Phd Program in Transportation

Transport Demand Modeling

Filipe Moura

Session 4

Factor Analysis

Highlights of Factorial Analysis

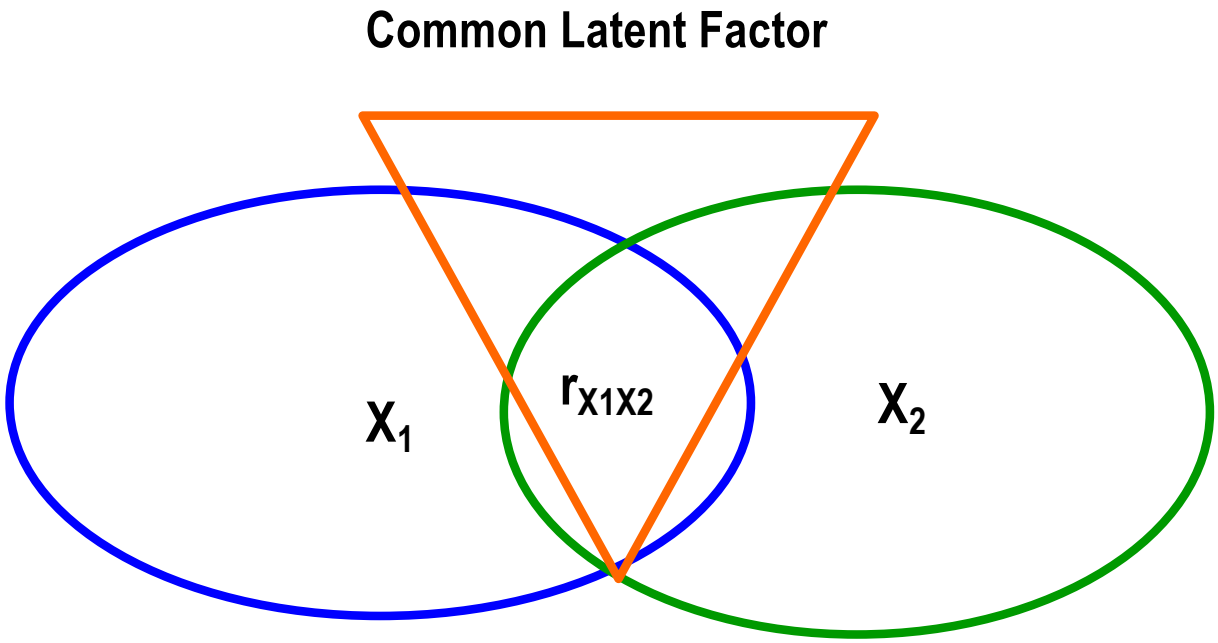
- Exploratory technique aimed at **defining the underlying structure among a group of interrelated variables**.
- The objective of FA is to explain the variance-covariance between a set of relatedvariables
- Factor analysis allows the **construction of a measurement scale for the factors that control the original variables**.
- It is also a technique used to **reduce the number of variables** in other multivariate methods.
- It uses the **correlations between variables to estimate the common underlying factors**
  - "*Something behind the numbers that we can't measure, relates the variables*"

Uses of factor analysis

- **Data summarization** – *understanding relationship between variables (a theory)*
  - Define a **small number of factors that adequately represent the original set of variables**.
  - Each dependent variable is a **function of an underlying and latent set of factors**.
  - **Each variable is predicted by all the factors** (and indirectly by all the other variables)
- **Data reduction** – *reduce variables to a smaller surrogate set of variables*
  - **Identify representative variables** from a much larger set of variables for use in subsequent analysis or create a new set of variables, much smaller in number to replace them.
  - **Data reduction relies on factor loadings** but uses them as the basis for identifying variables or making estimates of the factors for subsequent analysis

Definition and Purpose (I)

- If **two variables are correlated** this association results from the fact that both share a **common characteristic** which cannot be directly observed (a **latent common factor**).



Definition and Purpose (II)

- The general purpose of factor analysis is to **find a way to condense the information contained in a group of variables** into a **smaller set of composite dimensions** loosing the minimum amount of information in this process – search for the fundamental dimensions that underlie the original variables
- R type factor analysis
  - Analyses the **correlation matrices of the variables**
- Q type factor analysis
  - Analyses the **correlation matrix of the individual respondents** based on their characteristics

Definition and Purpose (III)

- Factor analysis is an **interdependence technique**
  - All variables are considered **simultaneously**, without any distinctions between dependent and independent variables
- A variate (or factor) is a **linear composite of variables**.
- It is formed in order to maximize the explanation of the variance of the entire set of variables

$$x_1 = a_{11}f_1 + a_{12}f_2 + e_1,$$
$$x_2 = a_{21}f_1 + a_{22}f_2 + e_2,$$
$$x_3 = a_{31}f_1 + a_{32}f_2 + e_3.$$

- Communalities

- Specific and errors

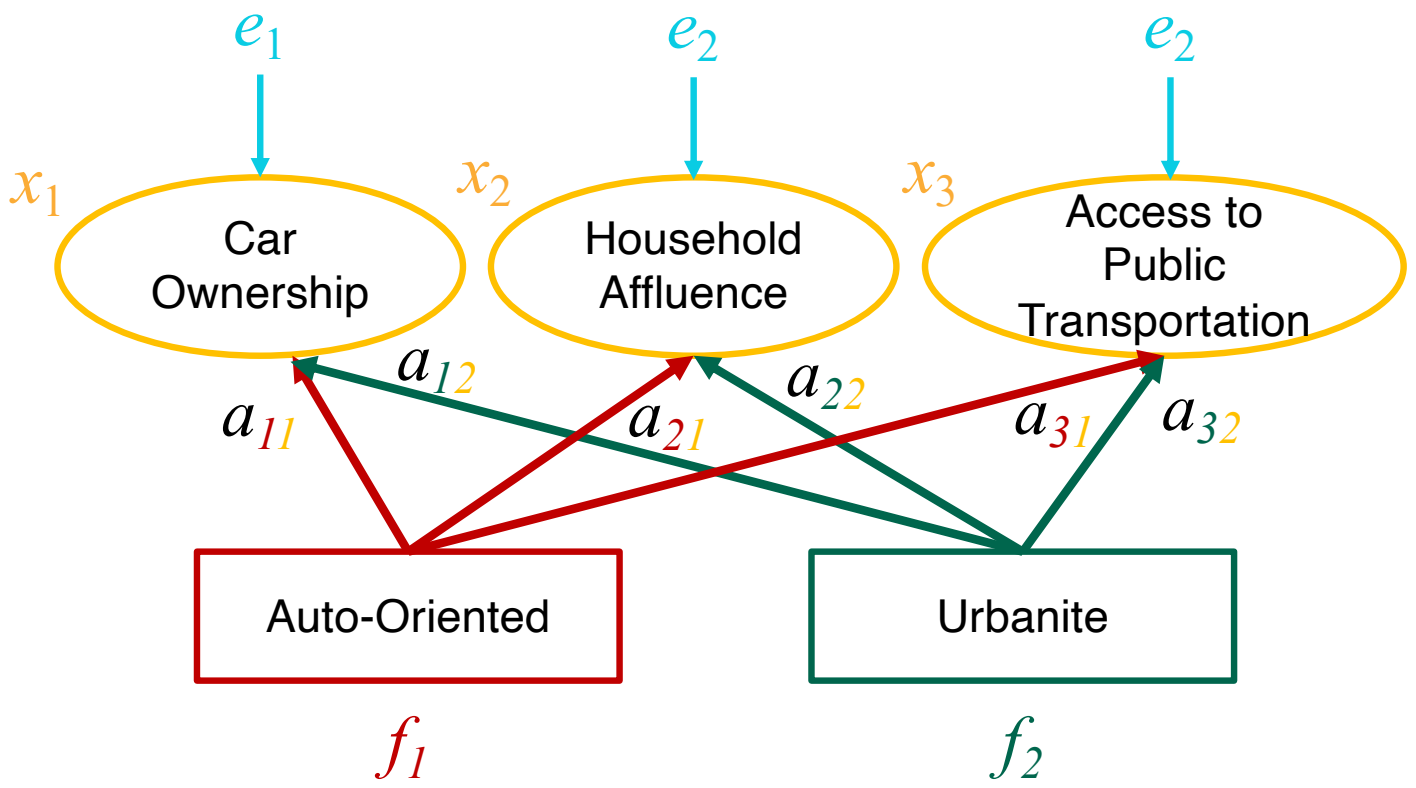
, where x_p are observed variables

f_m are common factors (latent variables), and desirably m<p

e_p are specific factors of each variable or errors

a_{pm} represent the contribution of each factor to the explanation of each variable (factor loadings).

Definition and Purpose (III)



Vector of equations:

$$x_1 = a_{11}f_1 + a_{12}f_2 + e_1,$$
$$x_2 = a_{21}f_1 + a_{22}f_2 + e_2,$$
$$x_3 = a_{31}f_1 + a_{32}f_2 + e_3.$$

Definition and Purpose (V)

$$
\begin{aligned}
x_1 &= a_{11} f_1 + a_{12} f_2 + e_1, \\
x_2 &= a_{21} f_1 + a_{22} f_2 + e_2, \\
x_3 &= a_{31} f_1 + a_{32} f_2 + e_3.
\end{aligned}
$$

- Communalities
- Specific and errors

- x_p can be analyzed as is, or in their standardized form (assuming normal distribution of variables).
  - As such, factor loadings are direct or standardized, respectively.
  - When standardized, the analysis rely on the correlation matrix (which is the same as the standardized variance-covariance matrix).
- In factor analysis,
  - Common factors f_m are orthogonal (independent) and equally distributed with mean 0 and variance 1
  - Specific factors e_p are orthogonal (independent) and equally distributed with mean 0 and variance Ψ
  - f_m *and* e_p are orthogonal (independent)

Matrix notation

- In **matrix notation**, the factor analysis model is

$$Z = \Lambda f + \eta$$

  - Z is the vector of *p* standardized variables (with $z_p = (x_p - \mu_p)/\sigma_p$)
  - *f* is the vector of common factors f_m (μ=0 and σ=I)
  - η is the vector of specific factors e_p (μ=0 and $\sigma=\Psi$)
  - Λ is the matrix of factor loadings a_{mp} (many times referred to λ_{mp})
- Assuming that *f* and η are independent and ψ is the diagonal, we can deduct that

$$\Pi = \Lambda\Lambda^t + \Psi$$

  - Π is the correlation matrix (standardized version of the VAR-COVAR matrix)

Variable selection

- Factor analysis produces factors thus special care should be taken against "**garbage in garbage out**" phenomena
- The **quality and meaning of the derived factors reflect the conceptual underpinnings of the variables considered for the factor analysis**
- Factor analysis could be used to **introduce in other statistical techniques** a smaller number of new variables **(e.g., IV's in MLR)** either using representative variables or the factor scores

Variable selection

- **Nonmetric variables could be problematic**
  - It is prudent to avoid nonmetric variables and substitute them by dummy variables.
  - If all variables included in the factor analysis are dummy then other methods should be used

- Since factor analysis aims to find patterns among groups of variables, **factors with only one variable don´t make sense**

