

Phd Program in Transportation

Transport Demand Modeling

Filipe Moura

Session 4

Factor Analysis

(Slides prepared by João de Abreu e Silva)

Highlights of Factorial Analysis

- Exploratory technique aimed at **defining the underlying structure among a group of interrelated variables.**
- 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 factors**

Uses of factor analysis

□ Data summarization

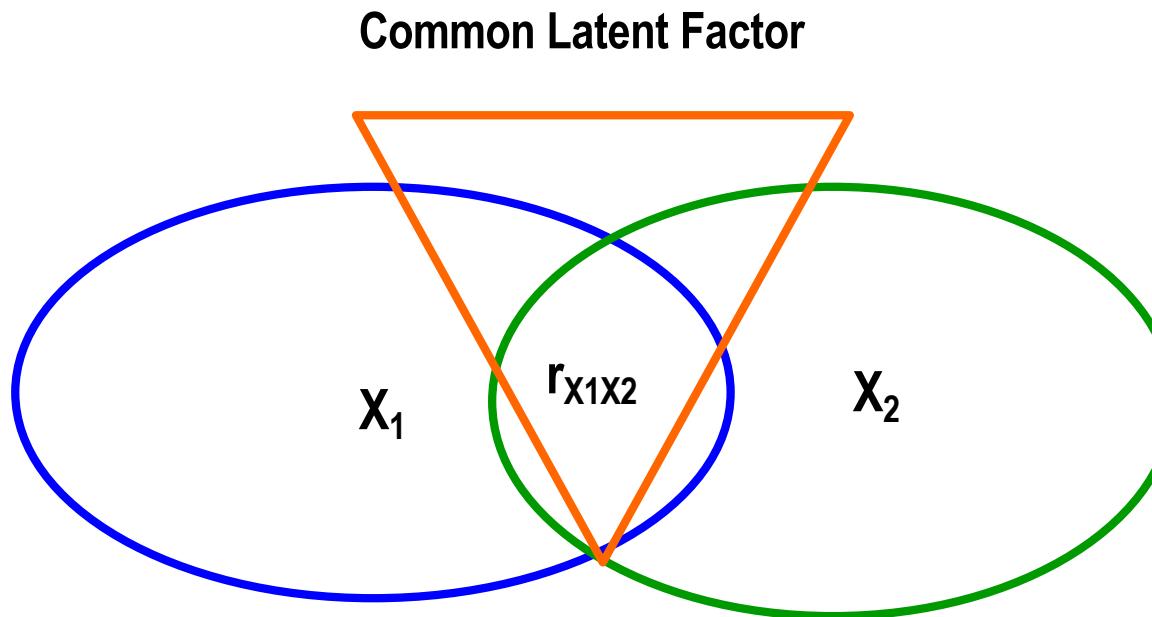
- Define a **small number of factors that adequately represent the original set of variables.**
- Each dependent variable is a function of an **underlying and latent set of factors.**
- **Each variable is predicted by all the factors** (and indirectly by all the other variables)

□ Data reduction

- **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

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

, 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 (IV)

$$x_1 = a_{11}f_1 + a_{12}f_2 + e_1, \quad \boxed{a_{11}f_1 + a_{12}f_2} - \text{Communalities}$$

$$x_2 = a_{21}f_1 + a_{22}f_2 + e_2, \quad \boxed{a_{21}f_1 + a_{22}f_2} - \text{Specific and errors}$$

$$x_3 = a_{31}f_1 + a_{32}f_2 + e_3.$$

- 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 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 ($\mu=0$ and $\sigma=I$)
 - η is the vector of specific factors e_p ($\mu=0$ 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

$$\Pi = \boxed{\Lambda \Lambda^t} + \Psi$$

- 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)

- The values in the **diagonal of the Anti-Image Correlation Matrix** are a Measure of Sampling Adequacy.
 - They could be interpreted in a way similar to KMO (i.e., **they should be at least bigger than 0,6**)
- **KMO (Kaiser-Meyer-Olkin)**
 - **Measure of homogeneity**, which compares simple with partial correlations observed between variables

