated (h1) model          Structured
##
## Latent Variables:
##                          Estimate  Std.Err  z-value  P(>|z|)
##   metabolic_health =~
##     SkinThickness           1.000
##     Insulin                 0.932    0.100    9.326    0.000
##     BMI                     0.693    0.079    8.736    0.000
##   reproductive_health =~
##     Pregnancies             1.000
##     Age                     1.799    0.327    5.505    0.000
##   glucose_metabolism =~
##     Glucose                 1.000
##
## Covariances:
##                          Estimate  Std.Err  z-value  P(>|z|)
##   metabolic_health ~~
##     reprodctv_hlth         -0.035    0.018   -1.918    0.055
##     glucose_mtblsm          0.208    0.033    6.284    0.000
##   reproductive_health ~~
##     glucose_mtblsm          0.146    0.033    4.365    0.000
##
## Variances:
##                    Estimate  Std.Err  z-value  P(>|z|)
##    .SkinThickness     0.537    0.054    9.985    0.000
##    .Insulin           0.598    0.051   11.774    0.000
##    .BMI               0.777    0.047   16.585    0.000
##    .Pregnancies       0.697    0.063   11.007    0.000
##    .Age               0.021    0.170    0.121    0.904
##    .Glucose           0.000
##     metabolic_hlth    0.461    0.063    7.303    0.000
##     reprodctv_hlth    0.302    0.064    4.733    0.000
##     glucose_mtblsm    0.999    0.051   19.596    0.000
```

```r
cfa.est<-cfa(cfa.model, data=data, std.lv=TRUE)

summary(cfa.est, fit=TRUE)
```

```
## lavaan 0.6.17 ended normally after 23 iterations
##
##   Estimator                                         ML
##   Optimization method                           NLMINB
##   Number of model parameters                        14
##
##   Number of observations                           768
##
## Model Test User Model:
##
##   Test statistic                               118.142
##   Degrees of freedom                                 7
##   P-value (Chi-square)                           0.000
##
## Model Test Baseline Model:
##
##   Test statistic                               776.259
##   Degrees of freedom                                15
##   P-value                                        0.000
##
## User Model versus Baseline Model:
##
##   Comparative Fit Index (CFI)                    0.854
##   Tucker-Lewis Index (TLI)                       0.687
##
## Loglikelihood and Information Criteria:
##
##   Loglikelihood user model (H0)              -6206.409
##   Loglikelihood unrestricted model (H1)      -6147.337
##
##   Akaike (AIC)                               12440.817
##   Bayesian (BIC)                             12505.830
##   Sample-size adjusted Bayesian (SABIC)      12461.374
##
## Root Mean Square Error of Approximation:
##
##   RMSEA                                          0.144
##   90 Percent confidence interval - lower         0.122
##   90 Percent confidence interval - upper         0.167
##   P-value H_0: RMSEA <= 0.050                    0.000
##   P-value H_0: RMSEA >= 0.080                    1.000
##
## Standardized Root Mean Square Residual:
##
##   SRMR                                           0.061
##
## Parameter Estimates:
##
##   Standard errors                             Standard
##   Information                                 Expected
```

```
##   Information saturated (h1) model          Structured
##
## Latent Variables:
##                          Estimate  Std.Err  z-value  P(>|z|)
##   metabolic_health =~
##     SkinThickness           0.679    0.046   14.606    0.000
##     Insulin                 0.633    0.045   13.910    0.000
##     BMI                     0.470    0.043   10.907    0.000
##   reproductive_health =~
##     Pregnancies             0.550    0.058    9.466    0.000
##     Age                     0.989    0.089   11.053    0.000
##   glucose_metabolism =~
##     Glucose                 0.999    0.025   39.192    0.000
##
## Covariances:
##                          Estimate  Std.Err  z-value  P(>|z|)
##   metabolic_health ~~
##     reprodctv_hlth         -0.094    0.046   -2.042    0.041
##     glucose_mtblsm          0.307    0.042    7.333    0.000
##   reproductive_health ~~
##     glucose_mtblsm          0.266    0.041    6.525    0.000
##
## Variances:
##                    Estimate  Std.Err  z-value  P(>|z|)
##    .SkinThickness     0.537    0.054    9.985    0.000
##    .Insulin           0.598    0.051   11.774    0.000
##    .BMI               0.777    0.047   16.585    0.000
##    .Pregnancies       0.697    0.063   11.007    0.000
##    .Age               0.021    0.169    0.121    0.904
##    .Glucose           0.000
##     metabolic_hlth    1.000
##     reprodctv_hlth    1.000
##     glucose_mtblsm    1.000
```

```
# Define CFA model
cfa.model <- '
  metabolic_health =~ SkinThickness + Insulin + BMI
  reproductive_health =~ Pregnancies + Age
  glucose_metabolism =~ Glucose
'

cfa.est <- cfa(cfa.model, data = data)

summary(cfa.est, fit = TRUE)
```

```
## lavaan 0.6.17 ended normally after 32 iterations
##
##   Estimator                                         ML
##   Optimization method                           NLMINB
##   Number of model parameters                        14
##
```