Sampling and Assumptions of factor analysis

- **Sample size**
  - N<50 unacceptable (N>200 is recommended)
  - Preferably at least 20 observations for each item (20:1; although as from 5:1 is still doable)
- **Assumptions of factor analysis**
  - More **conceptual** than statistical (doesn't mean that the statistical assumptions shouldn't be met)
    - Meaning that you should have an *a priori* understanding of communalities between observed variables
  - There is an **underlying structure in the data** (correlation between variables does not ensure the existence of this structure)
  - The **sample should be homogeneous with respect to the underlying factor structure** (e.g., variables that are different between men and women)

Statistical assumptions

- **Normality:** Statistical inference is improved if the variables are multivariate normal (although not necessary)
- **Linearity** between variables (examine bivariate scatterplots)
- **Some multicollinearity (homoscedasticity) is desirable**
- When **correlations** among variables are **small** (<0,3) or are all **equal**, then **factor analysis is not appropriate**
- Partial correlations
  - Correlation that is unexplained when the effects of other variables are taken into account.
  - **If they are high (>0,7) factor analysis is irrelevant**.
- Anti-Image correlation matrix
  - It is the negative value of the partial correlation.
  - **Large values of the diagonal indicate that the variables are independent**

Measures of Sampling Adequacy (MSA) (I)

- **KMO (Kaiser-Meyer-Olkin)**
  - **Measure of homogeneity**, which compares simple with partial correlations observed between variables

, where

$$KMO = \frac{\sum_{i=1}^{p-1}\sum_{j=i+1}^{p} r_{x_i x_j}^2}{\sum_{i=1}^{p-1}\sum_{j=i+1}^{p} r_{x_i x_j}^2 + \sum_{i=1}^{p-1}\sum_{j=i+1}^{p} r_{x_i x_j|x_k}^2}$$

$$r_{x_i x_j|x_k} = \frac{\left(r_{x_i x_j} - r_{x_i x_k} r_{x_j x_k}\right)}{\sqrt{\left(1 - r_{x_i x_j}^2\right)\left(1 - r_{x_j x_k}^2\right)}}$$

Measures of Sampling Adequacy (MSA) (II)

KMO value	Recommendations relative to Factor Analysis
]0,9;1,0]	Marvelous
]0,8;0,9]	Meritorious
]0,7;0,8]	Middling
]0,6;0,7]	Mediocre
]0,5;0,6]	Bad but still acceptable
≤ 0,5	Unacceptable

Measures of Sampling Adequacy (MSA) (III)

❒ Bartlett test of sphericity

- Statistical test for the presence of **correlation among variables**.
- Null hypothesis:
  - The **correlation matrix is not different from the identity matrix**, i.e. **none of the variables** being tested **may correlate with each other**.
  - You want to **reject the null hypothesis**, therefore get a **Chi-square test above threshold**
- It is sensible to sample size (more sensible in detecting correlations)

$$H_0: \Pi = I$$

$$H_a: \Pi \neq I$$

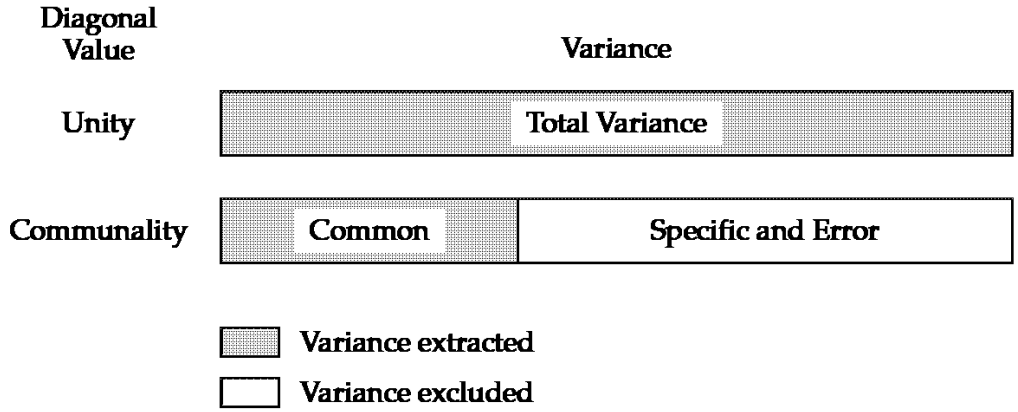
$$X^2 = -\left(N - 2 - \frac{2p+5}{6}\right) ln|R|$$

$$X^2 \geq \chi^2_{1-\alpha;(p(p-1)/2)} \; then \; H_0 \; is \; rejected$$

Extraction Methods

❒ Principal Components

- ➢ Considers the total variance of each component and derives factors that contain small proportions of unique variance
- ➢ Appropriate when data reduction is a primary concern.



❒ Common factor analysis

- ➢ Considers only the common or shared variance
- ➢ Error and specific variance are not of interest
- ➢ Objective is to identify the latent dimensions or constructs

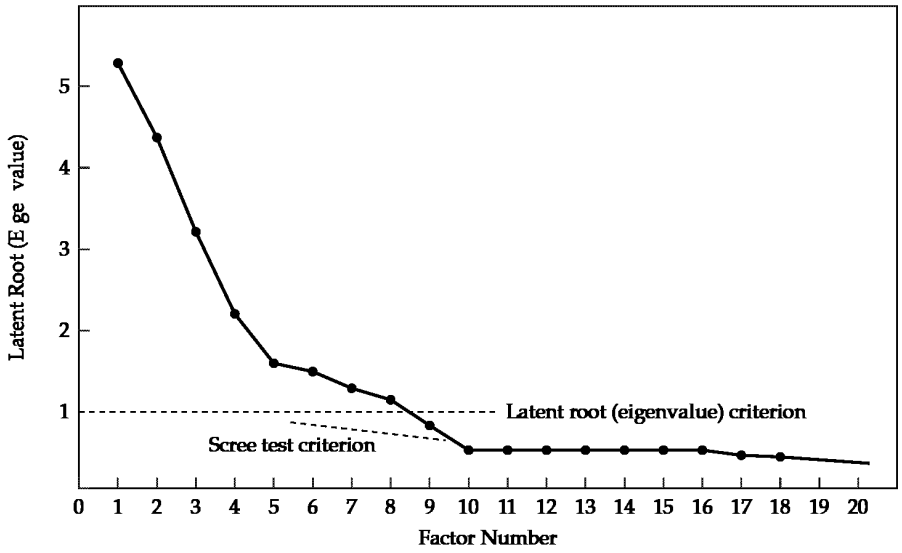
Factor extraction

- The number of factors to be extracted:
  - **Eigenvalue criterion**
    - Any individual factor should account for the variance of at least one single variable – Eingenvalue >1.
  - **A Priori criterion**
    - Define *a prio*ri the number of factors to be extracted (testing a hypothesis about the number of factors).
  - **Percentage of Variance criterion**
    - Achieving a specified cumulative percentage of total variance.
      - 95% in natural sciences;
      - >60% is not uncommon in social sciences .

Factor extraction

❒ Scree plot test

- ➢ In the scree plot graph, look for the point where there is an inflexion that usually correspond to eigenvalue 1 (aka, *latent root*).
- ➢ The point where the curve begins to straighten corresponds to the maximum number of factors to consider.

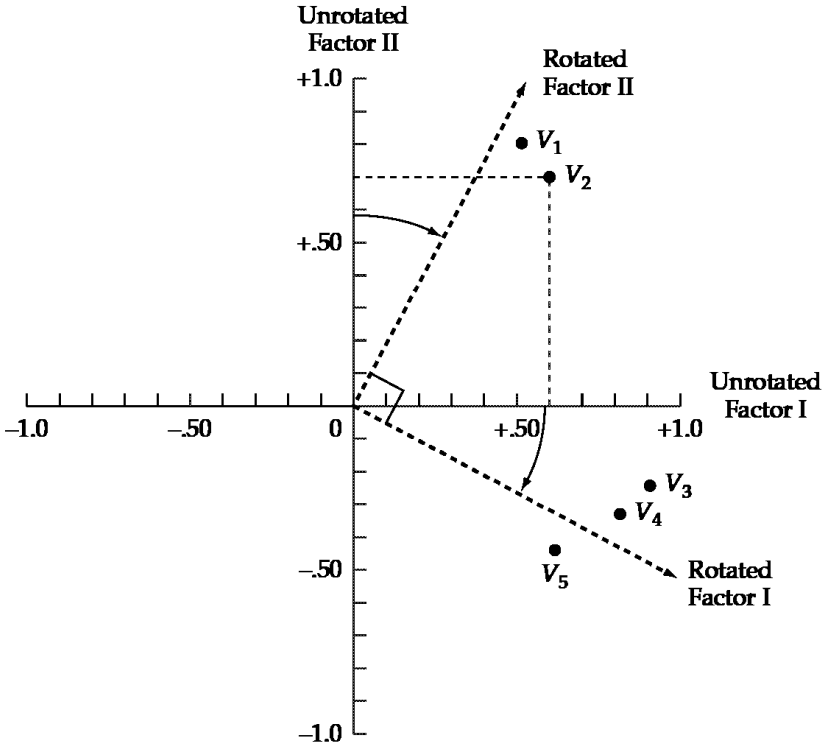


❒ Parsimony is important

- ➢ Have the most representative and parsimonious set of factors.

Factor Rotation

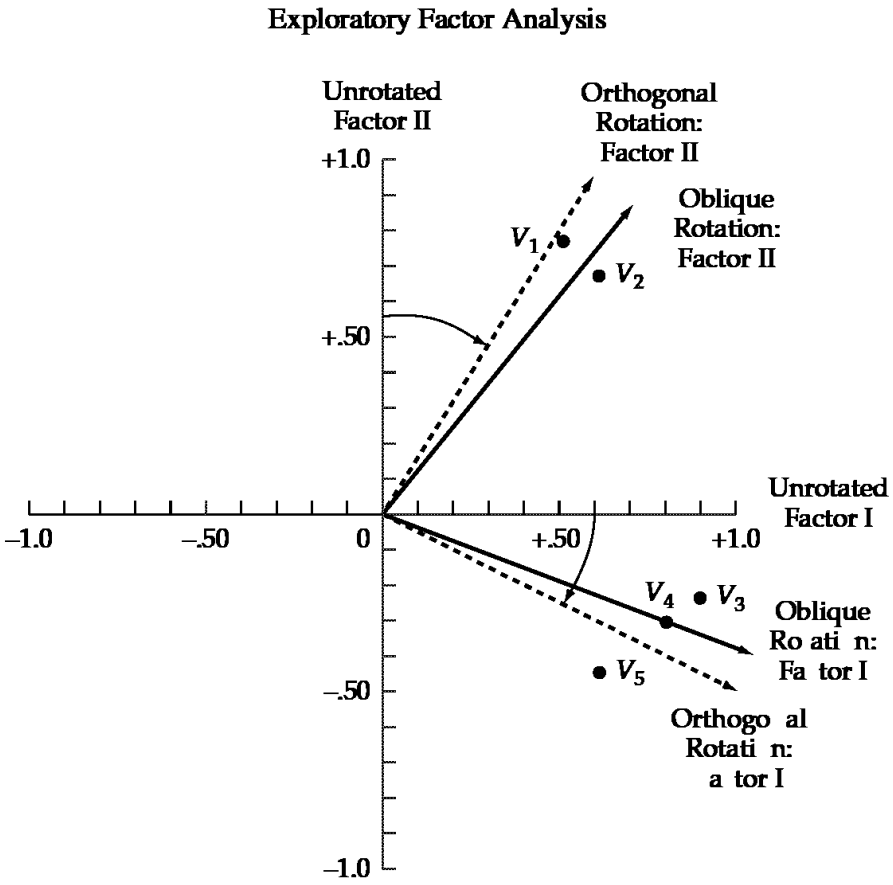
- Most of the times **rotation improves the factor interpretation**
- From a mathematical point of view the extracted factors are not unique.
- They could be translated in order to rotate the factor axis
  - Doesn´t change the data structure



- The **ultimate effect** of rotation is to **redistribute the variance from earlier factors to latter ones**
  - Simpler and more meaningful.