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

, where

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

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
$\leq 0,5$	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**, therefor get a **Chi-square test above threshold**
- It is sensible to sample size (more sensible in detecting correlations)

$$H_0: \Pi = I$$

$$H_\alpha: \Pi \neq I$$

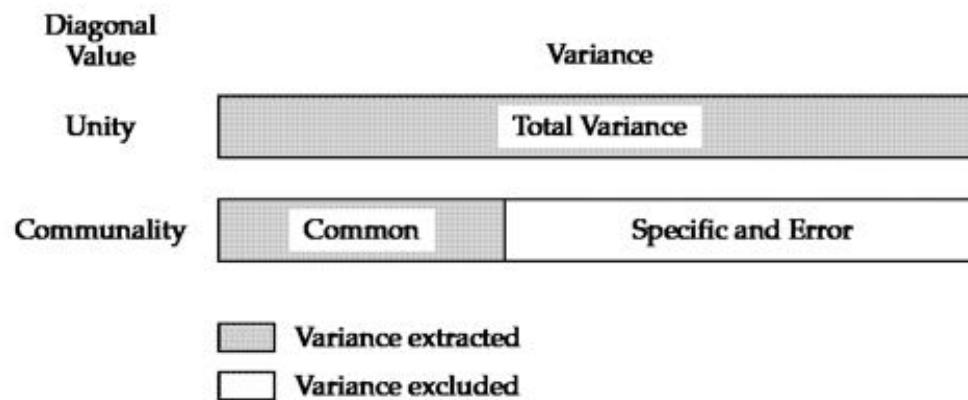
$$\chi^2 = -\left(N - 2 - \frac{2p + 5}{6}\right) \ln|R|$$

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

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 – Eigenvalue >1 .

➤ **A Priori criterion**

- Define *a priori* the number of factors to be extracted (testing an hypothesis about the number of factors).

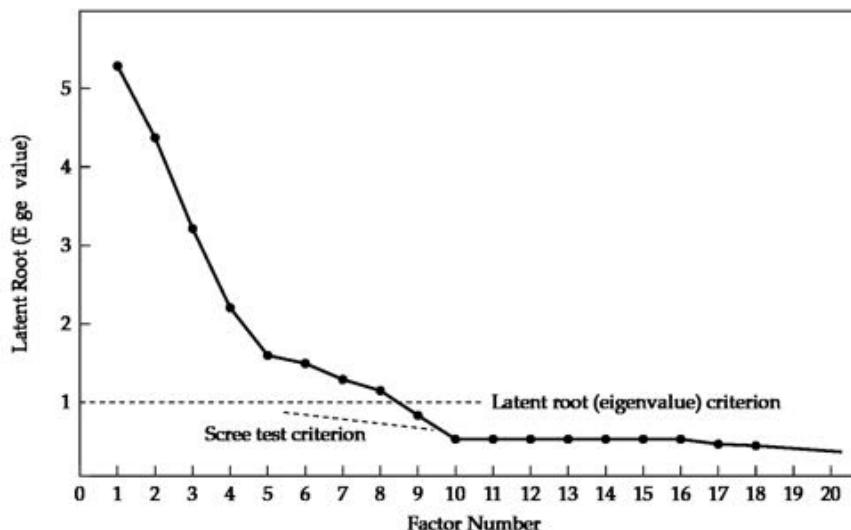
➤ **Percentage of Variance criterion**

- 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 inflection 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.

