What are the key latent variables able to describe our cities?
A case study based on 92 Italian Cities measured by Eurostat.

Tommaso Malaguti

1. Introduction

The aim of this study is to answer the research question “What are the key latent variables able to describe our cities” through the use of Factor Analysis.

In this study, we utilize factor analysis to explore the key latent variables that describe our cities, using data from 92 Italian cities measured by Eurostat. In our opinion the best path to finding the solution is by applying Factor analysis to this dataset; we aim to uncover the underlying dimensions that contribute to the variability in urban characteristics and provide insights into the factors shaping our cities.

1.1 Key concepts

Factor analysis is a statistical method used in multivariate analysis to identify underlying factors or latent variables that explain patterns of correlations among observed variables. The goal of factor analysis is to reduce a large number of variables to a smaller and more manageable set of new informative variables called factors, these capture the essential information present in the original variables and are useful for understanding the underlying structure or dimensions of the data.
So, basically it is the practice of condensing many variables into just a few, so that your research data is easier to work with.

Factors are useful to us because they are uncorrelated with each other, and they exhibit certain key properties that help us in explaining the variance and relationships among observed variables. They are latent variables so are not directly observed but are inferred from the patterns of covariance among observed variables and represent the underlying constructs or dimensions that contribute to the observed variability in the data.

Central to factor analysis is variance, it is used to quantify the degree to which numerical values deviate from the mean. Essentially, the goal of factor analysis is to determine how underlying factors affect the variation between our variables. Since they more closely reflect the variables, they are made of, certain factors will explain more variance than others.
Two other key concepts to know are Eigenvalues and Factor loadings.

- The eigenvalue expresses the amount of variance a factor explains. If a factor solution (unobserved or latent variables) has an eigenvalue of 1 or above, it indicates that a factor explains more variance than a single observed variable, which can be useful in reducing the number of variables in your analysis. Factors with eigenvalues less than 1 account for less variability than a single variable and are generally not included in the analysis.

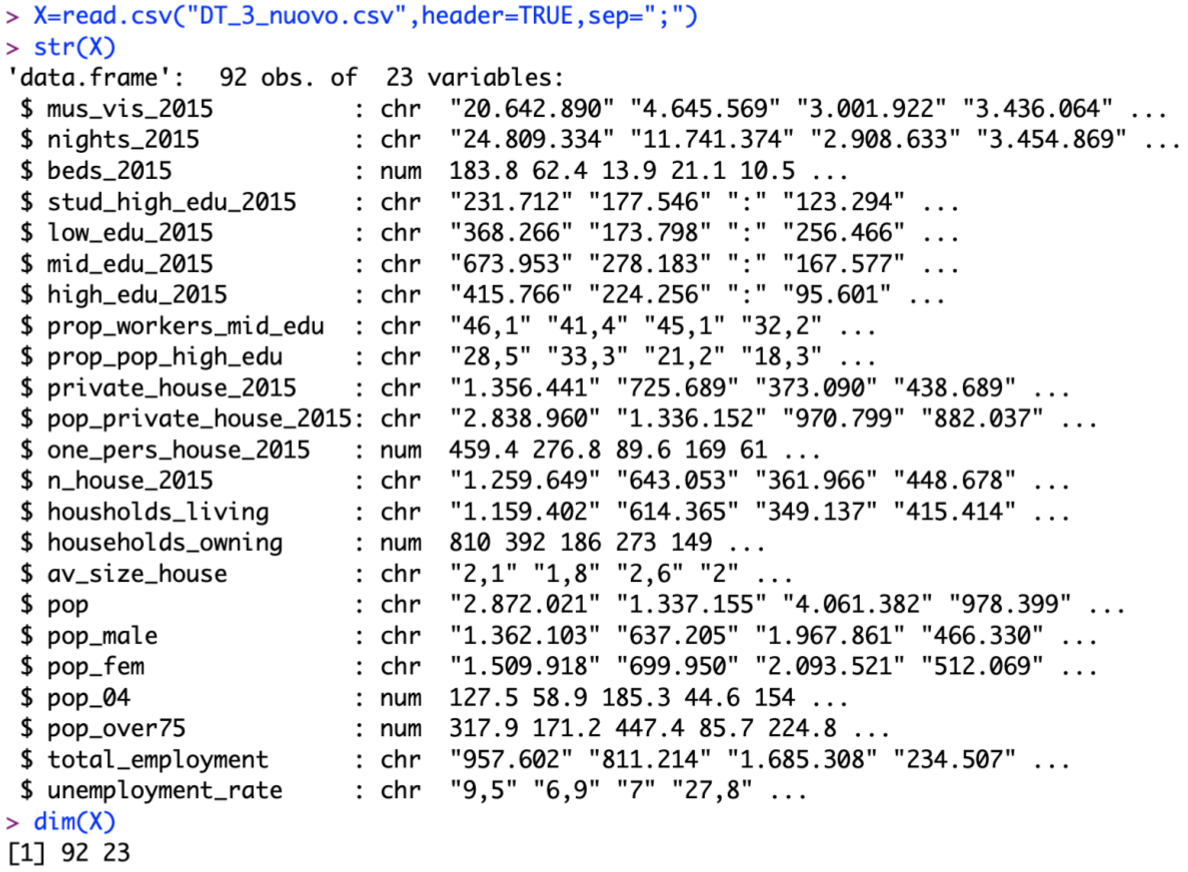
- Factor loading is the correlation coefficient for the variable and factor. Like the factor score, factor loadings give an indication of how much of the variance in an observed variable can be explained by the factor. High factor loadings (close to 1 or -1) mean the factor strongly influences the variable.

Factor analysis often produces factors that are difficult to interpret or don't align well with theoretical expectations. Factor rotation is used to improve the interpretability of the factors by adjusting their orientation in the multidimensional space. So, it’s useful to identify underlying dimensions or factors that explain patterns of correlations among observed variables.
Different rotations can be used; one of the most used is Varimax, which seeks to maximize the variance of squared loadings within each factor. We’ll have a factor structure where each variable has a high loading on one factor and a low loading on the others. Because of their increased orthogonality (uncorrelatedness), factors may be interpreted more easily after this rotation. After having done Varimax we can also use the Promax rotation which is particularly useful when there is reason to believe that factors are correlated. These two maximize variance but we can also go another route and minimize the complexity of the factor structure instead, and so allowing for correlations between factors, with Oblimin when a factor's correlation is suggested by empirical or theoretical data. Or we can do none of that and leave the factor structure as it was derived from the initial extraction.

All these notions are fundamental for our work and will be used to reach the goal, answering our research question.

2. Dataset definition

Every statistical research starts with the construction of the dataset, and so does ours.
The given file **“DT_3.xlsx”** is formed by 96 Italian cities and 23 variables that describe several factors. From this dataset, that embodies data like population, house market or tourism, we start our work.



```
> X=read.csv("DT_3_nuovo.csv",header=TRUE,sep=";")
> str(X)
'data.frame':	92 obs. of  23 variables:
 $ mus_vis_2015          : chr  "20.642.890" "4.645.569" "3.001.922" "3.436.064" ...
 $ nights_2015           : chr  "24.809.334" "11.741.374" "2.908.633" "3.454.869" ...
 $ beds_2015             : num  183.8 62.4 13.9 21.1 10.5 ...
 $ stud_high_edu_2015    : chr  "231.712" "177.546" ":" "123.294" ...
 $ low_edu_2015          : chr  "368.266" "173.798" ":" "256.466" ...
 $ mid_edu_2015          : chr  "673.953" "278.183" ":" "167.577" ...
 $ high_edu_2015         : chr  "415.766" "224.256" ":" "95.601" ...
 $ prop_workers_mid_edu  : chr  "46,1" "41,4" "45,1" "32,2" ...
 $ prop_pop_high_edu     : chr  "28,5" "33,3" "21,2" "18,3" ...
 $ private_house_2015    : chr  "1.356.441" "725.689" "373.090" "438.689" ...
 $ pop_private_house_2015: chr  "2.838.960" "1.336.152" "970.799" "882.037" ...
 $ one_pers_house_2015   : num  459.4 276.8 89.6 169 61 ...
 $ n_house_2015          : chr  "1.259.649" "643.053" "361.966" "448.678" ...
 $ housholds_living      : chr  "1.159.402" "614.365" "349.137" "415.414" ...
 $ households_owning     : num  810 392 186 273 149 ...
 $ av_size_house         : chr  "2,1" "1,8" "2,6" "2" ...
 $ pop                   : chr  "2.872.021" "1.337.155" "4.061.382" "978.399" ...
 $ pop_male              : chr  "1.362.103" "637.205" "1.967.861" "466.330" ...
 $ pop_fem               : chr  "1.509.918" "699.950" "2.093.521" "512.069" ...
 $ pop_04                : num  127.5 58.9 185.3 44.6 154 ...
 $ pop_over75            : num  317.9 171.2 447.4 85.7 224.8 ...
 $ total_employment      : chr  "957.602" "811.214" "1.685.308" "234.507" ...
 $ unemployment_rate     : chr  "9,5" "6,9" "7" "27,8" ...
> dim(X)
[1] 92 23
```

Fig.1 : Dataset before data manipulation

2.1 Data manipulation and standardization

First of all, we changed the name of each variable in order to become more readable. Then we transformed our dataset to make it suitable for factor analysis and for the program “R”, so we start by transforming all our variables from “character” to “numeric” using the function:

X_numeric <- data.frame(apply(X, 2, function(x) as.numeric(gsub(",", ".", gsub("\\.", "", x)))))

Next, even if this dataset is made by institutions like Eurostat as in our case, we came across some missing variables, so we checked for the presence of those in the dataset, since we cannot proceed with the analysis if present. Seeing that missing values are present in 5 cities, we considered two methods in order to tackle this problem: omitting variables with missing values, certainly the easiest and the first method we tried during our analysis; or filling the missing spaces with values forecasted by the interpolation of the variables.

In the first method (which we ended up not using), the missing values in the cities were firstly highlighted by the code **any_missing <- any(is.na(X_numeric)).** And as far as the result "R" gave back was "TRUE", we used the code **X_good <- na.omit(X_numeric)** to omit the cities with the missing values. This was our first method used during our computation and by far the easiest, but we saw that it could have been improved.