Factor Rotation

- **Orthogonal factor rotation**
  - Preserves orthogonality (not correlated)
  - It is the most widely used
- **Oblique factor rotation**
  - No restrictions as being orthogonal.
  - It also provides information about the extent to which the factors are correlated
  - Sometimes it is more difficult to interpret
  - Violates the initial idea of orthogonality between factors



Orthogonal Rotation Methods

- **Varimax**
  - Obtain a factor structure in which only one of the original variables is strongly associated with only one factor (the associations with other factors are much less strong).
  - Clearer separation of the factors.
  - Simplifies the factor matrix columns.
- **Quartimax**
  - Obtain a factor structure in which all variables have strong weights in one factor (general factor) and each variable has strong factor loadings in another factor (common factor) and small loadings in the other factors.
  - It assumes that the data structure could be explained by one general factor and one or more common factors. Simplifies the factor matrix rows.
- **Equimax**
  - Compromise between Varimax and Quartimax
  - It is not frequently used.

Oblique Rotation Methods

❒ Oblique rotation methods

- Similar to the orthogonal rotations, except that they allow correlations between the factors (e.g. **Oblimin in SPSS**).
- Care must be taken in the analysis.
- The nonorthogonality could be another way of becoming specific to the sample and non generalizable.
- Used when the goal is to obtain several theoretical meaningful factors.

Practical significance of factor loadings

- The **factor loading** is the correlation between each variable and a factor:
  - Loadings on the range of +-0,3 or +-0,4 are considered to meet the minimal level for interpretation of structure;
  - Loadings ≥+-0,5 are considered practically significant;
  - Loadings ≥+-0,7 are indicative of a well defined structure.

Practical significance

TABLE 2 Guidelines for Identifying Significant Factor Loadings Based on Sample Size

Factor Loading	Sample Size Needed for Significance[a]
.30	350
.35	250
.40	200
.45	150
.50	120
.55	100
.60	85
.65	70
.70	60
.75	50

[a] Significance is based on a .05 significance level (α), a power level of 80 percent, and standard errors assumed to be twice those of conventional correlation coefficients.

Source: Computations made with SOLO *Power Analysis,* BMDP Statistical Software, Inc., 1993.

Example

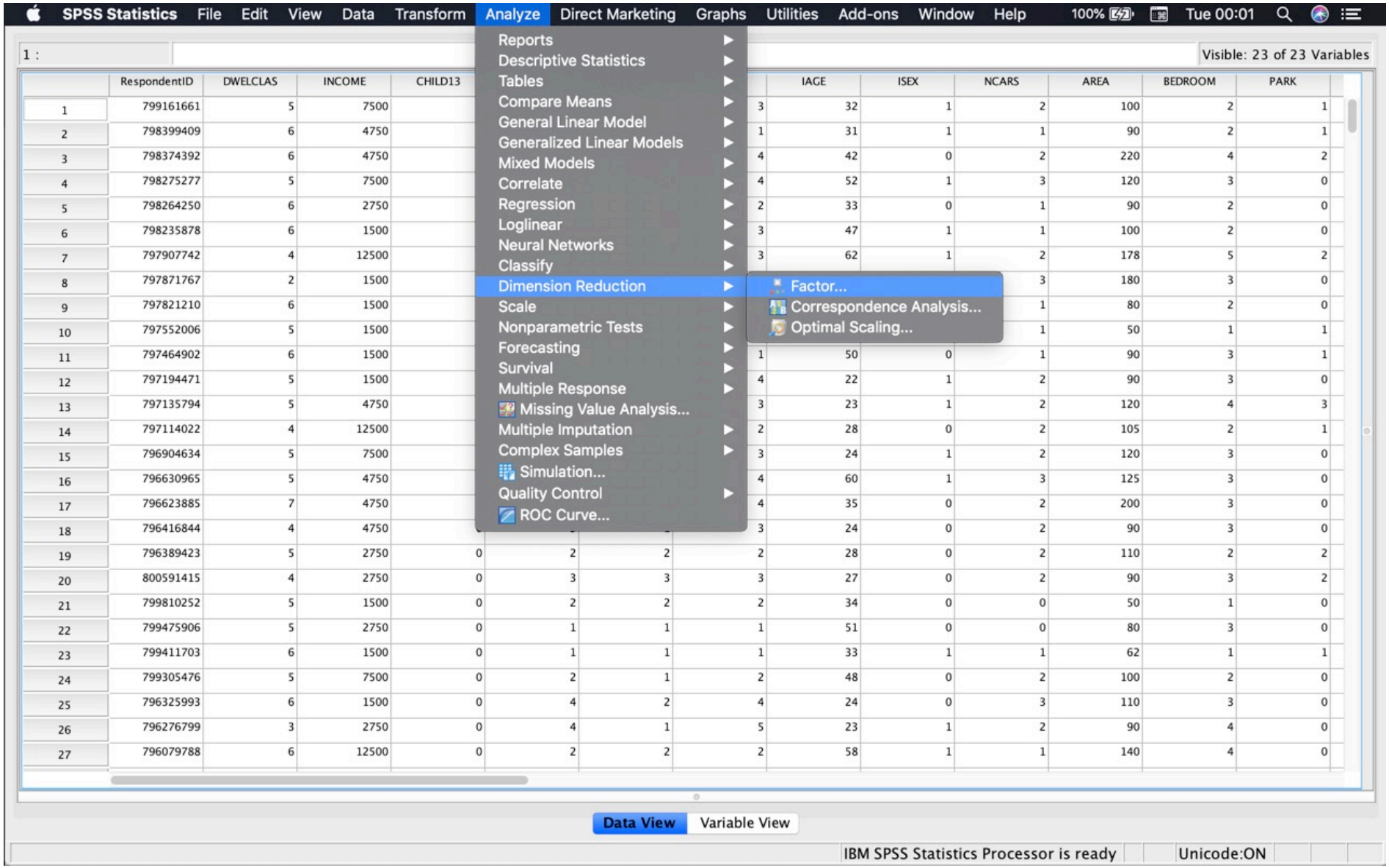
"Residential location satisfaction in the Lisbon metropolitan area"

- The database is from a study perform at IST:
  - Martínez, L. G., de Abreu e Silva, J., & Viegas, J. M. (2010). *Assessment of residential location satisfaction in the Lisbon metropolitan area, TRB* (No. 10-1161).
- Objective
  - The aim of this study was to examine the perception of households towards their residential location considering several land use and accessibility factors as well as household socioeconomic and attitudinal characteristics.

Example
Metadata

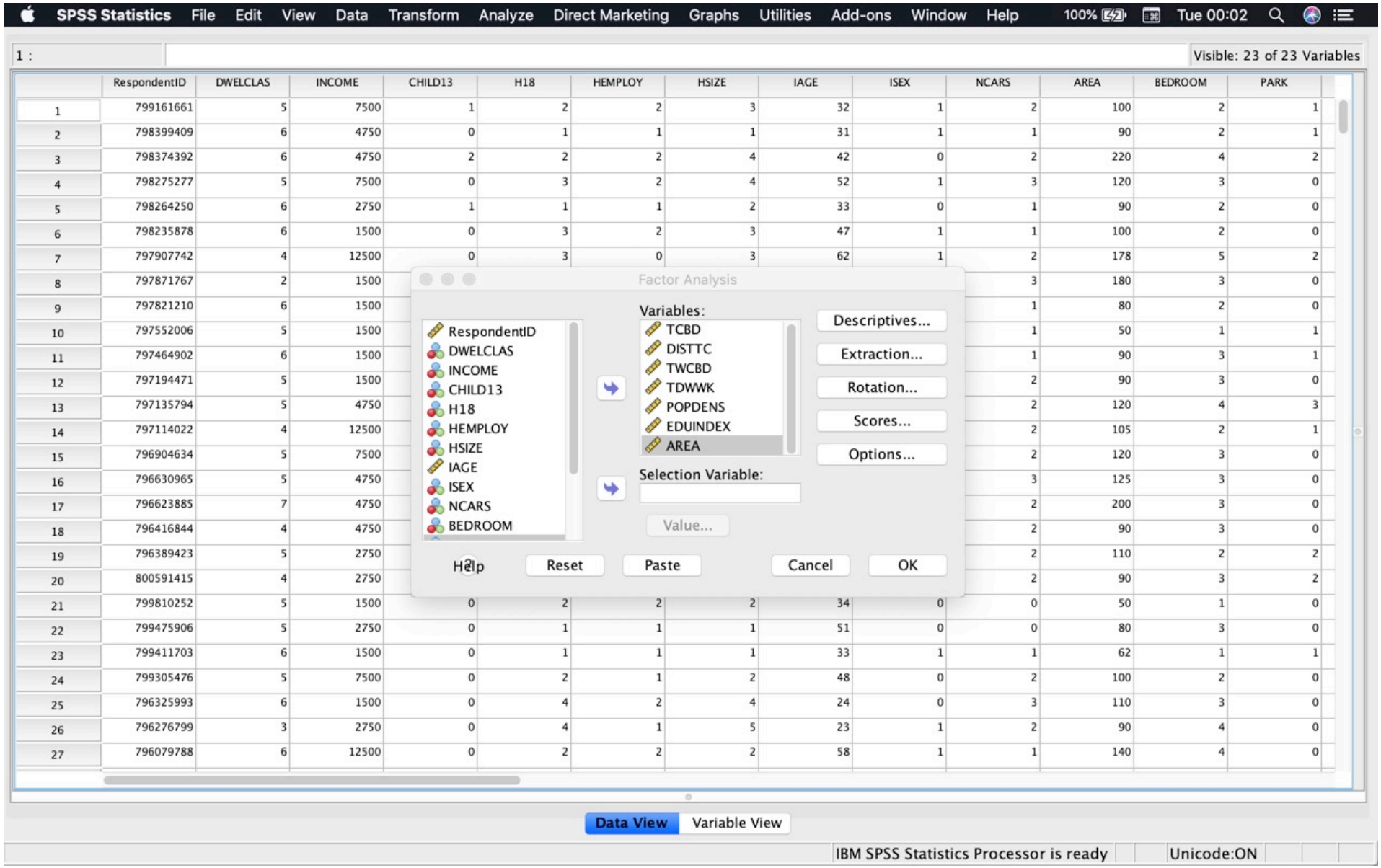
Symbol	Description
DWELCLAS	Classification of the Dwelling
INCOME	Income of the household
CHILD13	# Children <=13
H18	# Household members >=18
HEMPLOY	# Household members employed
HSIZE	Household size
IAGE	Sex of the respondent
ISEX	Age of the respondent
NCARS	# Car in the household
AREA	Area of the dwelling
BEDROOM	# Bedrooms in the dwelling
PARK	# Parking spaces in the dwelling
BEDSIZE	BEDROOM/HSIZE
PARKSIZE	PARK/NCARS
RAGE10	1 If Dwelling age <=10
TCBD	Private Car distance in time to CBD
DISTTC	Euclidean distance to heavy public transport system stops
TWCBD	Private car distance in time of work place to CBD
TDWWK	Private car distance in time of dwelling to work place
HEADH	1 If Head of the Household
POPDENS	Population density per hectare
EDUINDEX	Number of undergraduate persons/Population over 20 years old (500 meters)