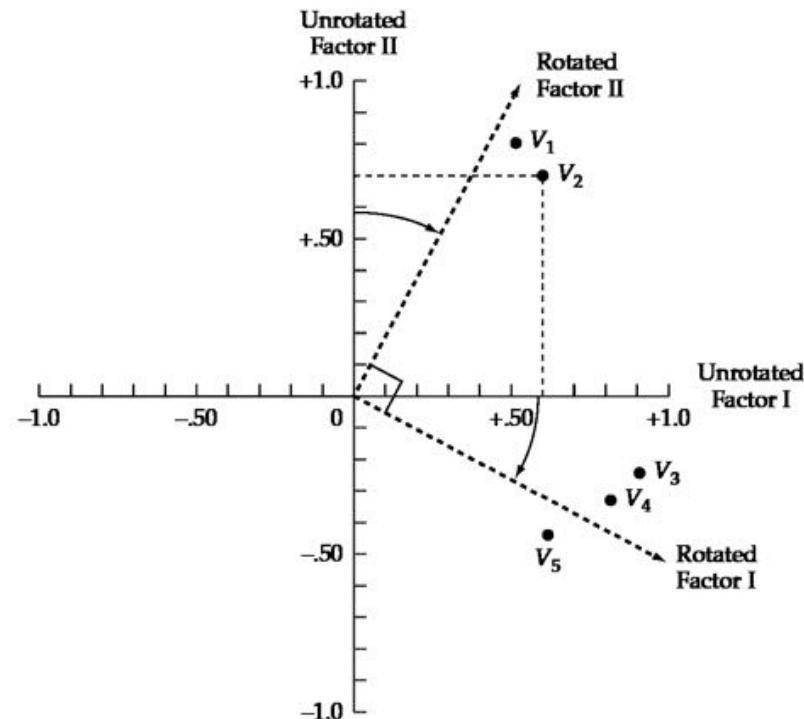


□ Parsimony is important

- Have the most representative and parsimonious set of factors.

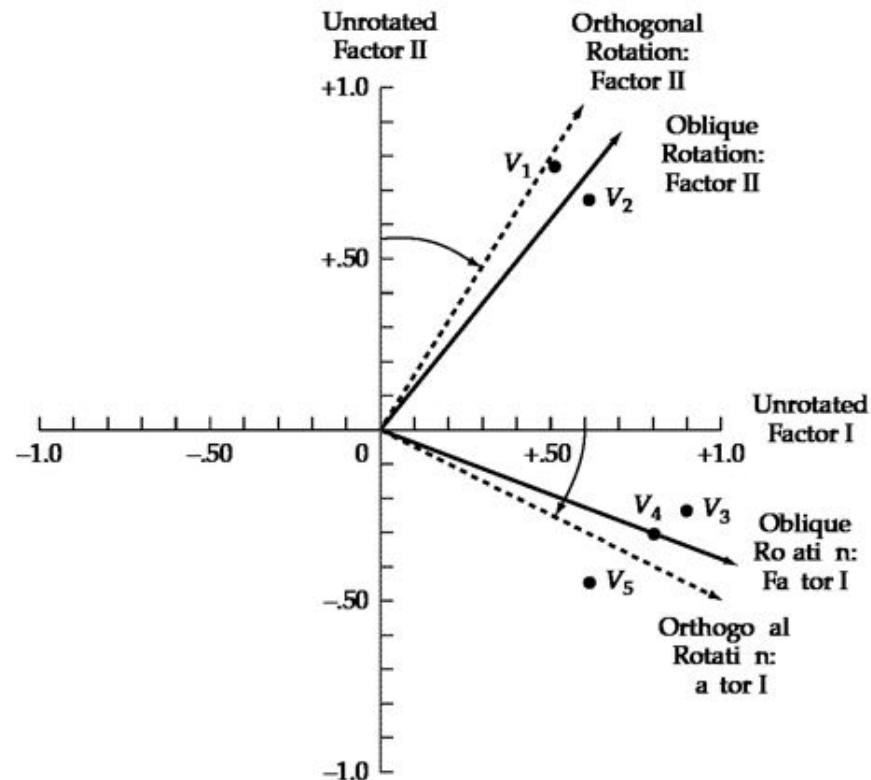
Factor Rotation

- 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

Exploratory Factor Analysis



□ 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 $\pm 0,3$ or $\pm 0,4$ are considered to meet the minimal level for interpretation of structure;
 - Loadings $\geq \pm 0,5$ are considered practically significant;
 - Loadings $\geq \pm 0,7$ 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

SPSS Statistics File Edit View Data Transform Analyze Direct Marketing Graphs Utilities Add-ons Window Help 100% 848 Tue 00:01