Since two major Italian cities in terms of population present missing values, that are Naples and Palermo, we decided to opt for filling the missing values using the function: **X_interpolated = na.approx(X_numeric) thanks to the library "zoo". This function replaces NA by linear interpolation.**

For the second step we standardize all the dataset using the function **scale(X_interpolated).**

Then we computed the correlation matrix shown in Fig. 2

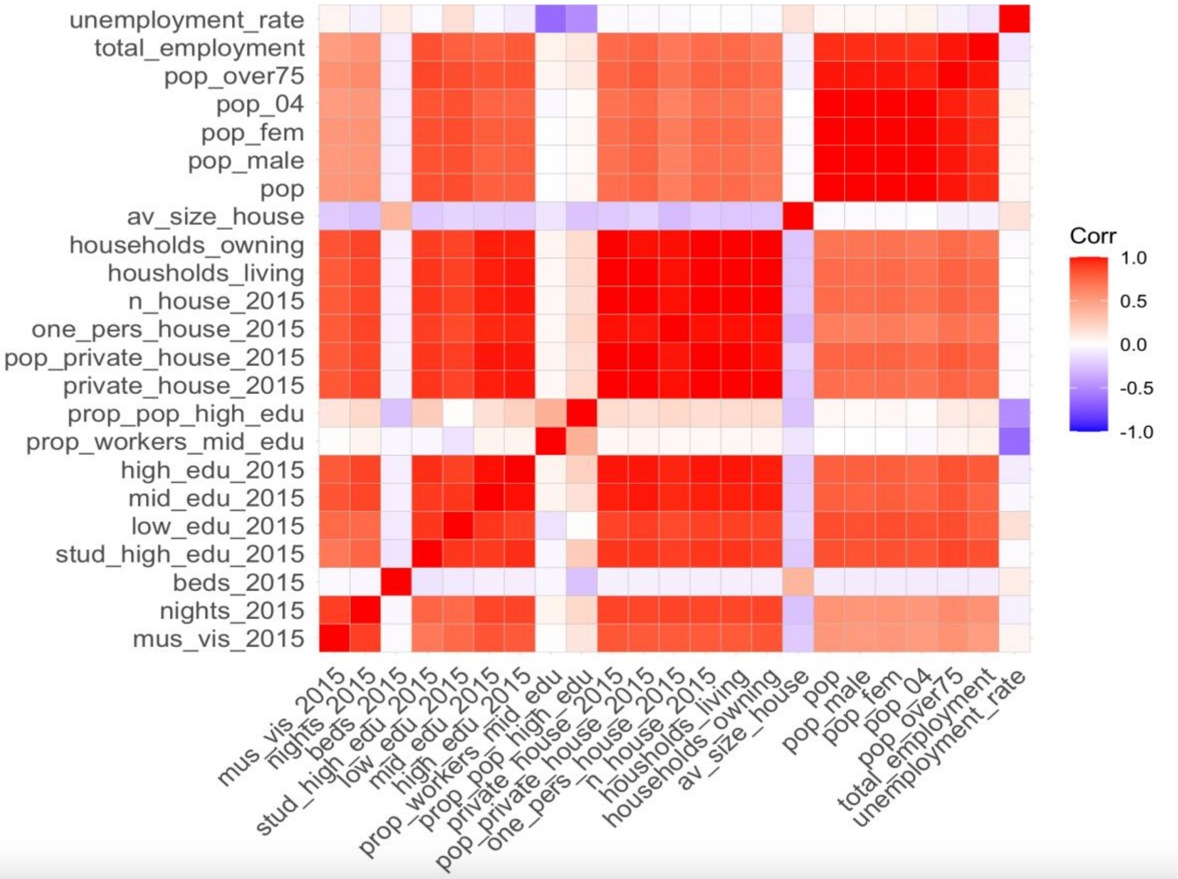


Fig.2 : Correlation matrix with all 23 variables

As seen in Fig.2 there are some variables that share high correlation and some even have a perfect linear correlation. This situation is unsuitable for Factor analysis since when attempting to compute factor analysis, the feedback given by the program is negative, so our dataset cannot be computed, because the correlation matrix is computationally singular.

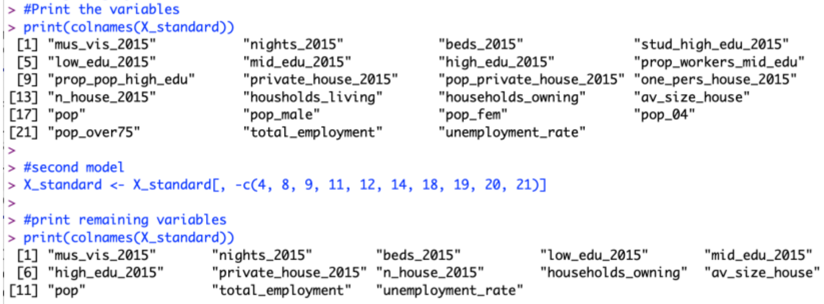
In addressing this issue, we explored two distinct approaches.

Initially, we used a code that established a correlation threshold, intended to exclude variables surpassing this threshold. However, this method presented two primary challenges. Firstly, we found it necessary to set an exceedingly high correlation threshold (0.99) to retain the most pertinent variables, given the abundance of variables exhibiting very high correlations. Secondly, the decision-making process for dropping variables was delegated to the code, resulting in arbitrary exclusions without a systematic criterion. Consequently, this approach led to the removal of variables crucial to our research objectives.

In our alternative approach that we decided to use, we exercised discretion by applying our own criteria to determine the variables worthy of retention. This involved categorizing the variables into five distinct groups: tourism, education, housing, population, and employment. Within each category, we meticulously evaluated and selected the variables deemed most pertinent to our research objectives. Through this process, we reshaped our dataset, making it more amenable to Factor Analysis. We excluded 10 variables displaying excessive correlation, unsuitable for Factor Analysis, resulting in the creation of our refined dataset named "Excell," featuring dimensions 92 by 11. During this refinement, redundant variables such as "male pop." or "female pop." were eliminated due to perfect linear correlation, as they proved redundant and irrelevant for our research goals.

To obtain this result we used the codes: **print(colnames(X_standard))** and **X_standard <- X_standard[, -c(4, 8, 9, 11, 12, 14, 18, 19, 20, 21)]**

Fig. 3 and 4 provides a visual representation of the chosen variables for each category.



```
> #manage highly correlated variables
> #Print the variables
> print(colnames(X_standard))
 [1] "mus_vis_2015"          "nights_2015"           "beds_2015"             "stud_high_edu_2015"
 [5] "low_edu_2015"          "mid_edu_2015"          "high_edu_2015"         "prop_workers_mid_edu"
 [9] "prop_pop_high_edu"     "private_house_2015"    "pop_private_house_2015" "one_pers_house_2015"
[13] "n_house_2015"          "housholds_living"      "households_owning"     "av_size_house"
[17] "pop"                   "pop_male"              "pop_fem"               "pop_04"
[21] "pop_over75"            "total_employment"      "unemployment_rate"
>
> #second model
> X_standard <- X_standard[, -c(4, 8, 9, 11, 12, 14, 18, 19, 20, 21)]
>
> #print remaining variables
> print(colnames(X_standard))
 [1] "mus_vis_2015"       "nights_2015"        "beds_2015"          "low_edu_2015"       "mid_edu_2015"
 [6] "high_edu_2015"      "private_house_2015" "n_house_2015"       "households_owning"  "av_size_house"
[11] "pop"                "total_employment"   "unemployment_rate"
```