Example

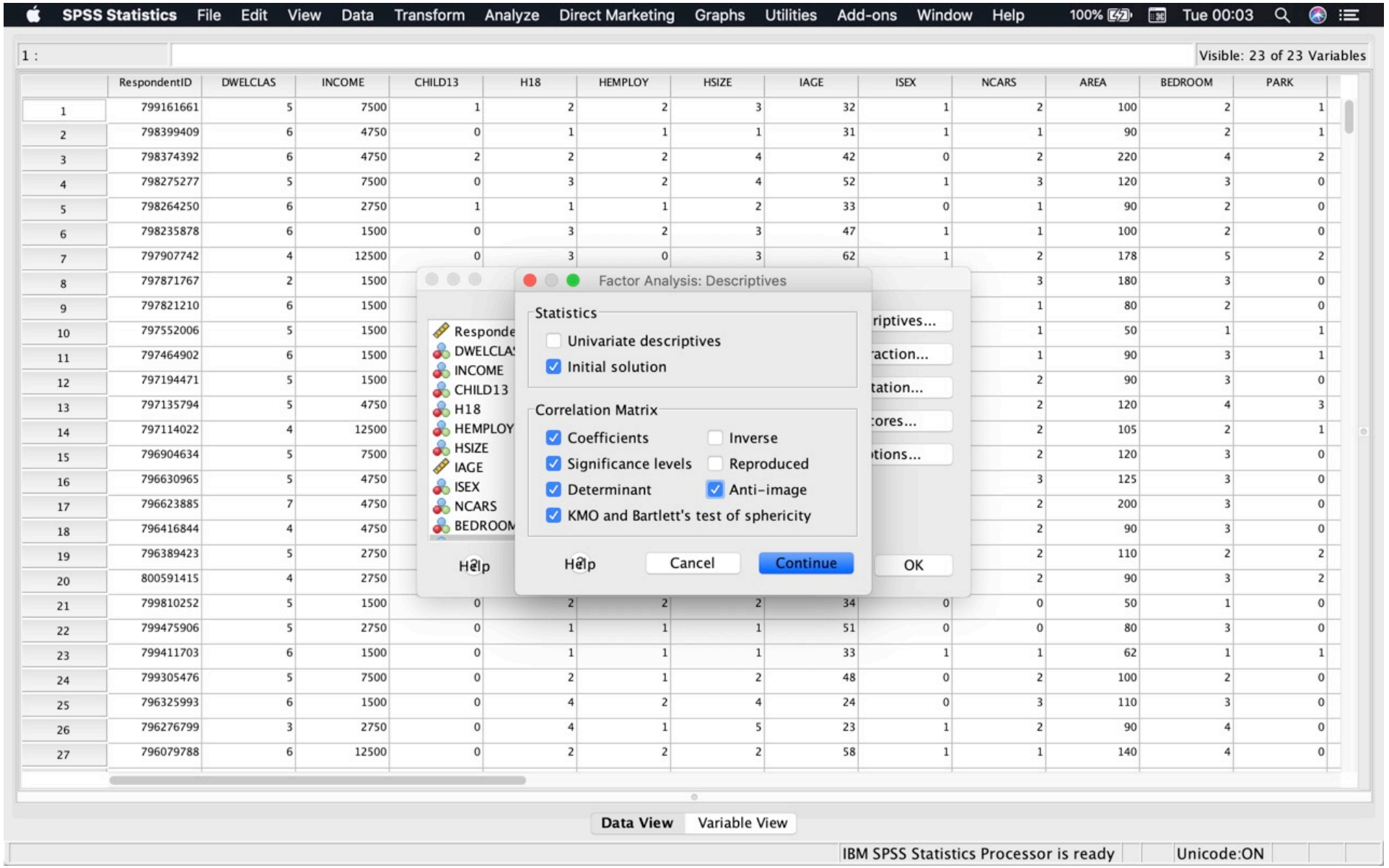


	RespondentID	DWELCLAS	INCOME	CHILD13				IAGE	ISEX	NCARS	AREA	BEDROOM	PARK
1	799161661	5	7500				3	32	1	2	100	2	1
2	798399409	6	4750				1	31	1	1	90	2	1
3	798374392	6	4750				4	42	0	2	220	4	2
4	798275277	5	7500				4	52	1	3	120	3	0
5	798264250	6	2750				2	33	0	1	90	2	0
6	798235878	6	1500				3	47	1	1	100	2	0
7	797907742	4	12500				3	62	1	2	178	5	2
8	797871767	2	1500							3	180	3	0
9	797821210	6	1500							1	80	2	0
10	797552006	5	1500							1	50	1	1
11	797464902	6	1500				1	50	0	1	90	3	1
12	797194471	5	1500				4	22	1	2	90	3	0
13	797135794	5	4750				3	23	1	2	120	4	3
14	797114022	4	12500				2	28	0	2	105	2	1
15	796904634	5	7500				3	24	1	2	120	3	0
16	796630965	5	4750				4	60	1	3	125	3	0
17	796623885	7	4750				4	35	0	2	200	3	0
18	796416844	4	4750				3	24	0	2	90	3	0
19	796389423	5	2750	0	2	2	2	28	0	2	110	2	2
20	800591415	4	2750	0	3	3	3	27	0	2	90	3	2
21	799810252	5	1500	0	2	2	2	34	0	0	50	1	0
22	799475906	5	2750	0	1	1	1	51	0	0	80	3	0
23	799411703	6	1500	0	1	1	1	33	1	1	62	1	1
24	799305476	5	7500	0	2	1	2	48	0	2	100	2	0
25	796325993	6	1500	0	4	2	4	24	0	3	110	3	0
26	796276799	3	2750	0	4	1	5	23	1	2	90	4	0
27	796079788	6	12500	0	2	2	2	58	1	1	140	4	0