Data View Variable View IBM SPSS Statistics Processor is ready Unicode ON

Visible: 23 of 23 Variables

	RespondentID	DM1CLAS	INCOME	CHILD1	AGE	SEX	NCARS	AREA	BEDROOM	PARK	
1	799061660	5	7500		3	12	1	2	100	2	1
2	798399409	6	4750		4	33	1	1	80	2	0
3	798376392	6	4750		4	42	0	2	220	4	2
4	798275277	5	7500		4	52	1	3	120	3	0
5	798264258	6	2750		2	33	0	1	90	2	0
6	798215878	6	15000		3	49	1	1	100	2	0
7	797907742	4	12500		3	62	1	2	170	5	2
8	797871767	2	15000		1	180	3	0			
9	797821210	6	15000		4	80	2	0			
10	797512006	5	15000		1	50	1	1			
11	797464902	6	15000		4	22	1	2	90	3	0
12	797294471	5	15000		3	23	1	2	120	4	0
13	797115794	5	4750		2	28	0	2	100	2	0
14	797114622	4	12500		3	24	1	2	120	3	0
15	796904634	5	7500		4	60	1	3	125	3	0
16	796430965	5	4750		4	35	0	2	200	3	0
17	796423885	7	4750		3	24	0	2	90	3	0
18	796415844	4	4750		1	28	0	1	150	2	2
19	796399423	5	2750	0	2	2	1	1	100	2	0
20	800591415	4	2750	0	3	3	1	2	90	3	2
21	799815212	5	15000	0	2	2	2	0	50	3	0
22	799475906	5	2750	0	1	1	1	0	80	3	0
23	799411370	6	15000	0	1	1	1	1	60	3	1
24	799305476	5	7500	0	2	1	2	2	100	2	0
25	796325993	6	15000	0	4	2	4	3	100	3	0
26	796276799	3	2750	0	1	1	1	2	90	4	0
27	796279788	6	12500	0	2	2	2	1	140	4	0

Example

SPSS Statistics File Edit View Data Transform Analyze Direct Marketing Graphs Utilities Add-ons Window Help 100% Tue 03/02 11:45 1: Visible: 23 of 23 Variables

	RespondentID	DWELCLAS	INCOME	CHILD13	HIB	HEMPLOY	HSIZE	AGE	SEX	NCARS	AREA	BEDROOM	PARK
1	799161861	5	7500	1	2	2	3	32	1	2	100	2	1
2	798189409	6	4750	0	1	1	1	31	1	1	90	2	1
3	798124382	6	4750	2	2	2	4	42	0	2	220	4	2
4	798271277	5	7500	0	3	2	4	32	1	3	120	3	0
5	798264259	6	2750	1	1	1	2	39	0	1	90	2	0
6	798231878	6	3500	0	3	2	3	47	1	1	100	2	0
7	797907742	4	12500	0	1	0	3	62	1	2	170	5	2
8	797871767	2	3100	0	0	0	0	31	1	3	180	3	0
9	797821230	6	3100	0	0	0	0	31	1	2	80	2	0
10	797152906	5	3100	0	0	0	0	31	1	3	50	1	1
11	797464962	6	3100	0	0	0	0	31	1	3	90	3	1
12	797194471	5	3100	0	0	0	0	31	2	3	90	3	0
13	797131794	5	4750	0	0	0	0	31	2	3	120	4	3
14	797140322	4	12500	0	0	0	0	31	2	3	100	2	1
15	796904634	5	7500	0	0	0	0	31	2	3	120	3	0
16	796430965	5	4750	0	0	0	0	31	3	3	125	3	0
17	796423885	7	4750	0	0	0	0	31	2	3	200	3	0
18	796406846	6	4750	0	0	0	0	31	2	3	90	3	0
19	796389423	5	2750	0	0	0	0	31	2	3	130	2	2
20	800591435	4	2750	0	0	0	0	31	2	3	90	3	2
21	799800252	5	3100	0	2	2	2	34	0	0	50	1	0
22	799471906	5	2750	0	1	1	1	31	0	0	80	3	0
23	799431708	6	3100	0	3	3	3	33	1	1	62	1	1
24	799301476	5	7500	0	2	2	2	48	0	2	100	2	0
25	796121993	6	3100	0	4	2	4	24	0	3	130	3	0
26	796276799	3	2750	0	4	3	5	23	1	2	90	4	0
27	796079788	6	12500	0	2	2	2	18	1	1	140	4	0

Help Reset Paste Cancel OK Data View Variable View IBM SPSS Statistics Processor is ready Unicode:ON