Fig.3 Manual manipulation of highly correlated variables

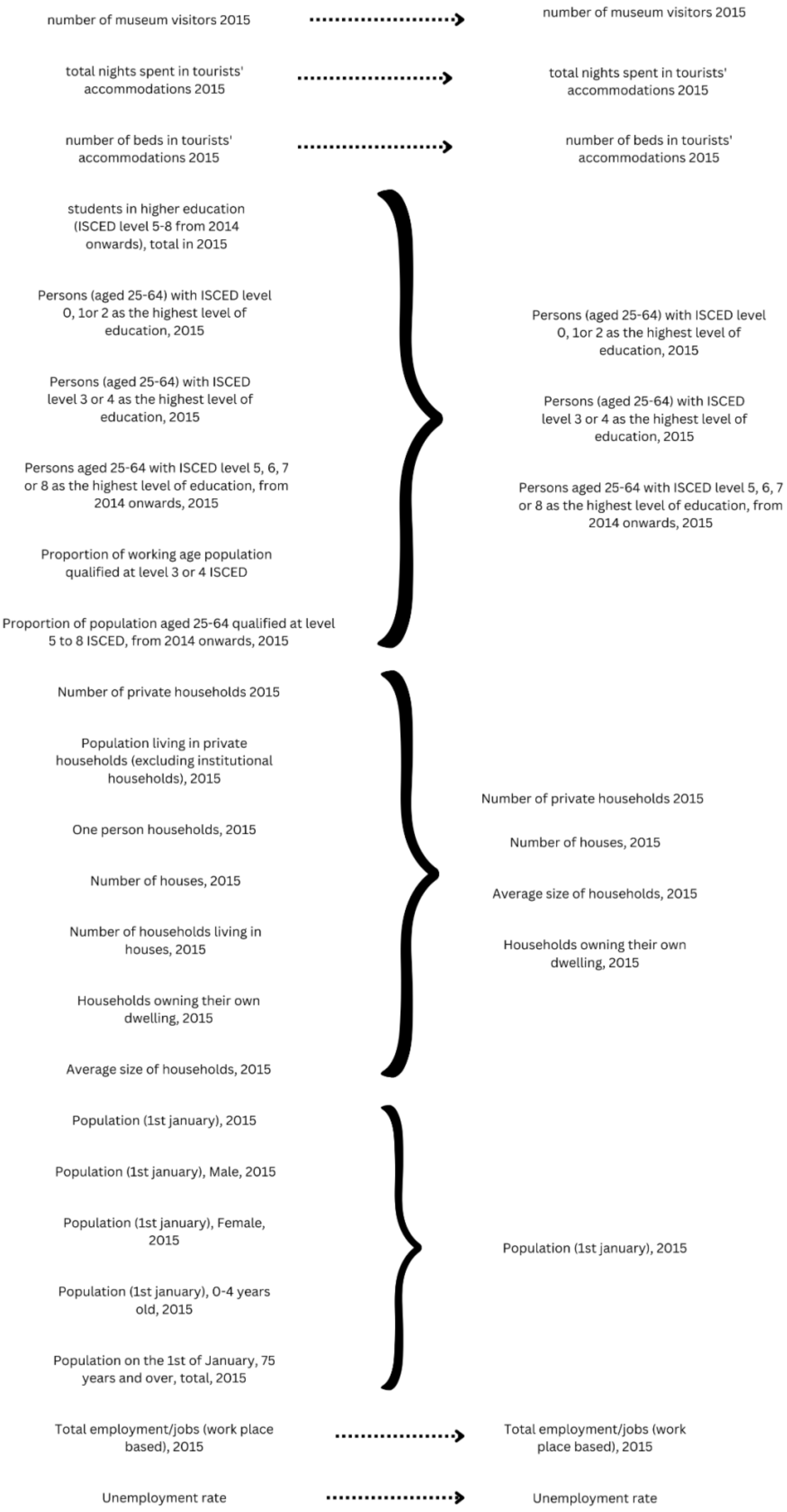
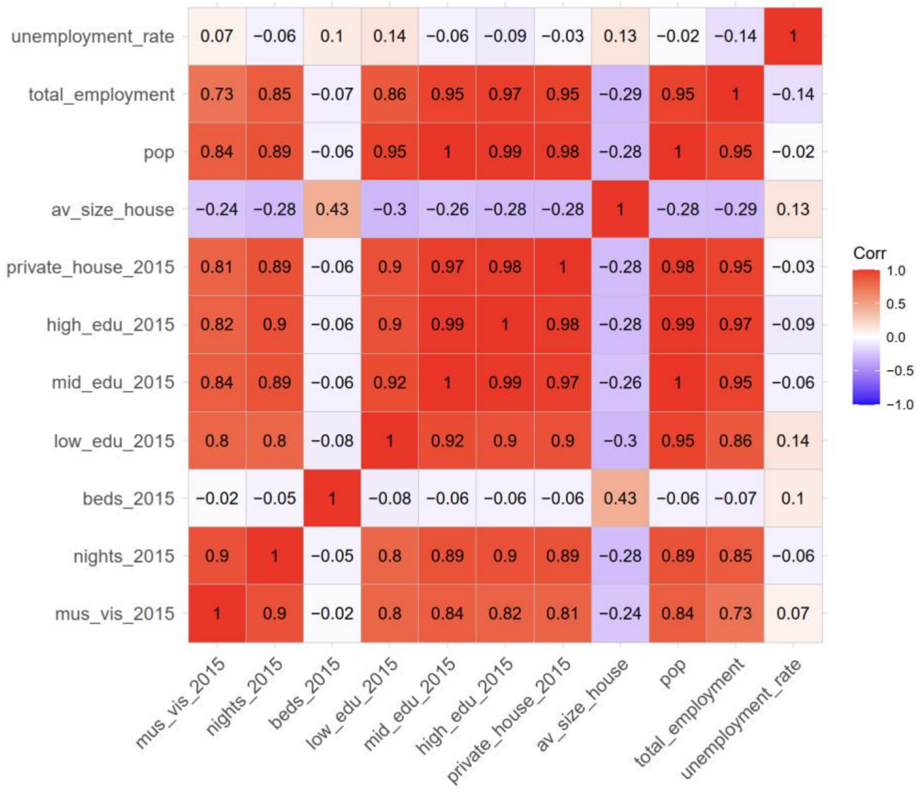

number of museum visitors 2015
number of museum visitors 2015
total nights spent in tourists' accommodations 2015
total nights spent in tourists' accommodations 2015
number of beds in tourists' accommodations 2015
number of beds in tourists' accommodations 2015
students in higher education (ISCED level 5-8 from 2014 onwards), total in 2015
Persons (aged 25-64) with ISCED level 0, 1or 2 as the highest level of education, 2015
Persons (aged 25-64) with ISCED level 3 or 4 as the highest level of education, 2015
Persons aged 25-64 with ISCED level 5, 6, 7 or 8 as the highest level of education, from 2014 onwards, 2015
Proportion of working age population qualified at level 3 or 4 ISCED
Proportion of population aged 25-64 qualified at level 5 to 8 ISCED, from 2014 onwards, 2015
Persons (aged 25-64) with ISCED level 0, 1or 2 as the highest level of education, 2015
Persons (aged 25-64) with ISCED level 3 or 4 as the highest level of education, 2015
Persons aged 25-64 with ISCED level 5, 6, 7 or 8 as the highest level of education, from 2014 onwards, 2015
Number of private households 2015
Population living in private households (excluding institutional households), 2015
One person households, 2015
Number of houses, 2015
Number of households living in houses, 2015
Households owning their own dwelling, 2015
Average size of households, 2015
Number of private households 2015
Number of houses, 2015
Average size of households, 2015
Households owning their own dwelling, 2015
Population (1st january), 2015
Population (1st january), Male, 2015
Population (1st january), Female, 2015
Population (1st january), 0-4 years old, 2015
Population on the 1st of January, 75 years and over, total, 2015
Population (1st january), 2015
Total employment/jobs (work place based), 2015
Total employment/jobs (work place based), 2015
Unemployment rate
Unemployment rate

Fig.4 : Visual representation of variables manipulation

2.2 Correlation analysis

Once we have selected the type of factor analysis, the next step involves generating the correlation matrix for our manipulated dataset. This matrix displays the correlation coefficients among each variable pair and serves as the foundation for factor extraction. This procedural stage constitutes a pivotal step in developing our factor analysis model.

A correlation matrix serves as a statistical method for assessing the association between two variables within a dataset. This matrix presents a table wherein each cell houses a correlation coefficient. A coefficient of 1 indicates a robust positive relationship between variables, 0 signifies a neutral relationship, and -1 implies a strong negative relationship. Its primary application lies in constructing regression models.



	mus_vis_2015	nights_2015	beds_2015	low_edu_2015	mid_edu_2015	high_edu_2015	private_house_2015	av_size_house	pop	total_employment	unemployment_rate
unemployment_rate	0.07	−0.06	0.1	0.14	−0.06	−0.09	−0.03	0.13	−0.02	−0.14	1
total_employment	0.73	0.85	−0.07	0.86	0.95	0.97	0.95	−0.29	0.95	1	−0.14
pop	0.84	0.89	−0.06	0.95	1	0.99	0.98	−0.28	1	0.95	−0.02
av_size_house	−0.24	−0.28	0.43	−0.3	−0.26	−0.28	−0.28	1	−0.28	−0.29	0.13
private_house_2015	0.81	0.89	−0.06	0.9	0.97	0.98	1	−0.28	0.98	0.95	−0.03
high_edu_2015	0.82	0.9	−0.06	0.9	0.99	1	0.98	−0.28	0.99	0.97	−0.09
mid_edu_2015	0.84	0.89	−0.06	0.92	1	0.99	0.97	−0.26	1	0.95	−0.06
low_edu_2015	0.8	0.8	−0.08	1	0.92	0.9	0.9	−0.3	0.95	0.86	0.14
beds_2015	−0.02	−0.05	1	−0.08	−0.06	−0.06	−0.06	0.43	−0.06	−0.07	0.1
nights_2015	0.9	1	−0.05	0.8	0.89	0.9	0.89	−0.28	0.89	0.85	−0.06
mus_vis_2015	1	0.9	−0.02	0.8	0.84	0.82	0.81	−0.24	0.84	0.73	0.07

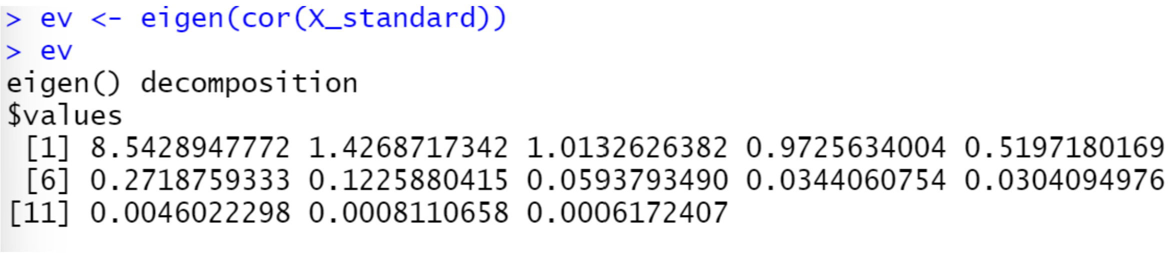
Fig.5 : Correlation matrix

For *Figure* 2 and 5 we used the libraries **(tidyverse)** and **(ggcorrplot).**

In Figure 5, it is evident that certain variables maintain a high degree of correlation. However, the noteworthy distinction lies in the fact that the correlation matrix is no longer computationally singular. This crucial transformation enables us to effectively utilize the factor analysis function.

2.3 Eigenvalues

The following step is to analyze the eigenvalues of our factor analysis, in our case we use this value to determine the number of factors to retain from the data.



```
> ev <- eigen(cor(X_standard))
> ev
eigen() decomposition
$values
 [1] 8.5428947772 1.4268717342 1.0132626382 0.9725634004 0.5197180169
 [6] 0.2718759333 0.1225880415 0.0593793490 0.0344060754 0.0304094976
[11] 0.0046022298 0.0008110658 0.0006172407
```

Fig.6 : Eigenvalues

Eigenvalues play a crucial role in factor extraction by representing the variance explained by each factor; factors are extracted sequentially, and each factor corresponds to an eigenvalue. The first factor extracted, in our case 8, is the one associated with the largest eigenvalue and this accounts for the greatest proportion of variance in the data. Subsequent factors are extracted in order of decreasing eigenvalues, meaning each subsequent factor explains a progressively smaller amount of variance. This definition is clearer if we look at *Figure 7*.