Example

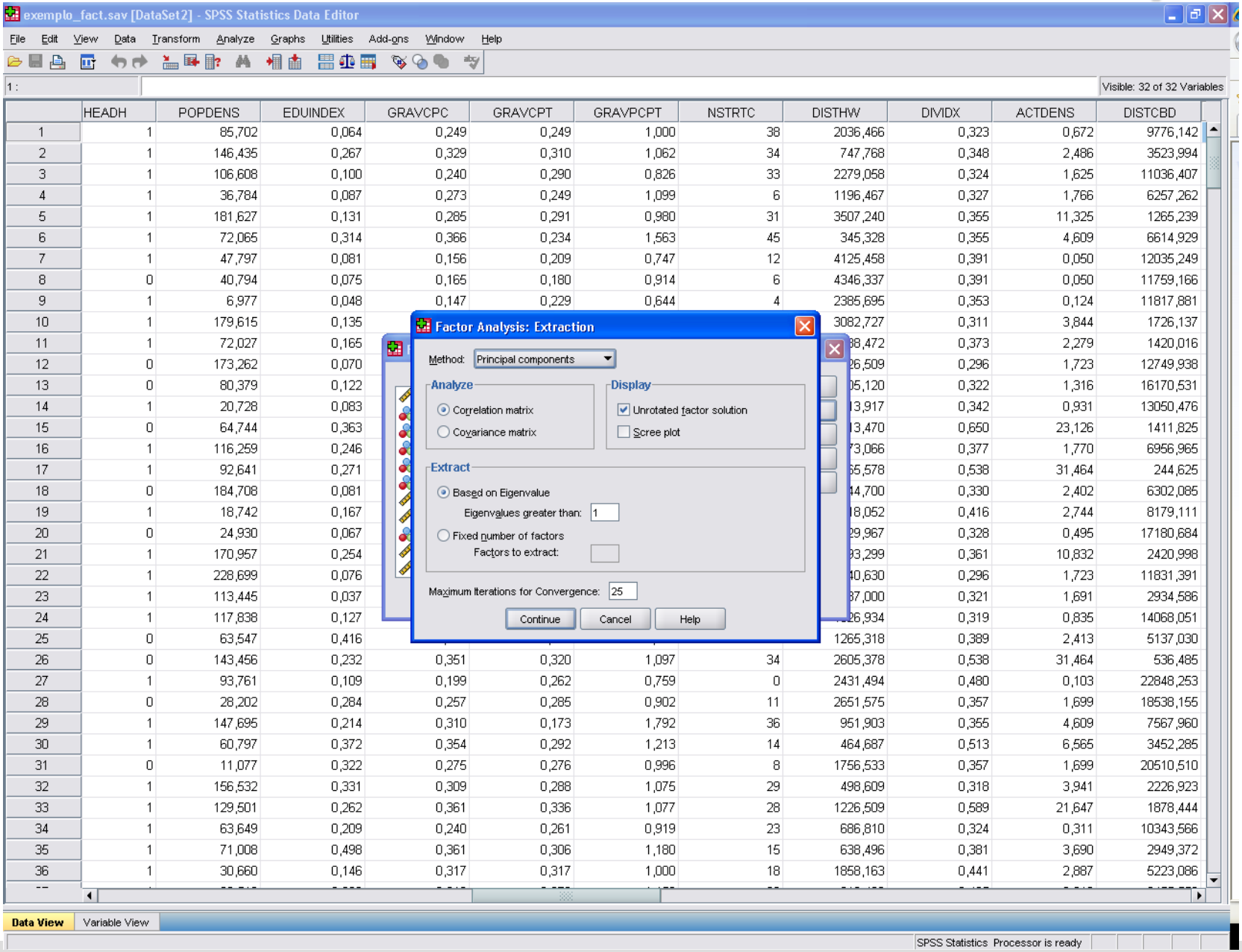

SPSS Statistics File Edit View Data Transform Analyze Direct Marketing Graphs Utilities Add-ons Window Help
Tue 00:02
1 :
Visible: 23 of 23 Variables
RespondentID DWELCLAS INCOME CHILD13 H18 HEMPLOY HSIZE IAGE ISEX NCARS AREA BEDROOM PARK
1 799161661 5 7500 1 2 2 3 32 1 2 100 2 1
2 798399409 6 4750 0 1 1 1 31 1 1 90 2 1
3 798374392 6 4750 2 2 2 4 42 0 2 220 4 2
4 798275277 5 7500 0 3 2 4 52 1 3 120 3 0
5 798264250 6 2750 1 1 1 2 33 0 1 90 2 0
6 798235878 6 1500 0 3 2 3 47 1 1 100 2 0
7 797907742 4 12500 0 3 0 3 62 1 2 178 5 2
8 797871767 2 1500 3 180 3 0
9 797821210 6 1500 1 80 2 0
10 797552006 5 1500 1 50 1 1
11 797464902 6 1500 1 90 3 1
12 797194471 5 1500 2 90 3 0
13 797135794 5 4750 2 120 4 3
14 797114022 4 12500 2 105 2 1
15 796904634 5 7500 2 120 3 0
16 796630965 5 4750 3 125 3 0
17 796623885 7 4750 2 200 3 0
18 796416844 4 4750 2 90 3 0
19 796389423 5 2750 2 110 2 2
20 800591415 4 2750 2 90 3 2
21 799810252 5 1500 0 2 2 2 34 0 0 50 1 0
22 799475906 5 2750 0 1 1 1 51 0 0 80 3 0
23 799411703 6 1500 0 1 1 1 33 1 1 62 1 1
24 799305476 5 7500 0 2 1 2 48 0 2 100 2 0
25 796325993 6 1500 0 4 2 4 24 0 3 110 3 0
26 796276799 3 2750 0 4 1 5 23 1 2 90 4 0
27 796079788 6 12500 0 2 2 2 58 1 1 140 4 0
Factor Analysis
RespondentID
DWELCLAS
INCOME
CHILD13
H18
HEMPLOY
HSIZE
IAGE
ISEX
NCARS
BEDROOM
Variables:
TCBD
DISTTC
TWCBD
TDWWK
POPDENS
EDUINDEX
AREA
Selection Variable:
Value...
Descriptives...
Extraction...
Rotation...
Scores...
Options...
Help
Reset
Paste
Cancel
OK
Data View
Variable View
IBM SPSS Statistics Processor is ready
Unicode:ON

Example



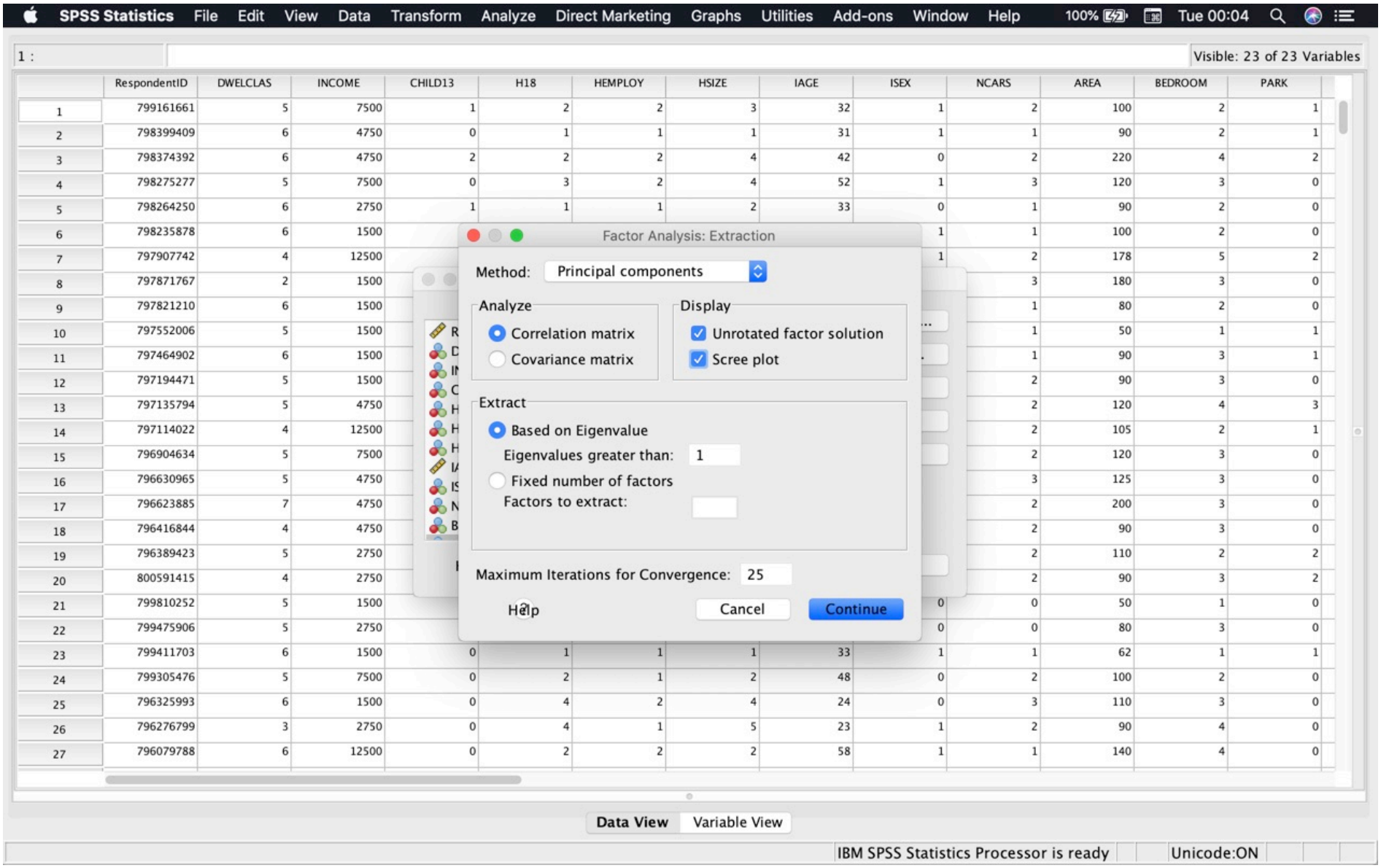
	RespondentID	DWELCLAS	INCOME	CHILD13	H18	HEMPLOY	HSIZE	IAGE	ISEX	NCARS	AREA	BEDROOM	PARK
1	799161661	5	7500	1	2	2	3	32	1	2	100	2	1
2	798399409	6	4750	0	1	1	1	31	1	1	90	2	1
3	798374392	6	4750	2	2	2	4	42	0	2	220	4	2
4	798275277	5	7500	0	3	2	4	52	1	3	120	3	0
5	798264250	6	2750	1	1	1	2	33	0	1	90	2	0
6	798235878	6	1500	0	3	2	3	47	1	1	100	2	0
7	797907742	4	12500	0	3	0	3	62	1	2	178	5	2
8	797871767	2	1500							3	180	3	0
9	797821210	6	1500							1	80	2	0
10	797552006	5	1500							1	50	1	1
11	797464902	6	1500							1	90	3	1
12	797194471	5	1500							2	90	3	0
13	797135794	5	4750							2	120	4	3
14	797114022	4	12500							2	105	2	1
15	796904634	5	7500							2	120	3	0
16	796630965	5	4750							3	125	3	0
17	796623885	7	4750							2	200	3	0
18	796416844	4	4750							2	90	3	0
19	796389423	5	2750							2	110	2	2
20	800591415	4	2750							2	90	3	2
21	799810252	5	1500	0	2	2	2	34	0	0	50	1	0
22	799475906	5	2750	0	1	1	1	51	0	0	80	3	0
23	799411703	6	1500	0	1	1	1	33	1	1	62	1	1
24	799305476	5	7500	0	2	1	2	48	0	2	100	2	0
25	796325993	6	1500	0	4	2	4	24	0	3	110	3	0
26	796276799	3	2750	0	4	1	5	23	1	2	90	4	0
27	796079788	6	12500	0	2	2	2	58	1	1	140	4	0