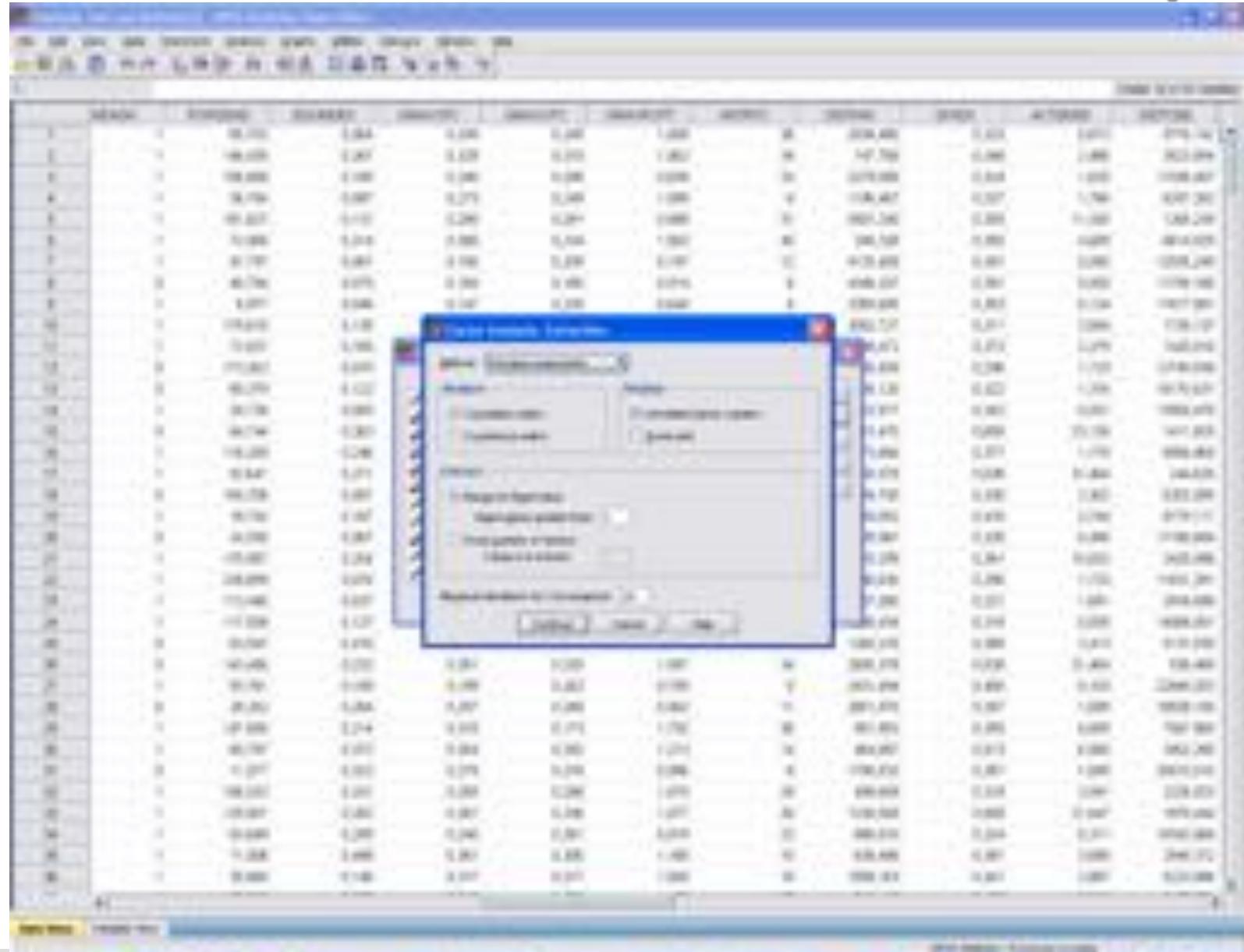
Example

SPSS Statistics File Edit View Data Transform Analyze Direct Marketing Graphs Utilities Add-ons Window Help 100% Tue 00:03 Visible: 23 of 23 Variables

1:	RespondentID	DWELCLAS	INCOME	CHILD13	H18	HEMPLOY	HSIZE	IAGE	ISEX	NCARS	AREA	BEDROOM	PARK
1	798163481	5	7500	1	2	2	3	52	1	2	100	2	1
2	798399409	6	4750	0	1	1	1	31	1	1	90	2	1
3	798374192	6	4750	2	2	2	4	42	0	2	120	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	798255878	6	1500	0	3	2	3	47	1	1	100	2	0
7	797907742	4	32500	0	3	0	3	62	1	2	178	5	2
8	797871767	2	1500							3	100	3	0
9	797821230	6	1500							1	80	2	0
10	797752006	5	1500							1	100	3	1
11	797464932	6	1500							1	90	3	1
12	797194471	5	1500							2	90	3	0
13	797135794	5	4750							2	120	4	3
14	797134022	4	32500							2	105	2	1
15	796904634	5	7500							2	120	3	0
16	796403981	5	4750							3	125	3	0
17	796423881	7	4750							2	200	3	0
18	796406844	4	4750							2	90	3	0
19	796389423	5	2750							2	110	2	2
20	800191413	4	2750							2	90	3	2
21	799800252	5	1500	0	2	2	2	34	0	0	50	3	0
22	799479308	5	2750	0	1	1	1	51	0	0	80	3	0
23	799401703	6	1500	0	1	1	1	33	1	1	62	3	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	796079188	6	32500	0	2	2	2	18	1	1	140	4	0