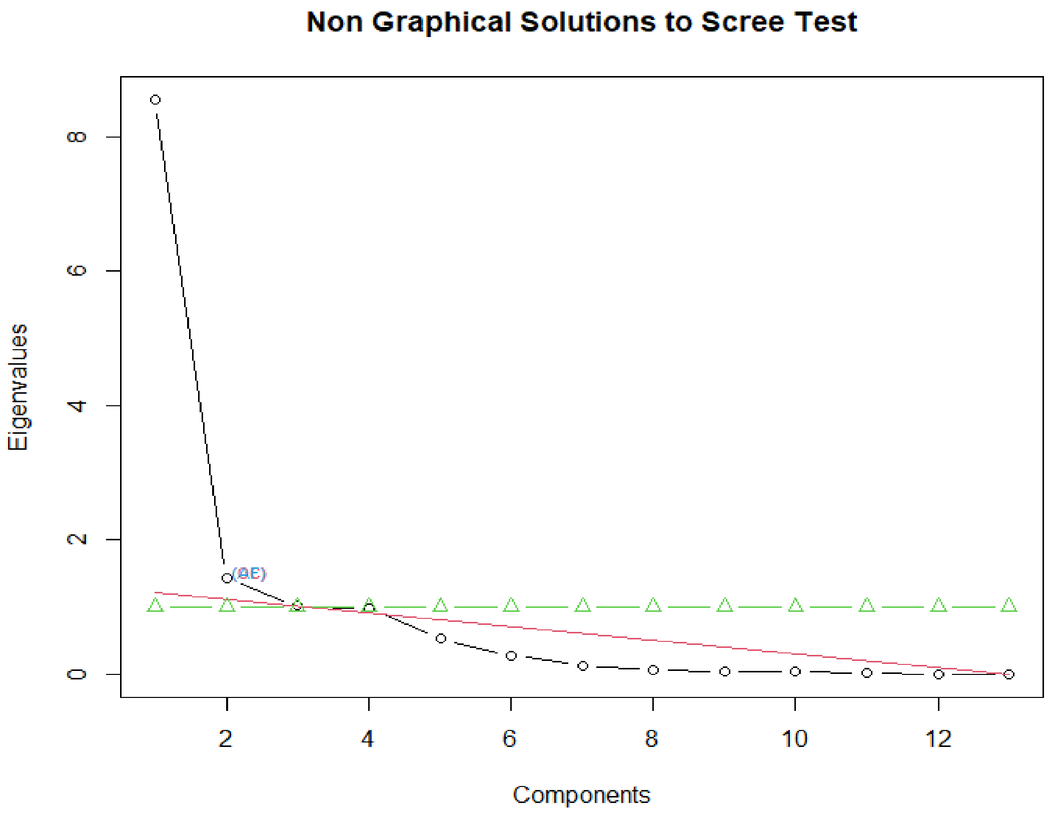


Fig.7 : Scree Plot of eigenvalues

There are different methods to decide how many eigenvalues to retain. One of the most used methods is the Kaiser criterion, a rule of thumb used in factor analysis to determine the number of factors to retain. The criterion suggests retaining only those factors whose eigenvalues are greater than 1. For our case this would mean selecting 3 factors.

Another commonly utilized approach involves examining the scree plot, which graphically represents the eigenvalues in descending order. The point where the plot displays an "elbow" or a noticeable change in slope is often considered an indicator of the optimal number of factors to retain. In our case, the factors with eigenvalues above the elbow are retained, (above **4**) as they capture substantial variance in the data. Furthermore, even if only the first three eigenvalues have a value greater than one, we decided to also include the fourth because its value is close to 1 and we can see a natural break in the Scree Plot, with eigenvalues shifting low and to the right after the fourth eigenvalue. Also we will find later that the fourth eigenvalue is important to explain the variance of the factor that represent the unemployment variable, in our opinion very relevant for the research.

As it can easily be seen in our plot, the first eigenvalue is above 8, suggesting that the corresponding factor is highly important in explaining the underlying structure of the observed variables. It indicates a strong relationship among the variables loading onto that factor.

3. Factor analysis

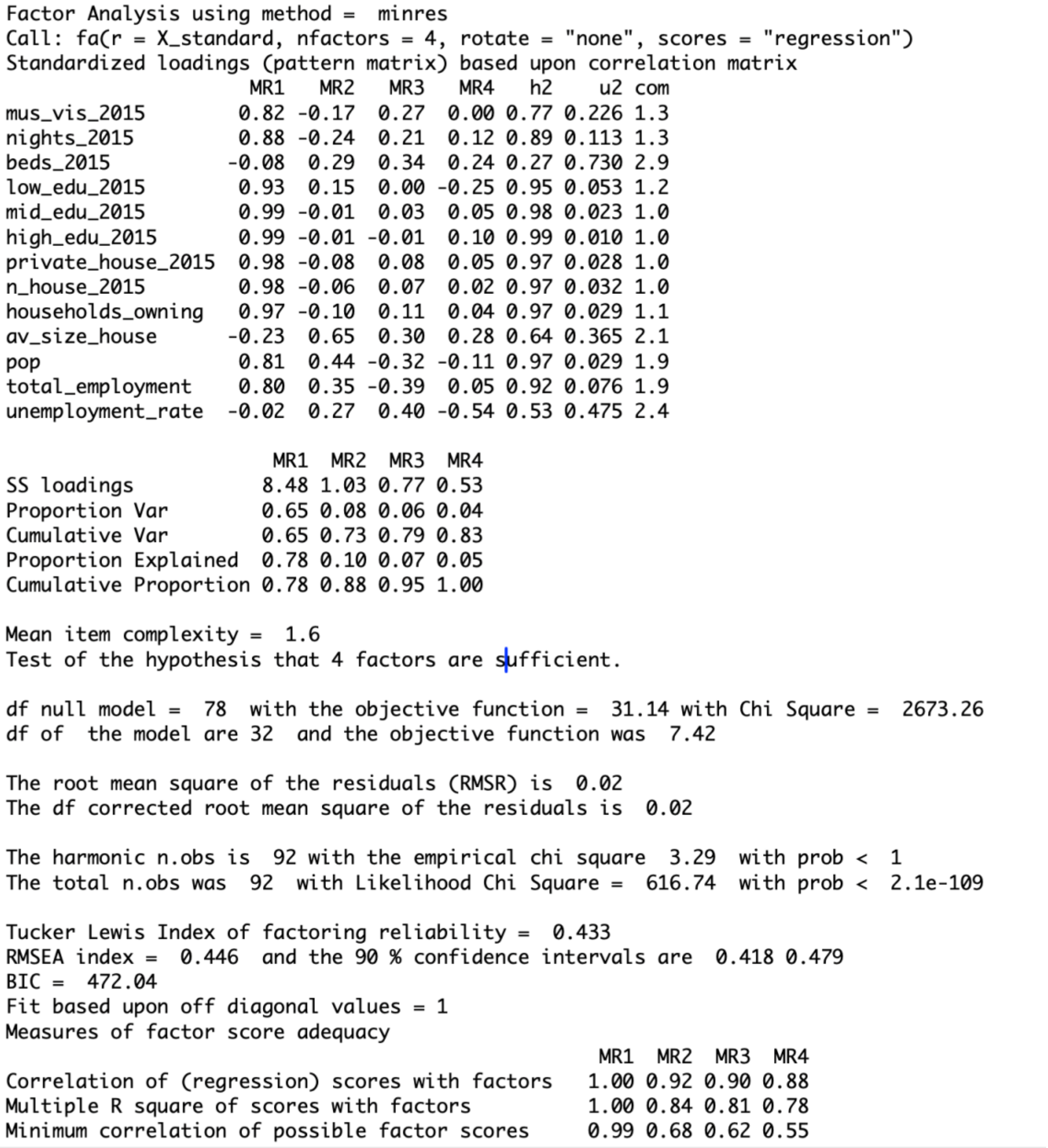
To conclude our analysis, now we will compute the factor analysis and try to give an answer to the 4 factors that explain our dataset of Italian cities. For this part we used the libraries “**psych”** and “**GPArotation”.**

3.1 Rotations

3.1.1 No rotations

The first FA was done without rotations:

M1 <- fa(X_standard, nfactors=4, rotate="none", scores="regression")



```
Factor Analysis using method =  minres
Call: fa(r = X_standard, nfactors = 4, rotate = "none", scores = "regression")
Standardized loadings (pattern matrix) based upon correlation matrix
                      MR1   MR2   MR3   MR4   h2    u2 com
mus_vis_2015         0.82 -0.17  0.27  0.00 0.77 0.226 1.3
nights_2015          0.88 -0.24  0.21  0.12 0.89 0.113 1.3
beds_2015           -0.08  0.29  0.34  0.24 0.27 0.730 2.9
low_edu_2015         0.93  0.15  0.00 -0.25 0.95 0.053 1.2
mid_edu_2015         0.99 -0.01  0.03  0.05 0.98 0.023 1.0
high_edu_2015        0.99 -0.01 -0.01  0.10 0.99 0.010 1.0
private_house_2015   0.98 -0.08  0.08  0.05 0.97 0.028 1.0
n_house_2015         0.98 -0.06  0.07  0.02 0.97 0.032 1.0
households_owning    0.97 -0.10  0.11  0.04 0.97 0.029 1.1
av_size_house       -0.23  0.65  0.30  0.28 0.64 0.365 2.1
pop                  0.81  0.44 -0.32 -0.11 0.97 0.029 1.9
total_employment     0.80  0.35 -0.39  0.05 0.92 0.076 1.9
unemployment_rate   -0.02  0.27  0.40 -0.54 0.53 0.475 2.4

                       MR1  MR2  MR3  MR4
SS loadings           8.48 1.03 0.77 0.53
Proportion Var        0.65 0.08 0.06 0.04
Cumulative Var        0.65 0.73 0.79 0.83
Proportion Explained  0.78 0.10 0.07 0.05
Cumulative Proportion 0.78 0.88 0.95 1.00

Mean item complexity =  1.6
Test of the hypothesis that 4 factors are sufficient.

df null model =  78  with the objective function =  31.14 with Chi Square =  2673.26
df of  the model are 32  and the objective function was  7.42

The root mean square of the residuals (RMSR) is  0.02
The df corrected root mean square of the residuals is  0.02

The harmonic n.obs is  92 with the empirical chi square  3.29  with prob <  1
The total n.obs was  92  with Likelihood Chi Square =  616.74  with prob <  2.1e-109

Tucker Lewis Index of factoring reliability =  0.433
RMSEA index =  0.446  and the 90 % confidence intervals are  0.418 0.479
BIC =  472.04
Fit based upon off diagonal values = 1
Measures of factor score adequacy
                                                   MR1  MR2  MR3  MR4
Correlation of (regression) scores with factors   1.00 0.92 0.90 0.88
Multiple R square of scores with factors          1.00 0.84 0.81 0.78
Minimum correlation of possible factor scores     0.99 0.68 0.62 0.55
```

Factor 1 (MR1):

This factor is by far the most prominent one, with variables like mus_vis_2015, nights_2015, education and housing related variables, and pop show high positive loadings, suggesting a strong association with Factor 1.