Example 1

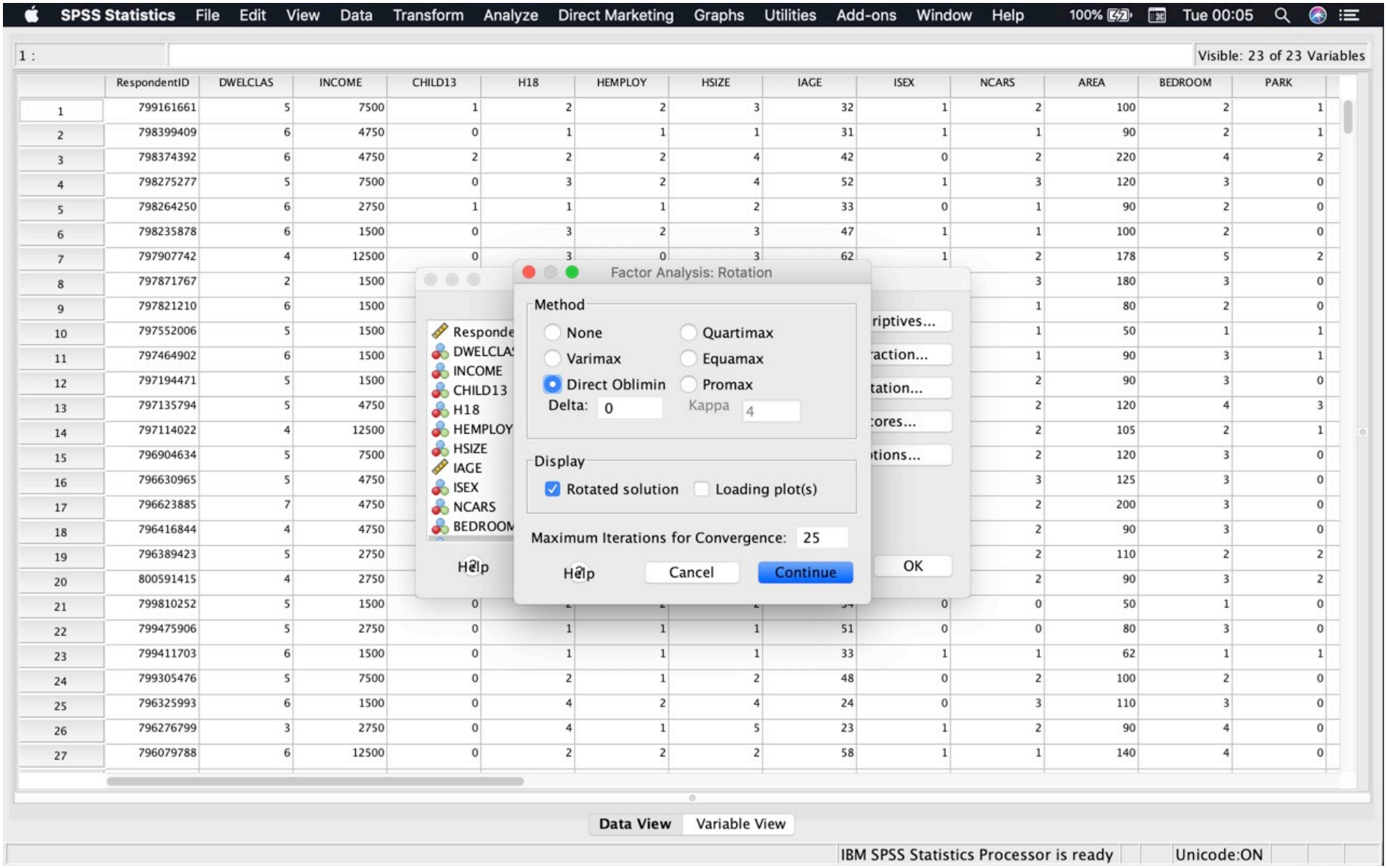


	HEADH	POPDENS	EDUINDEX	GRAVCPC	GRAVCPT	GRAVPCPT	NSTRTC	DISTHW	DIVIDX	ACTDENS	DISTCBD
1	1	85,702	0,064	0,249	0,249	1,000	38	2036,466	0,323	0,672	9776,142
2	1	146,435	0,267	0,329	0,310	1,062	34	747,768	0,348	2,486	3523,994
3	1	106,608	0,100	0,240	0,290	0,826	33	2279,058	0,324	1,625	11036,407
4	1	36,784	0,087	0,273	0,249	1,099	6	1196,467	0,327	1,766	6257,262
5	1	181,627	0,131	0,285	0,291	0,980	31	3507,240	0,355	11,325	1265,239
6	1	72,065	0,314	0,366	0,234	1,563	45	345,328	0,355	4,609	6614,929
7	1	47,797	0,081	0,156	0,209	0,747	12	4125,458	0,391	0,050	12035,249
8	0	40,794	0,075	0,165	0,180	0,914	6	4346,337	0,391	0,050	11759,166
9	1	6,977	0,048	0,147	0,229	0,644	4	2385,695	0,353	0,124	11817,881
10	1	179,615	0,135					3082,727	0,311	3,844	1726,137
11	1	72,027	0,165					38,472	0,373	2,279	1420,016
12	0	173,262	0,070					26,509	0,296	1,723	12749,938
13	0	80,379	0,122					05,120	0,322	1,316	16170,531
14	1	20,728	0,083					13,917	0,342	0,931	13050,476
15	0	64,744	0,363					13,470	0,650	23,126	1411,825
16	1	116,259	0,246					73,066	0,377	1,770	6956,965
17	1	92,641	0,271					35,578	0,538	31,464	244,625
18	0	184,708	0,081					44,700	0,330	2,402	6302,085
19	1	18,742	0,167					18,052	0,416	2,744	8179,111
20	0	24,930	0,067					29,967	0,328	0,495	17180,684
21	1	170,957	0,254					33,299	0,361	10,832	2420,998
22	1	228,699	0,076					40,630	0,296	1,723	11831,391
23	1	113,445	0,037					37,000	0,321	1,691	2934,586
24	1	117,838	0,127					26,934	0,319	0,835	14068,051
25	0	63,547	0,416					1265,318	0,389	2,413	5137,030
26	0	143,456	0,232	0,351	0,320	1,097	34	2605,378	0,538	31,464	536,485
27	1	93,761	0,109	0,199	0,262	0,759	0	2431,494	0,480	0,103	22848,253
28	0	28,202	0,284	0,257	0,285	0,902	11	2651,575	0,357	1,699	18538,155
29	1	147,695	0,214	0,310	0,173	1,792	36	951,903	0,355	4,609	7567,960
30	1	60,797	0,372	0,354	0,292	1,213	14	464,687	0,513	6,565	3452,285
31	0	11,077	0,322	0,275	0,276	0,996	8	1756,533	0,357	1,699	20510,510
32	1	156,532	0,331	0,309	0,288	1,075	29	498,609	0,318	3,941	2226,923
33	1	129,501	0,262	0,361	0,336	1,077	28	1226,509	0,589	21,647	1878,444
34	1	63,649	0,209	0,240	0,261	0,919	23	686,810	0,324	0,311	10343,566
35	1	71,008	0,498	0,361	0,306	1,180	15	638,496	0,381	3,690	2949,372
36	1	30,660	0,146	0,317	0,317	1,000	18	1858,163	0,441	2,887	5223,086

Example

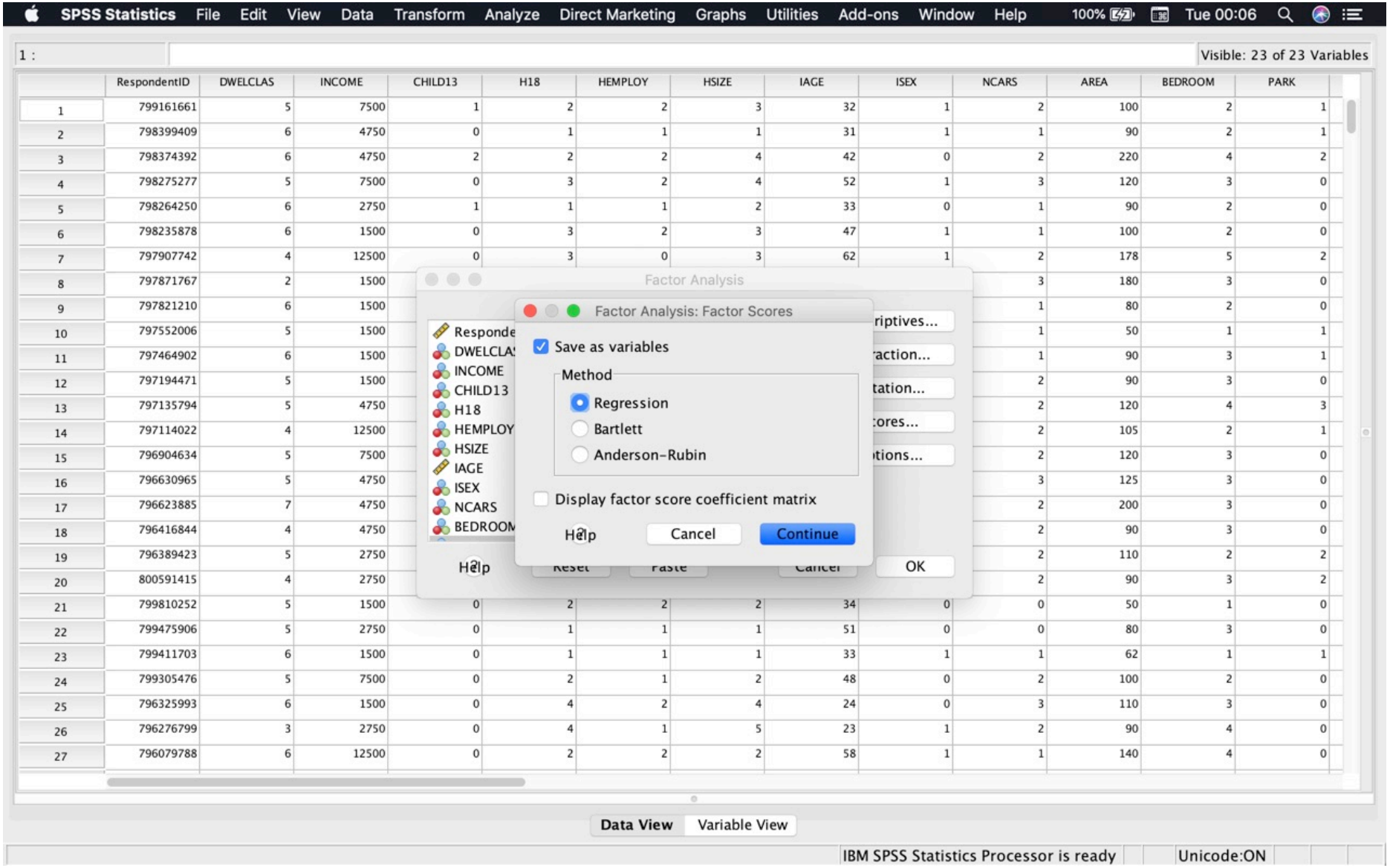

SPSS Statistics File Edit View Data Transform Analyze Direct Marketing Graphs Utilities Add-ons Window Help 100% Tue 00:04
1 :
Visible: 23 of 23 Variables
RespondentID DWELCLAS INCOME CHILD13 H18 HEMPLOY HSIZE IAGE ISEX NCARS AREA BEDROOM PARK
1 799161661 5 7500 1 2 2 3 32 1 2 100 2 1
2 798399409 6 4750 0 1 1 1 31 1 1 90 2 1
3 798374392 6 4750 2 2 2 4 42 0 2 220 4 2
4 798275277 5 7500 0 3 2 4 52 1 3 120 3 0
5 798264250 6 2750 1 1 1 2 33 0 1 90 2 0
6 798235878 6 1500 1 1 100 2 0
7 797907742 4 12500 1 2 178 5 2
8 797871767 2 1500 3 180 3 0
9 797821210 6 1500 1 80 2 0
10 797552006 5 1500 1 50 1 1
11 797464902 6 1500 1 90 3 1
12 797194471 5 1500 2 90 3 0
13 797135794 5 4750 2 120 4 3
14 797114022 4 12500 2 105 2 1
15 796904634 5 7500 2 120 3 0
16 796630965 5 4750 3 125 3 0
17 796623885 7 4750 2 200 3 0
18 796416844 4 4750 2 90 3 0
19 796389423 5 2750 2 110 2 2
20 800591415 4 2750 2 90 3 2
21 799810252 5 1500 0 0 50 1 0
22 799475906 5 2750 0 0 80 3 0
23 799411703 6 1500 0 1 1 1 33 1 1 62 1 1
24 799305476 5 7500 0 2 1 2 48 0 2 100 2 0
25 796325993 6 1500 0 4 2 4 24 0 3 110 3 0
26 796276799 3 2750 0 4 1 5 23 1 2 90 4 0
27 796079788 6 12500 0 2 2 2 58 1 1 140 4 0
Factor Analysis: Extraction
Method: Principal components
Analyze
Correlation matrix
Covariance matrix
Display
Unrotated factor solution
Scree plot
Extract
Based on Eigenvalue
Eigenvalues greater than: 1
Fixed number of factors
Factors to extract:
Maximum Iterations for Convergence: 25
Help
Cancel
Continue
Data View Variable View
IBM SPSS Statistics Processor is ready
Unicode:ON