```
##   Number of observations                           768
##
## Model Test User Model:
##
##   Test statistic                                118.142
##   Degrees of freedom                                  7
##   P-value (Chi-square)                            0.000
##
## Model Test Baseline Model:
##
##   Test statistic                                776.259
##   Degrees of freedom                                 15
##   P-value                                         0.000
##
## User Model versus Baseline Model:
##
##   Comparative Fit Index (CFI)                     0.854
##   Tucker-Lewis Index (TLI)                        0.687
##
## Loglikelihood and Information Criteria:
##
##   Loglikelihood user model (H0)              -6206.409
##   Loglikelihood unrestricted model (H1)      -6147.337
##
##   Akaike (AIC)                               12440.817
##   Bayesian (BIC)                             12505.830
##   Sample-size adjusted Bayesian (SABIC)      12461.374
##
## Root Mean Square Error of Approximation:
##
##   RMSEA                                           0.144
##   90 Percent confidence interval - lower          0.122
##   90 Percent confidence interval - upper          0.167
##   P-value H_0: RMSEA <= 0.050                     0.000
##   P-value H_0: RMSEA >= 0.080                     1.000
##
## Standardized Root Mean Square Residual:
##
##   SRMR                                            0.061
##
## Parameter Estimates:
##
##   Standard errors                             Standard
##   Information                                 Expected
##   Information saturated (h1) model          Structured
##
## Latent Variables:
##                          Estimate  Std.Err  z-value  P(>|z|)
##   metabolic_health =~
##     SkinThickness           1.000
```

```
##     Insulin                 0.932    0.100    9.326    0.000
##     BMI                     0.693    0.079    8.736    0.000
##   reproductive_health =~
##     Pregnancies             1.000
##     Age                     1.799    0.327    5.505    0.000
##   glucose_metabolism =~
##     Glucose                 1.000
##
## Covariances:
##                          Estimate  Std.Err  z-value  P(>|z|)
##   metabolic_health ~~
##     reprodctv_hlth         -0.035    0.018   -1.918    0.055
##     glucose_mtblsm          0.208    0.033    6.284    0.000
##   reproductive_health ~~
##     glucose_mtblsm          0.146    0.033    4.365    0.000
##
## Variances:
##                    Estimate  Std.Err  z-value  P(>|z|)
##    .SkinThickness     0.537    0.054    9.985    0.000
##    .Insulin           0.598    0.051   11.774    0.000
##    .BMI               0.777    0.047   16.585    0.000
##    .Pregnancies       0.697    0.063   11.007    0.000
##    .Age               0.021    0.170    0.121    0.904
##    .Glucose           0.000
##     metabolic_hlth    0.461    0.063    7.303    0.000
##     reprodctv_hlth    0.302    0.064    4.733    0.000
##     glucose_mtblsm    0.999    0.051   19.596    0.000
```

```
# Define the modified CFA model
cfa.model.mod1 <- '
  metabolic_health =~ SkinThickness + Insulin + BMI
  reproductive_health =~ Pregnancies + Age
  glucose_metabolism =~ Glucose
  metabolic_health ~~ reproductive_health
'

# Fit the modified CFA model
cfa.est.mod1 <- cfa(cfa.model.mod1, data = data)

summary(cfa.est.mod1, fit = TRUE)
```

```
## lavaan 0.6.17 ended normally after 32 iterations
##
##   Estimator                                         ML
##   Optimization method                           NLMINB
##   Number of model parameters                        14
##
##   Number of observations                           768
##
## Model Test User Model:
```

```
##
##   Test statistic                                118.142
##   Degrees of freedom                                  7
##   P-value (Chi-square)                            0.000
##
## Model Test Baseline Model:
##
##   Test statistic                                776.259
##   Degrees of freedom                                 15
##   P-value                                         0.000
##
## User Model versus Baseline Model:
##
##   Comparative Fit Index (CFI)                     0.854
##   Tucker-Lewis Index (TLI)                        0.687
##
## Loglikelihood and Information Criteria:
##
##   Loglikelihood user model (H0)              -6206.409
##   Loglikelihood unrestricted model (H1)      -6147.337
##
##   Akaike (AIC)                               12440.817
##   Bayesian (BIC)                             12505.830
##   Sample-size adjusted Bayesian (SABIC)      12461.374
##
## Root Mean Square Error of Approximation:
##
##   RMSEA                                          0.144
##   90 Percent confidence interval - lower         0.122
##   90 Percent confidence interval - upper         0.167
##   P-value H_0: RMSEA <= 0.050                    0.000
##   P-value H_0: RMSEA >= 0.080                    1.000
##
## Standardized Root Mean Square Residual:
##
##   SRMR                                           0.061
##
## Parameter Estimates:
##
##   Standard errors                             Standard
##   Information                                 Expected
##   Information saturated (h1) model          Structured
##
## Latent Variables:
##                          Estimate  Std.Err  z-value  P(>|z|)
##   metabolic_health =~
##     SkinThickness           1.000
##     Insulin                 0.932    0.100    9.326    0.000
##     BMI                     0.693    0.079    8.736    0.000
##   reproductive_health =~
```