Factor 2 (MR2):

This factor has moderate loadings across several variables, such as av_size_house, pop, and employment.

Factor 3 (MR3):

Variables like av_size_house and unemployment_rate have positive loadings on this factor, but they are not as strong as those in Factor 1.

Factor 4 (MR4):

This factor includes variables like high_edu_2015, low_edu_2015, and unemployment_rate with mixed loadings. The negative loading on unemployment_rate suggests an inverse relationship with high education and converse relationship with low education.

Model fit:

The proportion of variance explained by each factor is displayed, with Factor 1 (MR1) explaining the highest proportion.

To evaluate the model fit, we need to look at different statistics:

Root Mean Square Residual (RMSR):

The RMSR is 0.02, indicating a very low average discrepancy between the observed and predicted covariances, which suggests a good fit.

Empirical Chi-Square:

This is a chi-square statistic based on empirical data. The value of 3.29 with a probability less than 1 suggests that the model fits the data well.

Likelihood Chi-Square:

This is another chi-square statistic based on likelihood estimation. The extremely low probability (prob < 2.1e-109) indicates a very good fit, almost a perfect fit to the data.

Tucker Lewis Index (TLI):

TLI compares the fit of the specified model to a null model. Your TLI of 0.433 is below the commonly accepted threshold for good fit (0.90). A higher TLI is generally preferred.

In this case we don’t consider this model as good to explain the factor structure of our dataset and we want to consider rotations, since the loadings are concentrated into only one factor and the other three factors only have one variable for each with fairly high loadings, as seen in Fig. 8.

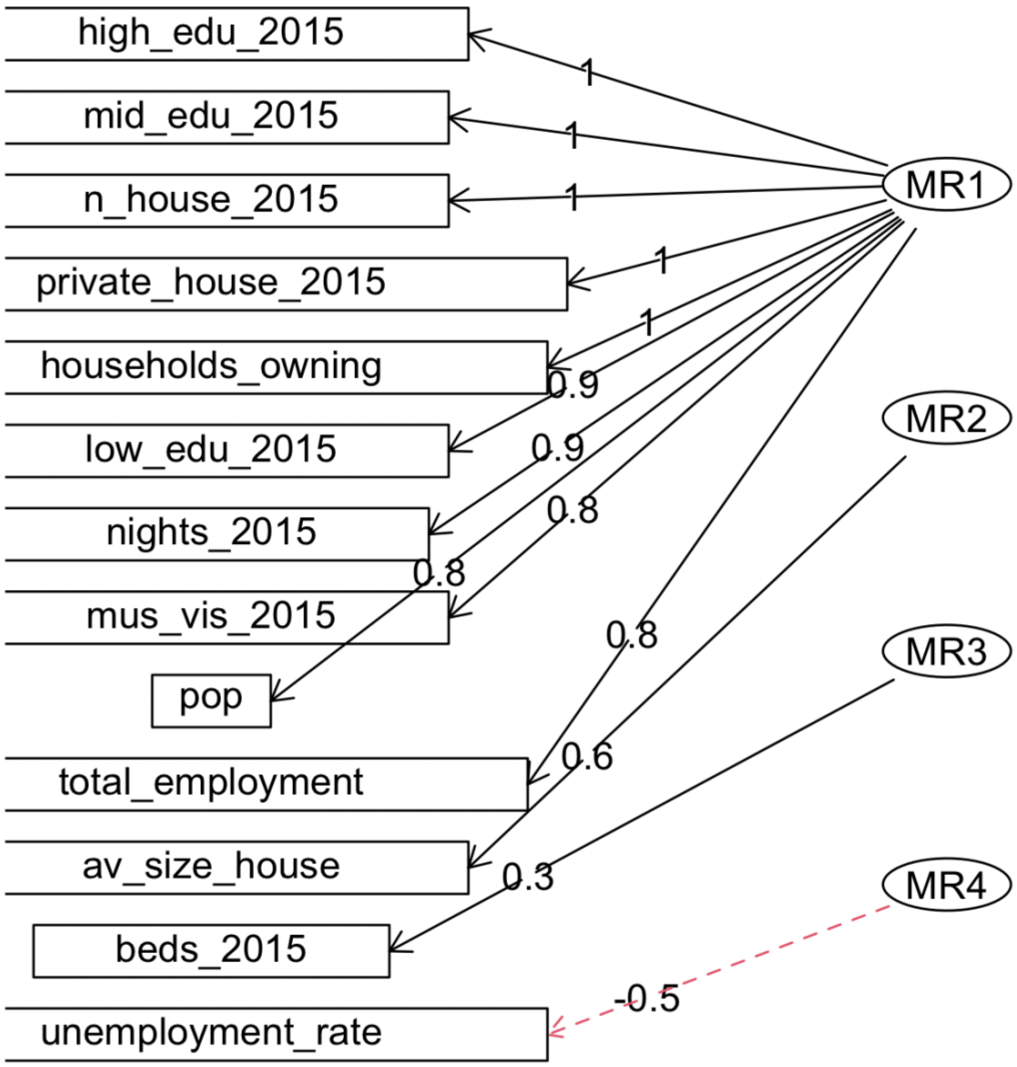


Fig.8 : Factor analysis with no rotation

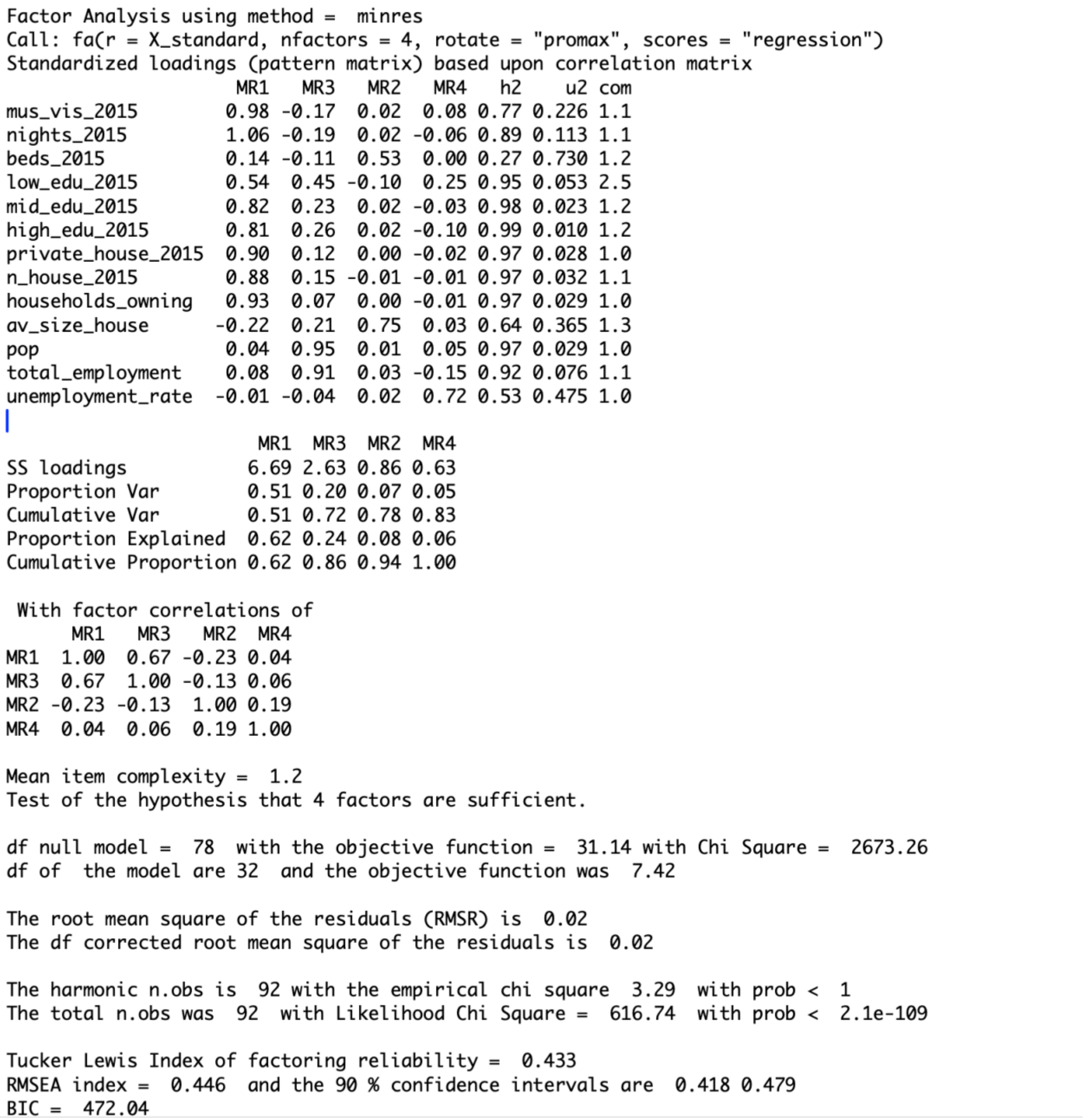
An important difference is that, in contrary to the other models with different rotations, unemployment rate has negative loadings in respect to factor 4. In all the other rotations, it has positive loadings..