Example



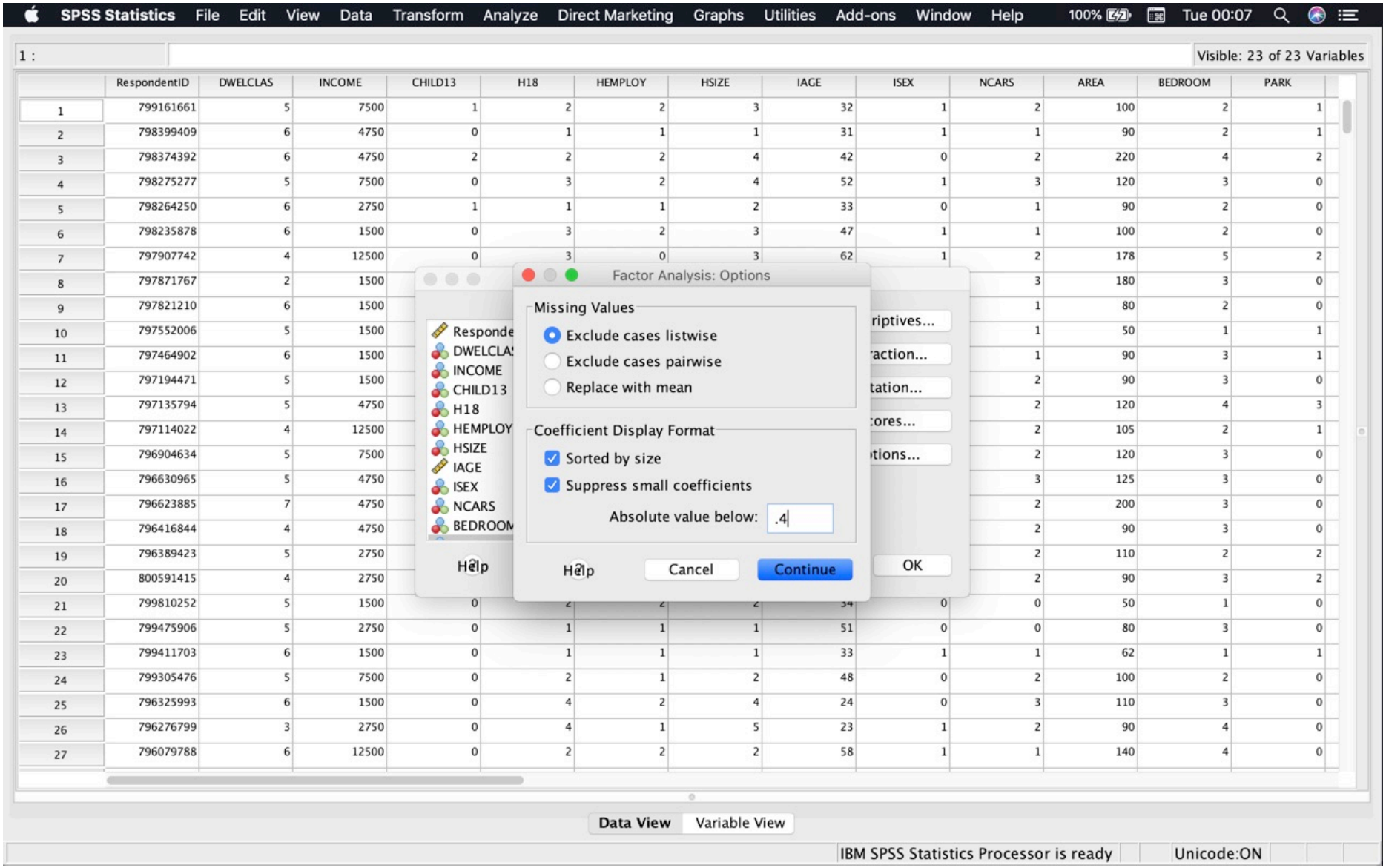
	RespondentID	DWELCLAS	INCOME	CHILD13	H18	HEMPLOY	HSIZE	IAGE	ISEX	NCARS	AREA	BEDROOM	PARK
1	799161661	5	7500	1	2	2	3	32	1	2	100	2	1
2	798399409	6	4750	0	1	1	1	31	1	1	90	2	1
3	798374392	6	4750	2	2	2	4	42	0	2	220	4	2
4	798275277	5	7500	0	3	2	4	52	1	3	120	3	0
5	798264250	6	2750	1	1	1	2	33	0	1	90	2	0
6	798235878	6	1500	0	3	2	3	47	1	1	100	2	0
7	797907742	4	12500	0	3	0	3	62	1	2	178	5	2
8	797871767	2	1500							3	180	3	0
9	797821210	6	1500							1	80	2	0
10	797552006	5	1500							1	50	1	1
11	797464902	6	1500							1	90	3	1
12	797194471	5	1500							2	90	3	0
13	797135794	5	4750							2	120	4	3
14	797114022	4	12500							2	105	2	1
15	796904634	5	7500							2	120	3	0
16	796630965	5	4750							3	125	3	0
17	796623885	7	4750							2	200	3	0
18	796416844	4	4750							2	90	3	0
19	796389423	5	2750							2	110	2	2
20	800591415	4	2750							2	90	3	2
21	799810252	5	1500	0					0	0	50	1	0
22	799475906	5	2750	0	1	1	1	51	0	0	80	3	0
23	799411703	6	1500	0	1	1	1	33	1	1	62	1	1
24	799305476	5	7500	0	2	1	2	48	0	2	100	2	0
25	796325993	6	1500	0	4	2	4	24	0	3	110	3	0
26	796276799	3	2750	0	4	1	5	23	1	2	90	4	0
27	796079788	6	12500	0	2	2	2	58	1	1	140	4	0

Example



	RespondentID	DWELCLAS	INCOME	CHILD13	H18	HEMPLOY	HSIZE	IAGE	ISEX	NCARS	AREA	BEDROOM	PARK
1	799161661	5	7500	1	2	2	3	32	1	2	100	2	1
2	798399409	6	4750	0	1	1	1	31	1	1	90	2	1
3	798374392	6	4750	2	2	2	4	42	0	2	220	4	2
4	798275277	5	7500	0	3	2	4	52	1	3	120	3	0
5	798264250	6	2750	1	1	1	2	33	0	1	90	2	0
6	798235878	6	1500	0	3	2	3	47	1	1	100	2	0
7	797907742	4	12500	0	3	0	3	62	1	2	178	5	2
8	797871767	2	1500							3	180	3	0
9	797821210	6	1500							1	80	2	0
10	797552006	5	1500							1	50	1	1
11	797464902	6	1500							1	90	3	1
12	797194471	5	1500							2	90	3	0
13	797135794	5	4750							2	120	4	3
14	797114022	4	12500							2	105	2	1
15	796904634	5	7500							2	120	3	0
16	796630965	5	4750							3	125	3	0
17	796623885	7	4750							2	200	3	0
18	796416844	4	4750							2	90	3	0
19	796389423	5	2750							2	110	2	2
20	800591415	4	2750							2	90	3	2
21	799810252	5	1500	0	2	2	2	34	0	0	50	1	0
22	799475906	5	2750	0	1	1	1	51	0	0	80	3	0
23	799411703	6	1500	0	1	1	1	33	1	1	62	1	1
24	799305476	5	7500	0	2	1	2	48	0	2	100	2	0
25	796325993	6	1500	0	4	2	4	24	0	3	110	3	0
26	796276799	3	2750	0	4	1	5	23	1	2	90	4	0
27	796079788	6	12500	0	2	2	2	58	1	1	140	4	0

Example



	RespondentID	DWELCLAS	INCOME	CHILD13	H18	HEMPLOY	HSIZE	IAGE	ISEX	NCARS	AREA	BEDROOM	PARK
1	799161661	5	7500	1	2	2	3	32	1	2	100	2	1
2	798399409	6	4750	0	1	1	1	31	1	1	90	2	1
3	798374392	6	4750	2	2	2	4	42	0	2	220	4	2
4	798275277	5	7500	0	3	2	4	52	1	3	120	3	0
5	798264250	6	2750	1	1	1	2	33	0	1	90	2	0
6	798235878	6	1500	0	3	2	3	47	1	1	100	2	0
7	797907742	4	12500	0	3	0	3	62	1	2	178	5	2
8	797871767	2	1500							3	180	3	0
9	797821210	6	1500							1	80	2	0
10	797552006	5	1500							1	50	1	1
11	797464902	6	1500							1	90	3	1
12	797194471	5	1500							2	90	3	0
13	797135794	5	4750							2	120	4	3
14	797114022	4	12500							2	105	2	1
15	796904634	5	7500							2	120	3	0
16	796630965	5	4750							3	125	3	0
17	796623885	7	4750							2	200	3	0
18	796416844	4	4750							2	90	3	0
19	796389423	5	2750							2	110	2	2
20	800591415	4	2750							2	90	3	2
21	799810252	5	1500	0	2	2	2	34	0	0	50	1	0
22	799475906	5	2750	0	1	1	1	51	0	0	80	3	0
23	799411703	6	1500	0	1	1	1	33	1	1	62	1	1
24	799305476	5	7500	0	2	1	2	48	0	2	100	2	0
25	796325993	6	1500	0	4	2	4	24	0	3	110	3	0
26	796276799	3	2750	0	4	1	5	23	1	2	90	4	0
27	796079788	6	12500	0	2	2	2	58	1	1	140	4	0

Example

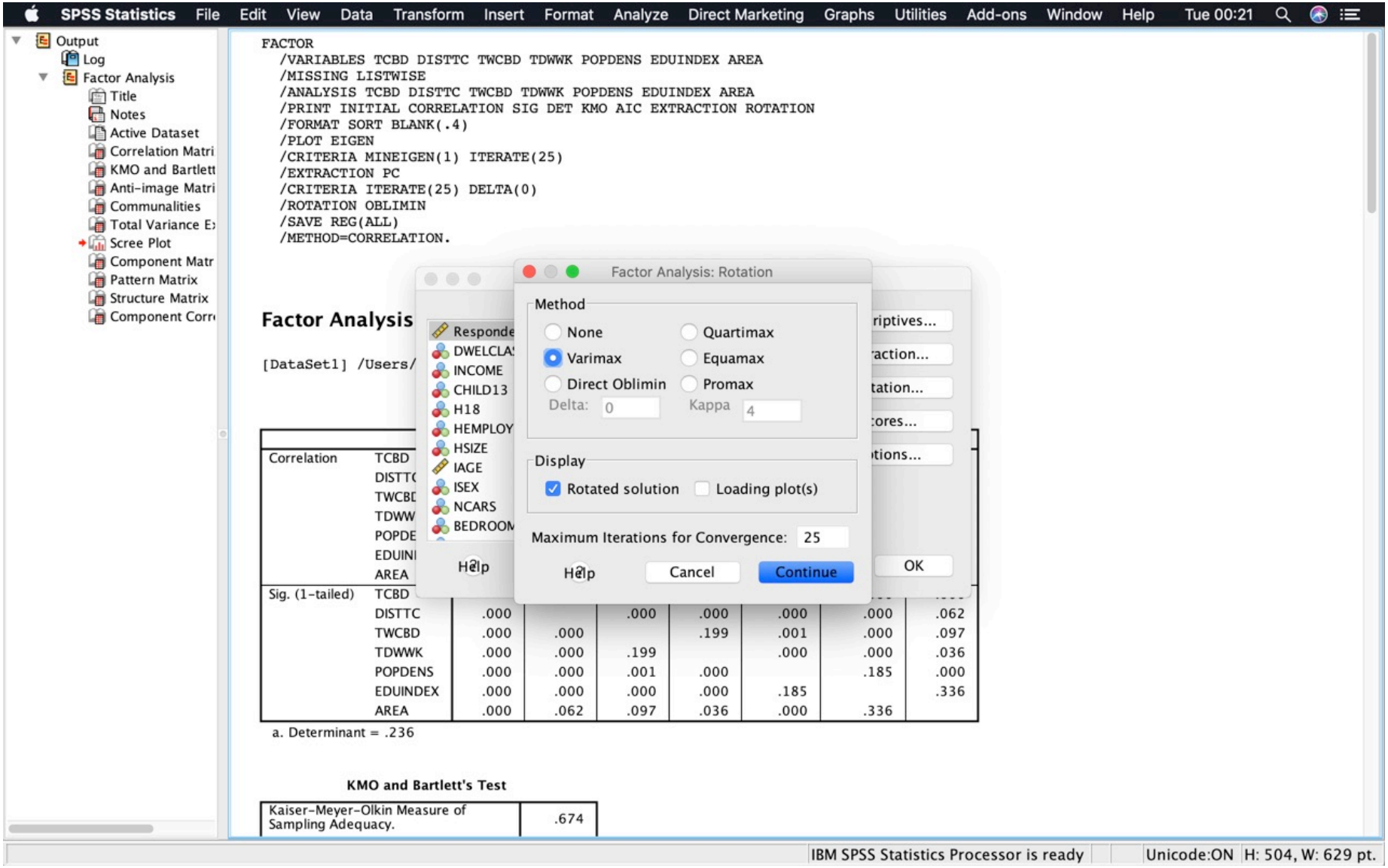
- Low correlation between factors indicates that they are independent (orthogonal)
  - Orthogonal rotation is advisable
  - Refer to the last table in the SPSS output

Component Correlation Matrix

Component	1	2	3
1	1.000	.068	.229
2	.068	1.000	.026
3	.229	.026	1.000

Extraction Method: Principal Component Analysis.
Rotation Method: Oblimin with Kaiser Normalization.