```
##     Pregnancies             1.000
##     Age                     1.799    0.327    5.505    0.000
##   glucose_metabolism =~
##     Glucose                 1.000
##
## Covariances:
##                          Estimate  Std.Err  z-value  P(>|z|)
##   metabolic_health ~~
##     reprodctv_hlth         -0.035    0.018   -1.918    0.055
##     glucose_mtblsm          0.208    0.033    6.284    0.000
##   reproductive_health ~~
##     glucose_mtblsm          0.146    0.033    4.365    0.000
##
## Variances:
##                    Estimate  Std.Err  z-value  P(>|z|)
##    .SkinThickness     0.537    0.054    9.985    0.000
##    .Insulin           0.598    0.051   11.774    0.000
##    .BMI               0.777    0.047   16.585    0.000
##    .Pregnancies       0.697    0.063   11.007    0.000
##    .Age               0.021    0.170    0.121    0.904
##    .Glucose           0.000
##     metabolic_hlth    0.461    0.063    7.303    0.000
##     reprodctv_hlth    0.302    0.064    4.733    0.000
##     glucose_mtblsm    0.999    0.051   19.596    0.000
```

```
# Define the modified CFA model
cfa.model.mod2 <- '
  metabolic_health =~ SkinThickness + Insulin + BMI
  reproductive_health =~ Pregnancies + Age + BMI
  glucose_metabolism =~ Glucose
'

# Fit the modified CFA model
cfa.est.mod2 <- cfa(cfa.model.mod2, data = data)

summary(cfa.est.mod2, fit = TRUE)
```

```
## lavaan 0.6.17 ended normally after 33 iterations
##
##   Estimator                                         ML
##   Optimization method                           NLMINB
##   Number of model parameters                        15
##
##   Number of observations                           768
##
## Model Test User Model:
##
##   Test statistic                               109.897
##   Degrees of freedom                                 6
##   P-value (Chi-square)                           0.000
```

```
##
## Model Test Baseline Model:
##
##   Test statistic                                776.259
##   Degrees of freedom                                 15
##   P-value                                         0.000
##
## User Model versus Baseline Model:
##
##   Comparative Fit Index (CFI)                     0.864
##   Tucker-Lewis Index (TLI)                        0.659
##
## Loglikelihood and Information Criteria:
##
##   Loglikelihood user model (H0)              -6202.286
##   Loglikelihood unrestricted model (H1)      -6147.337
##
##   Akaike (AIC)                               12434.572
##   Bayesian (BIC)                             12504.228
##   Sample-size adjusted Bayesian (SABIC)      12456.597
##
## Root Mean Square Error of Approximation:
##
##   RMSEA                                          0.150
##   90 Percent confidence interval - lower         0.126
##   90 Percent confidence interval - upper         0.175
##   P-value H_0: RMSEA <= 0.050                    0.000
##   P-value H_0: RMSEA >= 0.080                    1.000
##
## Standardized Root Mean Square Residual:
##
##   SRMR                                           0.055
##
## Parameter Estimates:
##
##   Standard errors                             Standard
##   Information                                 Expected
##   Information saturated (h1) model          Structured
##
## Latent Variables:
##                          Estimate  Std.Err  z-value  P(>|z|)
##   metabolic_health =~
##     SkinThickness           1.000
##     Insulin                 0.876    0.092    9.499    0.000
##     BMI                     0.714    0.080    8.934    0.000
##   reproductive_health =~
##     Pregnancies             1.000
##     Age                     1.659    0.264    6.281    0.000
##     BMI                     0.190    0.065    2.910    0.004
##   glucose_metabolism =~
```

```
##     Glucose                  1.000
##
## Covariances:
##                          Estimate  Std.Err  z-value  P(>|z|)
##   metabolic_health ~~
##     reprodctv_hlth         -0.055    0.021   -2.586    0.010
##     glucose_mtblsm          0.196    0.033    5.876    0.000
##   reproductive_health ~~
##     glucose_mtblsm          0.158    0.033    4.821    0.000
##
## Variances:
##                    Estimate  Std.Err  z-value  P(>|z|)
##    .SkinThickness     0.512    0.054    9.454    0.000
##    .Insulin           0.625    0.049   12.770    0.000
##    .BMI               0.754    0.047   16.114    0.000
##    .Pregnancies       0.671    0.060   11.154    0.000
##    .Age               0.097    0.136    0.714    0.475
##    .Glucose           0.000
##     metabolic_hlth    0.486    0.065    7.534    0.000
##     reprodctv_hlth    0.327    0.062    5.262    0.000
##     glucose_mtblsm    0.999    0.051   19.596    0.000
```