3.1.2 Promax rotation

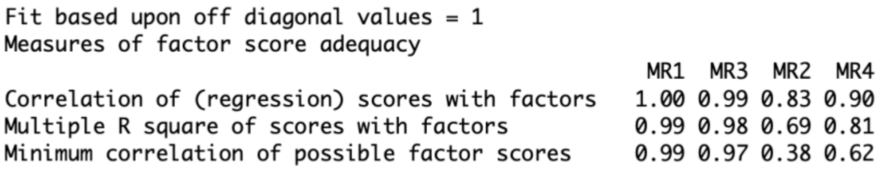
The second FA was done using the **Promax** rotation:

M2 <- fa(X_standard, nfactors=4, rotate="promax", scores="regression")

With the Promax rotation, we allow factors to be correlated between each other.



```
Factor Analysis using method =  minres
Call: fa(r = X_standard, nfactors = 4, rotate = "promax", scores = "regression")
Standardized loadings (pattern matrix) based upon correlation matrix
                     MR1   MR3   MR2   MR4   h2    u2 com
mus_vis_2015        0.98 -0.17  0.02  0.08 0.77 0.226 1.1
nights_2015         1.06 -0.19  0.02 -0.06 0.89 0.113 1.1
beds_2015           0.14 -0.11  0.53  0.00 0.27 0.730 1.2
low_edu_2015        0.54  0.45 -0.10  0.25 0.95 0.053 2.5
mid_edu_2015        0.82  0.23  0.02 -0.03 0.98 0.023 1.2
high_edu_2015       0.81  0.26  0.02 -0.10 0.99 0.010 1.2
private_house_2015  0.90  0.12  0.00 -0.02 0.97 0.028 1.0
n_house_2015        0.88  0.15 -0.01 -0.01 0.97 0.032 1.1
households_owning   0.93  0.07  0.00 -0.01 0.97 0.029 1.0
av_size_house      -0.22  0.21  0.75  0.03 0.64 0.365 1.3
pop                 0.04  0.95  0.01  0.05 0.97 0.029 1.0
total_employment    0.08  0.91  0.03 -0.15 0.92 0.076 1.1
unemployment_rate  -0.01 -0.04  0.02  0.72 0.53 0.475 1.0

                       MR1  MR3  MR2  MR4
SS loadings           6.69 2.63 0.86 0.63
Proportion Var        0.51 0.20 0.07 0.05
Cumulative Var        0.51 0.72 0.78 0.83
Proportion Explained  0.62 0.24 0.08 0.06
Cumulative Proportion 0.62 0.86 0.94 1.00

 With factor correlations of 
      MR1   MR3   MR2  MR4
MR1  1.00  0.67 -0.23 0.04
MR3  0.67  1.00 -0.13 0.06
MR2 -0.23 -0.13  1.00 0.19
MR4  0.04  0.06  0.19 1.00

Mean item complexity =  1.2
Test of the hypothesis that 4 factors are sufficient.

df null model =  78  with the objective function =  31.14 with Chi Square =  2673.26
df of  the model are 32  and the objective function was  7.42 

The root mean square of the residuals (RMSR) is  0.02 
The df corrected root mean square of the residuals is  0.02 

The harmonic n.obs is  92 with the empirical chi square  3.29  with prob <  1 
The total n.obs was  92  with Likelihood Chi Square =  616.74  with prob <  2.1e-109 

Tucker Lewis Index of factoring reliability =  0.433
RMSEA index =  0.446  and the 90 % confidence intervals are  0.418 0.479
BIC =  472.04
```



```
Fit based upon off diagonal values = 1
Measures of factor score adequacy             
                                                   MR1  MR3  MR2  MR4
Correlation of (regression) scores with factors   1.00 0.99 0.83 0.90
Multiple R square of scores with factors          0.99 0.98 0.69 0.81
Minimum correlation of possible factor scores     0.99 0.97 0.38 0.62
```

With this rotation the proportionate variability is more spread among the four different factors: MR1 dominates, accounting for 51% of the variance, followed by MR3 (20%), MR2 (7%), and MR4 (5%).

Factor correlation:

Factor correlations provide insights into the relationships between latent factors. MR1 and MR3 exhibit a positive correlation (0.67), while MR2 and MR4 are positively correlated (0.19). Notably, MR1 and MR2, as well as MR3 and MR4, display negative correlations.

The model fit as in the first model suggests a reasonable fit of the model to the data.

Overall, this rotated Factor Analysis provides a more nuanced understanding of the underlying structure in your dataset compared to the non-rotated version.

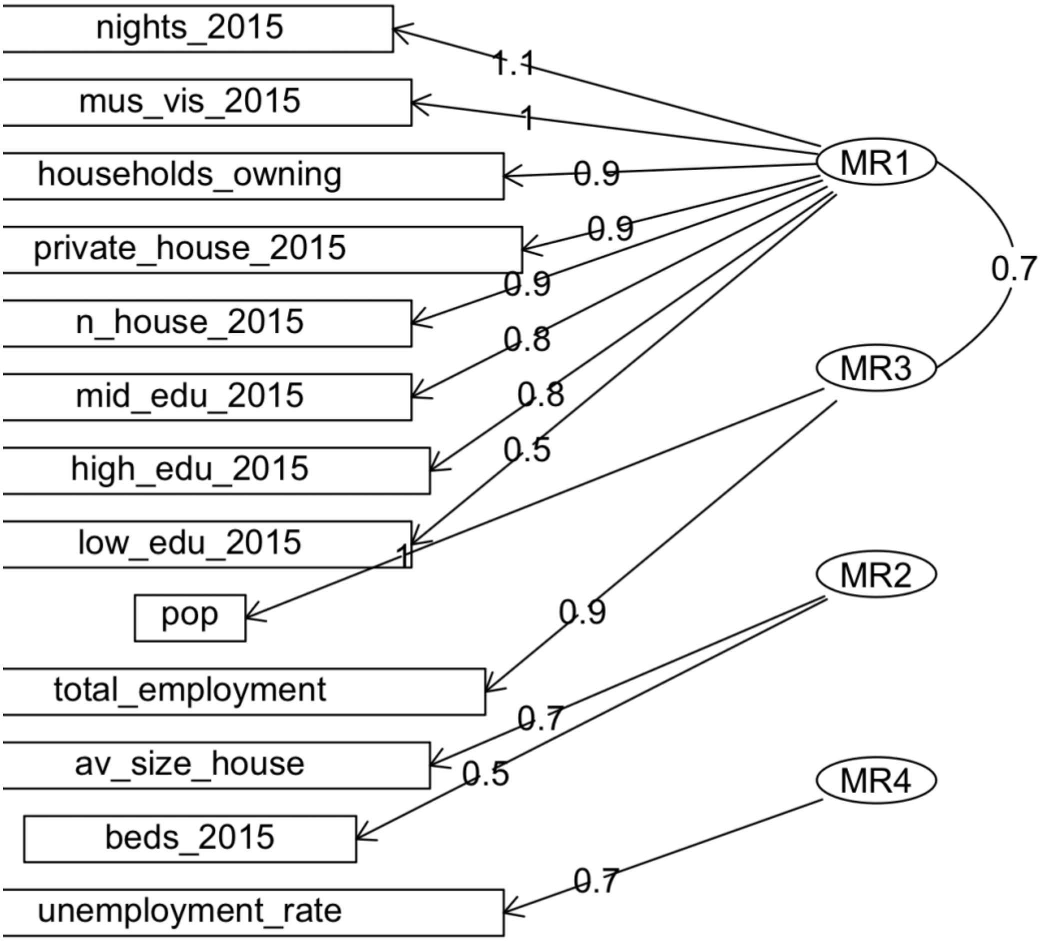


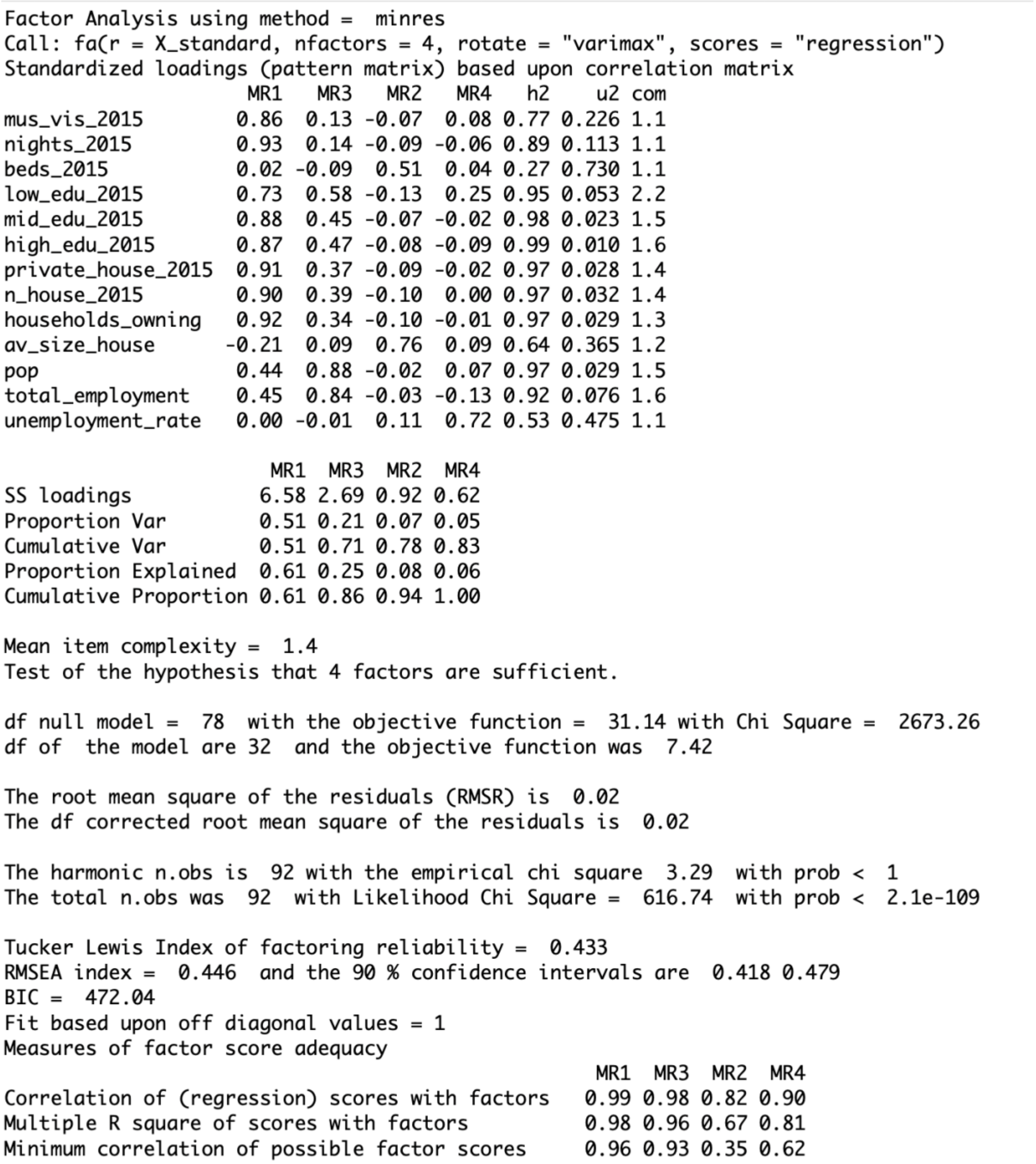
Fig.9 : Factor analysis with Promax rotation