Example



		TCBD	DISTTC	TWCBD	TDWWK	POPDENS	EDUINDEX	AREA
Correlation	TCBD		.000	.000	.000	.000	.000	.062
	DISTTC	.000		.000	.199	.001	.000	.097
	TWCBD	.000	.000			.000	.000	.036
	TDWWK	.000	.000	.199			.185	.000
	POPDENS	.000	.000	.001	.000			.336
	EDUINDEX	.000	.000	.000	.000	.185		
	AREA	.000	.062	.097	.036	.000	.336	
Sig. (1-tailed)	TCBD							

Example

Correlation Matrix[a]

		TCBD	DISTTC	TWCBD	TDWWK	POPDENS	EDUINDEX	AREA
Correlation	TCBD	1.000	.531	.433	.455	-.386	-.428	.168
	DISTTC	.531	1.000	.163	.334	-.442	-.288	.071
	TWCBD	.433	.163	1.000	-.039	-.145	-.222	.060
	TDWWK	.455	.334	-.039	1.000	-.238	-.259	.083
	POPDENS	-.386	-.442	-.145	-.238	1.000	.041	-.164
	EDUINDEX	-.428	-.288	-.222	-.259	.041	1.000	.020
	AREA	.168	.071	.060	.083	-.164	.020	1.000

KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.674
Bartlett's Test of Sphericity	Approx. Chi-Square	672.580
	df	21
	Sig.	.000

- AREA is not correlated with the remaining variables selected, such as EUINDEX (although more correlated than the others)
- KMO mean that the recommendation for FA is Mediocre/Middling
- Bartlett's Test of Sphericity is significant

Example

Anti-image Matrices

		TCBD	DISTTC	TWCBD	TDWWK	POPDENS	EDUINDEX	AREA
Anti-image Covariance	TCBD	.436	-.162	-.245	-.206	.098	.155	-.084
	DISTTC	-.162	.631	.044	-.048	.216	.086	.038
	TWCBD	-.245	.044	.736	.211	.020	.060	.007
	TDWWK	-.206	-.048	.211	.709	.043	.074	-.007
	POPDENS	.098	.216	.020	.043	.736	.133	.088
	EDUINDEX	.155	.086	.060	.074	.133	.770	-.069
	AREA	-.084	.038	.007	-.007	.088	-.069	.950
Anti-image Correlation	TCBD	.658[a]	.309	-.433	-.371	.172	.267	-.130
	DISTTC	-.309	.760[a]	.064	-.071	.317	.123	.049
	TWCBD	-.433	.064	.505[a]	.292	.027	.079	.009
	TDWWK	-.371	-.071	.292	.651[a]	.059	.100	-.009
	POPDENS	.172	.317	.027	.059	.719[a]	.177	.105
	EDUINDEX	.267	.123	.079	.100	.177	.732[a]	.081
	AREA	-.130	.049	.009	-.009	.105	-.081	.658[a]

a. Measures of Sampling Adequacy(MSA)

- Diagonal values indicate some sample size problems for TWCBD as it is below the threshold of 0,6 and thus Mediocre.

Example

Communalities

	Initial	Extraction
TCBD	1.000	.773
DISTTC	1.000	.591
TWCBD	1.000	.857
TDWWK	1.000	.680
POPDENS	1.000	.586
EDUINDEX	1.000	.638
AREA	1.000	.603

Extraction Method: Principal Component Analysis.

❒ Communalities

- ➢ They represent the amount of variance accounted for by the factor analysis
- ➢ At least half of the variance of each variable should be considered before inclusion in FA
- ➢ This means that variables with communalities smaller than 0,5 should be excluded

Example

Total Variance Explained

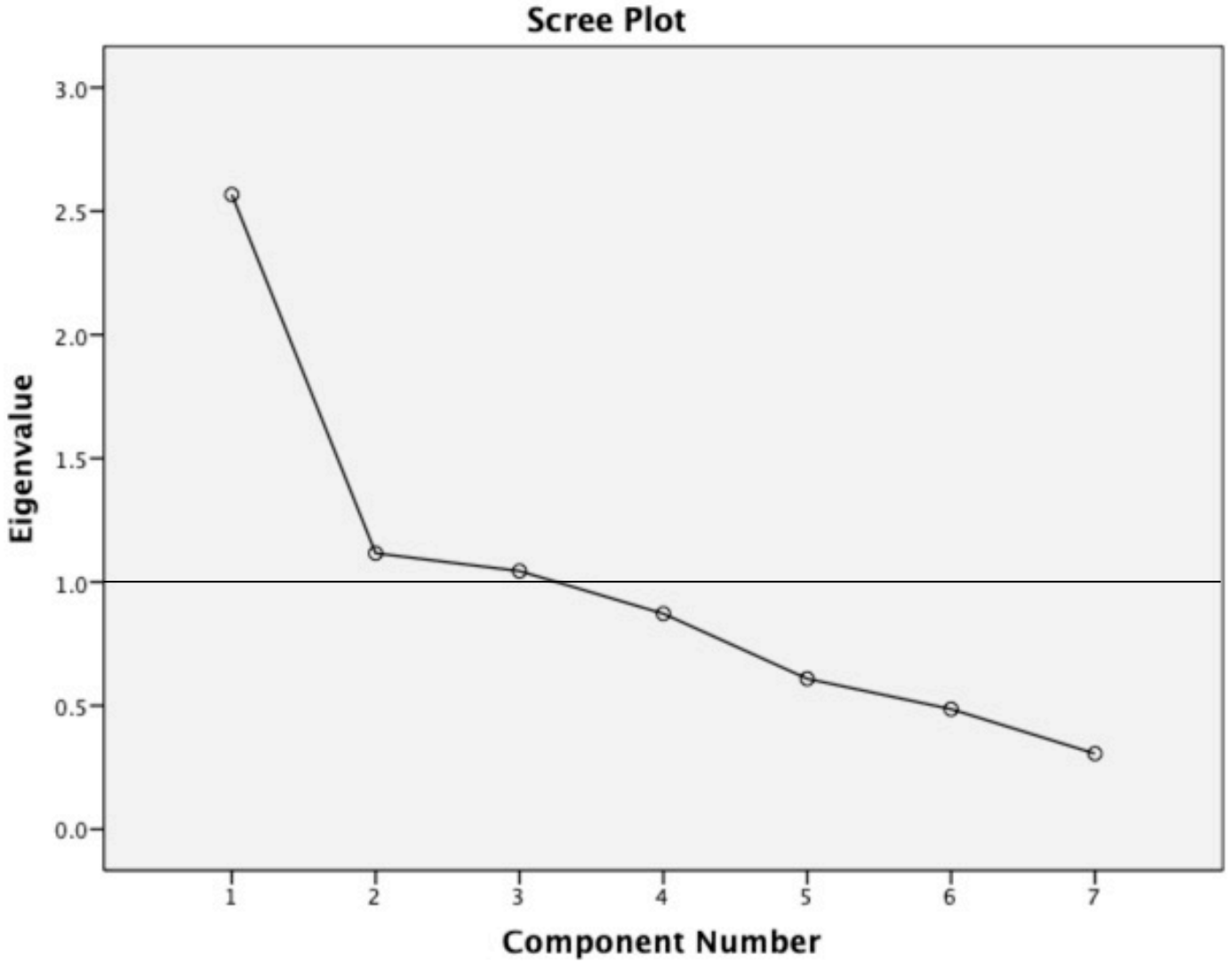
Component	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings[a]
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total
1	2.567	36.673	36.673	2.567	36.673	36.673	2.334
2	1.116	15.947	52.619	1.116	15.947	52.619	1.148
3	1.044	14.918	67.538	1.044	14.918	67.538	1.562
4	.872	12.456	79.993				
5	.609	8.696	88.690				
6	.486	6.941	95.631				
7	.306	4.369	100.000				

Extraction Method: Principal Component Analysis.

a. When components are correlated, sums of squared loadings cannot be added to obtain a total variance.

- Total variance explained is 67,5%, if we consider components with eigenvalues > 1
- Is it good?

Example


Scree Plot
Eigenvalue
3.0
2.5
2.0
1.5
1.0
0.5
0.0
1
2
3
4
5
6
7
Component Number

Example

Component Matrix[a]

	Component		
	1	2	3
TCBD	.868		
DISTTC	.746		
POPDENS	-.595	-.463	
TDWWK	.593		-.534
EDUINDEX	-.548	.536	
AREA		.542	.506
TWCBD	.444	-.498	.642

Extraction Method: Principal Component Analysis.

a. 3 components extracted.

Rotated Component Matrix[a]

	Component		
	1	2	3
TDWWK	.805		
DISTTC	.723		
TCBD	.686	.531	
EDUINDEX	-.495	-.462	.423
TWCBD		.919	
AREA			.772
POPDENS	-.510		-.566

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.

a. Rotation converged in 6 iterations.

Example

Component Transformation Matrix

Component	1	2	3
1	.851	.477	.222
2	.155	-.630	.761
3	-.503	.613	.610

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.

- This is the matrix that transforms the unrotated solution in the rotated component matrix (by matrix multiplication)

Underlying latent structure in the data?

Rotated Component Matrix[a]

	Component		
	1	2	3
TDWWK	.805		
DISTTC	.723		
TCBD	.686	.531	
EDUINDEX	-.495	-.462	.423
TWCBD		.919	
AREA			.772
POPDENS	-.510		-.566

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.

a. Rotation converged in 6 iterations.

TDWWK	Private car distance in time of dwelling to work place
DISTTC	Euclidean distance to heavy public transport system stops
TCBD	Private Car distance in time to CBD
EDUINDEX	Number of undergraduate persons/Population over 20 years old
TWCBD	Private car distance in time of work place to CBD
AREA	Area of the dwelling
POPDENS	Population density per hectare

Test without the variable AREA?

Recommended Readings

- Hair, Joseph P. et al (1995) “Multivariate Data Analysis with Readings”, Fourth Edition, Prentice Hall - Chapter 2

- Maroco, João (2003) “Análise Estatística com utilização do SPSS”, Ed. Sílabo– Capítulo 10