3.1.3 Varimax rotation

The third FA was done with the **Varimax** rotation:

M3 <- fa(X_standard, nfactors=4, rotate="varimax", scores="regression")

The Varimax rotation allows to maximize the variance of squared loadings within each factor, leading to a simpler and more interpretable factor structure.



```
Factor Analysis using method =  minres
Call: fa(r = X_standard, nfactors = 4, rotate = "varimax", scores = "regression")
Standardized loadings (pattern matrix) based upon correlation matrix
                     MR1   MR3   MR2   MR4   h2    u2 com
mus_vis_2015        0.86  0.13 -0.07  0.08 0.77 0.226 1.1
nights_2015         0.93  0.14 -0.09 -0.06 0.89 0.113 1.1
beds_2015           0.02 -0.09  0.51  0.04 0.27 0.730 1.1
low_edu_2015        0.73  0.58 -0.13  0.25 0.95 0.053 2.2
mid_edu_2015        0.88  0.45 -0.07 -0.02 0.98 0.023 1.5
high_edu_2015       0.87  0.47 -0.08 -0.09 0.99 0.010 1.6
private_house_2015  0.91  0.37 -0.09 -0.02 0.97 0.028 1.4
n_house_2015        0.90  0.39 -0.10  0.00 0.97 0.032 1.4
households_owning   0.92  0.34 -0.10 -0.01 0.97 0.029 1.3
av_size_house      -0.21  0.09  0.76  0.09 0.64 0.365 1.2
pop                 0.44  0.88 -0.02  0.07 0.97 0.029 1.5
total_employment    0.45  0.84 -0.03 -0.13 0.92 0.076 1.6
unemployment_rate   0.00 -0.01  0.11  0.72 0.53 0.475 1.1

                       MR1  MR3  MR2  MR4
SS loadings           6.58 2.69 0.92 0.62
Proportion Var        0.51 0.21 0.07 0.05
Cumulative Var        0.51 0.71 0.78 0.83
Proportion Explained  0.61 0.25 0.08 0.06
Cumulative Proportion 0.61 0.86 0.94 1.00

Mean item complexity =  1.4
Test of the hypothesis that 4 factors are sufficient.

df null model =  78  with the objective function =  31.14 with Chi Square =  2673.26
df of  the model are 32  and the objective function was  7.42

The root mean square of the residuals (RMSR) is  0.02
The df corrected root mean square of the residuals is  0.02

The harmonic n.obs is  92 with the empirical chi square  3.29  with prob <  1
The total n.obs was  92  with Likelihood Chi Square =  616.74  with prob <  2.1e-109

Tucker Lewis Index of factoring reliability =  0.433
RMSEA index =  0.446  and the 90 % confidence intervals are  0.418 0.479
BIC =  472.04
Fit based upon off diagonal values = 1
Measures of factor score adequacy
                                                   MR1  MR3  MR2  MR4
Correlation of (regression) scores with factors   0.99 0.98 0.82 0.90
Multiple R square of scores with factors          0.98 0.96 0.67 0.81
Minimum correlation of possible factor scores     0.96 0.93 0.35 0.62
```

With this rotation we can see a clearer representation of the factors: MR1 (Tourism), MR2 (Housing), MR3 (Education), and MR4 (Population and Employment) are distinct factors. Each factor contributes meaningfully, with Tourism being the dominant factor.

The model fits well, as indicated by various fit indices as in previous models.

In comparing the three previous models in terms of variance explained, Varimax rotation (Model 3) and No Rotation (Model 1) have similar cumulative variance (83%), while Promax rotation (Model 2) explains 100% of the variance.

The factor structure in Promax rotation leads to more evenly distributed loadings across factors, whereas the other models may have more concentrated loadings.

In summary, for now the Promax rotation may provide a more interpretable factor structure, especially when factors are correlated. Since we expect factors to be correlated, Promax rotation may provide a more realistic representation.

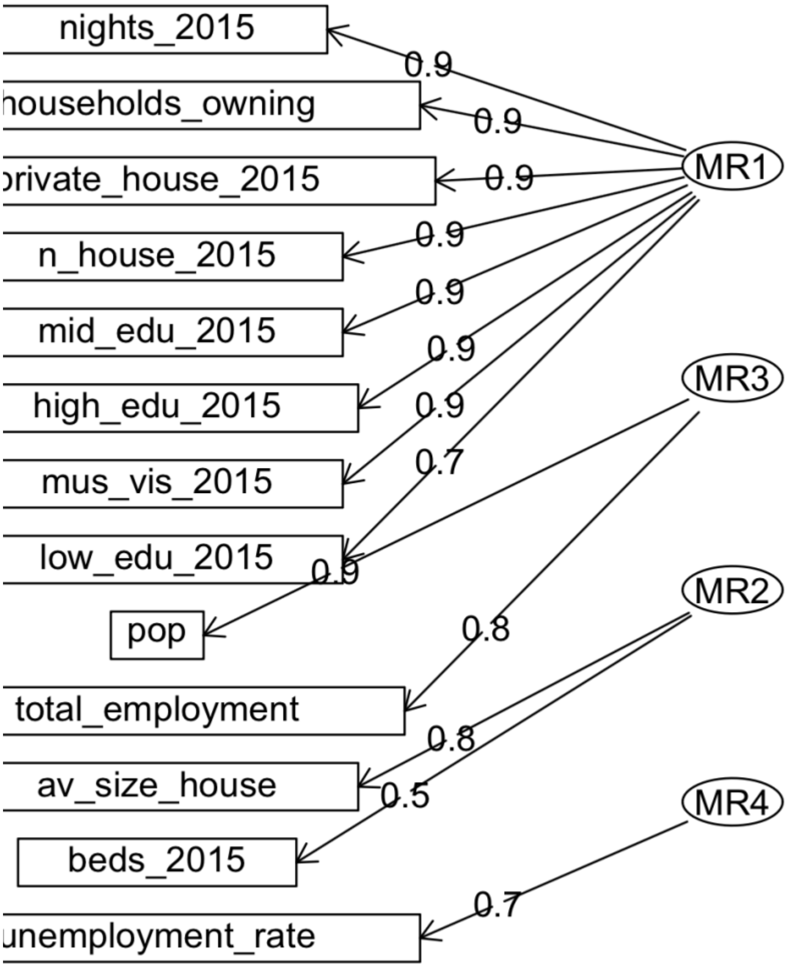


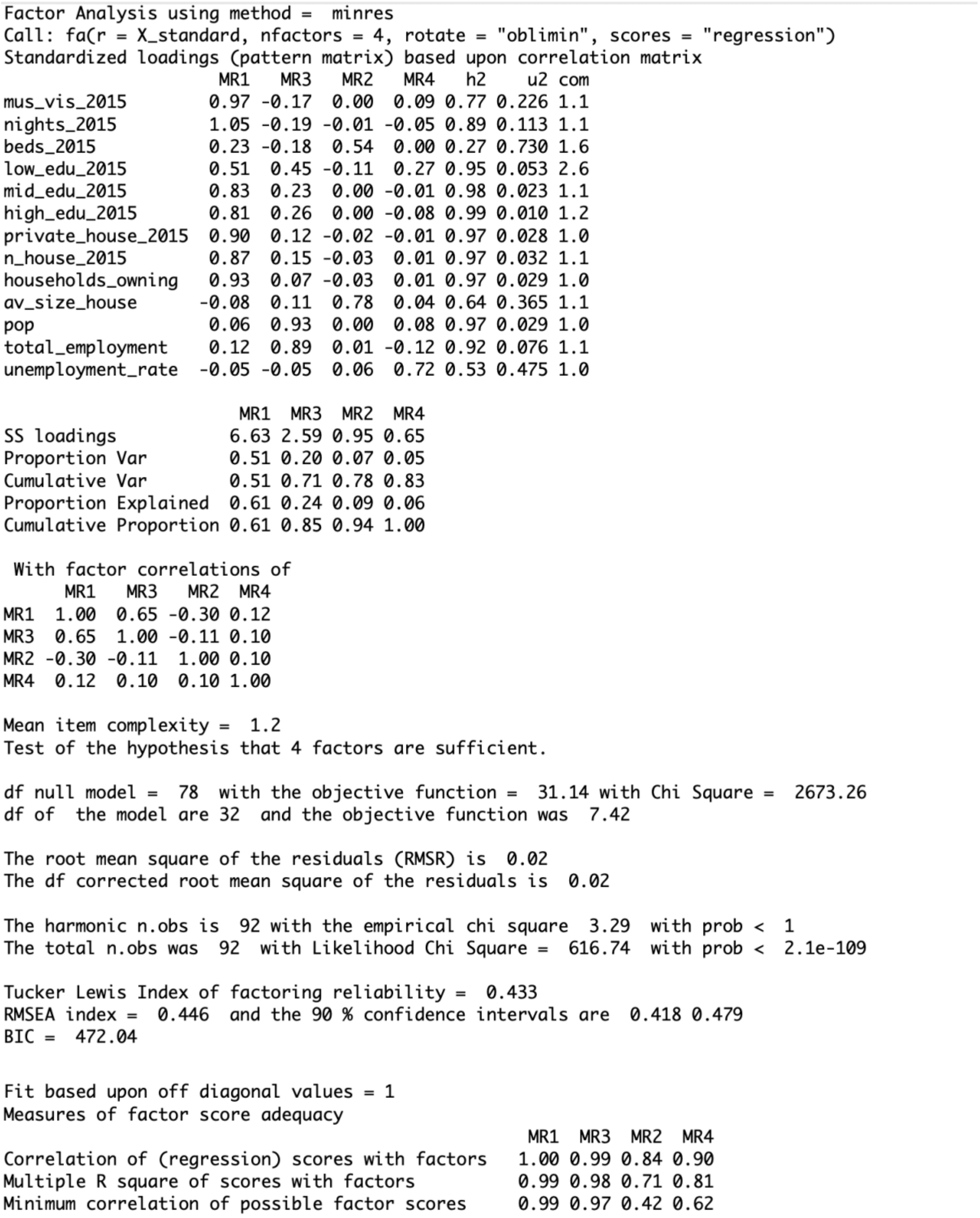
Fig.10 : Factor analysis with Varimax rotation

3.1.4 Oblimin rotation

The fourth FA was done by using the **Oblimin** rotation:

M4 <- fa(X_standard, nfactors=4, rotate="oblimin", scores="regression")

Oblimin is an oblique rotation that allows factors to be correlated.



```
Factor Analysis using method =  minres
Call: fa(r = X_standard, nfactors = 4, rotate = "oblimin", scores = "regression")
Standardized loadings (pattern matrix) based upon correlation matrix
                     MR1   MR3   MR2   MR4   h2    u2 com
mus_vis_2015        0.97 -0.17  0.00  0.09 0.77 0.226 1.1
nights_2015         1.05 -0.19 -0.01 -0.05 0.89 0.113 1.1
beds_2015           0.23 -0.18  0.54  0.00 0.27 0.730 1.6
low_edu_2015        0.51  0.45 -0.11  0.27 0.95 0.053 2.6
mid_edu_2015        0.83  0.23  0.00 -0.01 0.98 0.023 1.1
high_edu_2015       0.81  0.26  0.00 -0.08 0.99 0.010 1.2
private_house_2015  0.90  0.12 -0.02 -0.01 0.97 0.028 1.0
n_house_2015        0.87  0.15 -0.03  0.01 0.97 0.032 1.1
households_owning   0.93  0.07 -0.03  0.01 0.97 0.029 1.0
av_size_house      -0.08  0.11  0.78  0.04 0.64 0.365 1.1
pop                 0.06  0.93  0.00  0.08 0.97 0.029 1.0
total_employment    0.12  0.89  0.01 -0.12 0.92 0.076 1.1
unemployment_rate  -0.05 -0.05  0.06  0.72 0.53 0.475 1.0

                       MR1  MR3  MR2  MR4
SS loadings           6.63 2.59 0.95 0.65
Proportion Var        0.51 0.20 0.07 0.05
Cumulative Var        0.51 0.71 0.78 0.83
Proportion Explained  0.61 0.24 0.09 0.06
Cumulative Proportion 0.61 0.85 0.94 1.00

 With factor correlations of 
      MR1   MR3   MR2  MR4
MR1  1.00  0.65 -0.30 0.12
MR3  0.65  1.00 -0.11 0.10
MR2 -0.30 -0.11  1.00 0.10
MR4  0.12  0.10  0.10 1.00

Mean item complexity =  1.2
Test of the hypothesis that 4 factors are sufficient.

df null model =  78  with the objective function =  31.14 with Chi Square =  2673.26
df of  the model are 32  and the objective function was  7.42 

The root mean square of the residuals (RMSR) is  0.02 
The df corrected root mean square of the residuals is  0.02 

The harmonic n.obs is  92 with the empirical chi square  3.29  with prob <  1 
The total n.obs was  92  with Likelihood Chi Square =  616.74  with prob <  2.1e-109 

Tucker Lewis Index of factoring reliability =  0.433
RMSEA index =  0.446  and the 90 % confidence intervals are  0.418 0.479
BIC =  472.04

Fit based upon off diagonal values = 1
Measures of factor score adequacy             
                                                   MR1  MR3  MR2  MR4
Correlation of (regression) scores with factors   1.00 0.99 0.84 0.90
Multiple R square of scores with factors          0.99 0.98 0.71 0.81
Minimum correlation of possible factor scores     0.99 0.97 0.42 0.62
```

The factor analysis with Oblimin rotation provides a refined factor structure, highlighting meaningful relationships between variables and factors. The fit indices and factor score adequacy measures collectively suggest that the rotated model is a substantial improvement over the non-rotated version. The correlated factors enhance the interpretability of the underlying structure, contributing valuable insights to our understanding of the dataset.

The h2 column represents the communality of each variable, which is the proportion of each variable’s variance that can be explained by the factors. The u2 column represents the uniqueness of each variable, which is the proportion of each variable’s variance that cannot be explained by the factors. As we can see in the results, there are variables that aren't well represented in terms of variability explained by the factors, especially beds_2015, with a uniqueness of 0.73. The com column represents the complexity of each variable, which is a measure of how much a variable is represented by more than one factor. As we can see in the results, almost all variables have a complexity of about 1, meaning that they are represented mainly by one factor, driven mainly by the first factor. An exception by low_edu_2015 that has a complexity of 2.6.

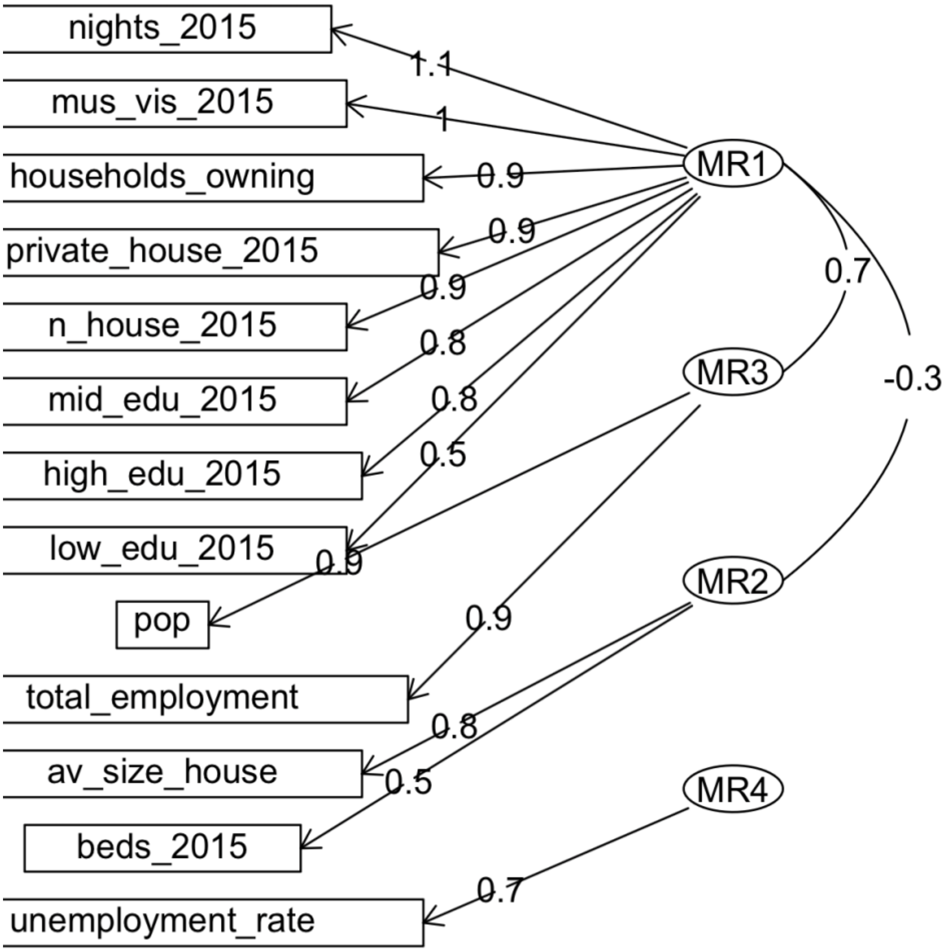


Fig.11 : Factor analysis with oblimin rotation

4. Interpretation of the results

Since we expect the factors to be correlated between each other, we decided to consider the Oblimin rotation for interpreting our results.

Based on the factor analysis results, the key latent variables (factors) that describe your cities could be interpreted as follows:

Factor MR1: This factor has high loadings from mus_vis_2015, nights_2015, low_edu_2015, mid_edu_2015, high_edu_2015, private_house_2015, n_house_2015, and households_owning. This suggests that MR1 might represent a “Socio-Economic Status” and “Tourism activity” factor, capturing aspects related to education level, housing, and tourism.

Factor MR3: This factor has high loadings from pop and total_employment. This suggests that MR3 might represent a “Population and Employment” factor, capturing aspects related to the size of the population and employment levels. Cities with higher population will have higher level of employment, which is a common expectation in urban areas with more economic opportunities and a larger job market.

Factor MR2: This factor has high loadings from av_size_house and beds_2015. This suggests that MR2 might represent a “Housing Characteristics” factor, capturing aspects related to the average size of houses and the number of beds. Cities with higher values on this factor may exhibit certain housing characteristics, such as larger average house sizes and more beds. This information can be valuable for understanding the housing landscape and infrastructure in these cities.

Factor MR4: This factor has high loadings from unemployment_rate. This factor has relatively low correlation with other factors, suggesting that the “Unemployment” factor is fairly independent from the other factors. To explore potential improvements, an alternative factor analysis with oblimin rotation was conducted, omitting the "Unemployment Rate" variable and employing a three-factor structure instead of four. However, the outcomes showed minimal variation, also in terms of model fit, leading to the decision to retain the original four-factor model due to its stability and consistency with the research objectives.

The correlations between factors should indeed be considered when interpreting our results. For instance, there is a moderate positive correlation of 0.67 between factor MR1 and MR3. This implies that cities with a higher socio-economic status, characterized by higher income and education levels, tend to also have larger populations and higher employment rates. This relationship is quite intuitive, as exemplified by cities like Milan, where a higher level of education and income is typically associated with a higher employment rate.

On the other hand, factors MR1 and MR2 exhibit a slightly negative correlation of - 0.30. This suggests that as socio-economic status and tourism activity increase, the average house size tends to decrease. This trend could potentially be attributed to the urban characteristics of Italian cities. Major Italian cities tend to have higher population density, due to historical and geographical reasons, and are generally concentrated in urban areas where space is at a premium, leading to the construction of smaller houses.

As socio-economic status rises and tourism activity increases, there's a greater demand for residential space within these urban centers. However, the limited availability of land and the premium placed on location in popular tourist destinations lead to a trend of optimizing space. This optimization can manifest in smaller average house sizes as residents prioritize convenience and proximity over larger living spaces. Additionally, the proliferation of short-term rental sites like Airbnb has encouraged property owners to divide bigger properties into several smaller apartments in order to meet the demand for short-term rentals. This further contributes to the decrease in average house size as properties are subdivided to maximize profitability.