Data View Variable View IBM SPSS Statistics Processor is ready Unicode ON

Example 1



The screenshot shows a software application window with a large grid of numerical data. A smaller dialog box titled "Run" is overlaid on the grid, containing several tabs and some descriptive text. The dialog box tabs include "General", "Advanced", "Parameters", "Variables", "Outputs", and "Help". The "General" tab is selected, displaying the text: "Run simulation", "Run parameters", and "Run variables". The "Outputs" tab is also visible.

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50	51	52	53	54	55	56	57	58	59	60	61	62	63	64	65	66	67	68	69	70	71	72	73	74	75	76	77	78	79	80	81	82	83	84	85	86	87	88	89	90	91	92	93	94	95	96	97	98	99	100	101	102	103	104	105	106	107	108	109	110	111	112	113	114	115	116	117	118	119	120	121	122	123	124	125	126	127	128	129	130	131	132	133	134	135	136	137	138	139	140	141	142	143	144	145	146	147	148	149	150	151	152	153	154	155	156	157	158	159	160	161	162	163	164	165	166	167	168	169	170	171	172	173	174	175	176	177	178	179	180	181	182	183	184	185	186	187	188	189	190	191	192	193	194	195	196	197	198	199	200	201	202	203	204	205	206	207	208	209	210	211	212	213	214	215	216	217	218	219	220	221	222	223	224	225	226	227	228	229	230	231	232	233	234	235	236	237	238	239	240	241	242	243	244	245	246	247	248	249	250	251	252	253	254	255	256	257	258	259	260	261	262	263	264	265	266	267	268	269	270	271	272	273	274	275	276	277	278	279	280	281	282	283	284	285	286	287	288	289	290	291	292	293	294	295	296	297	298	299	300	301	302	303	304	305	306	307	308	309	310	311	312	313	314	315	316	317	318	319	320	321	322	323	324	325	326	327	328	329	330	331	332	333	334	335	336	337	338	339	340	341	342	343	344	345	346	347	348	349	350	351	352	353	354	355	356	357	358	359	360	361	362	363	364	365	366	367	368	369	370	371	372	373	374	375	376	377	378	379	380	381	382	383	384	385	386	387	388	389	390	391	392	393	394	395	396	397	398	399	400	401	402	403	404	405	406	407	408	409	410	411	412	413	414	415	416	417	418	419	420	421	422	423	424	425	426	427	428	429	430	431	432	433	434	435	436	437	438	439	440	441	442	443	444	445	446	447	448	449	450	451	452	453	454	455	456	457	458	459	460	461	462	463	464	465	466	467	468	469	470	471	472	473	474	475	476	477	478	479	480	481	482	483	484	485	486	487	488	489	490	491	492	493	494	495	496	497	498	499	500	501	502	503	504	505	506	507	508	509	510	511	512	513	514	515	516	517	518	519	520	521	522	523	524	525	526	527	528	529	530	531	532	533	534	535	536	537	538	539	540	541	542	543	544	545	546	547	548	549	550	551	552	553	554	555	556	557	558	559	560	561	562	563	564	565	566	567	568	569	570	571	572	573	574	575	576	577	578	579	580	581	582	583	584	585	586	587	588	589	590	591	592	593	594	595	596	597	598	599	600	601	602	603	604	605	606	607	608	609	610	611	612	613	614	615	616	617	618	619	620	621	622	623	624	625	626	627	628	629	630	631	632	633	634	635	636	637	638	639	640	641	642	643	644	645	646	647	648	649	650	651	652	653	654	655	656	657	658	659	660	661	662	663	664	665	666	667	668	669	670	671	672	673	674	675	676	677	678	679	680	681	682	683	684	685	686	687	688	689	690	691	692	693	694	695	696	697	698	699	700	701	702	703	704	705	706	707	708	709	710	711	712	713	714	715	716	717	718	719	720	721	722	723	724	725	726	727	728	729	730	731	732	733	734	735	736	737	738	739	740	741	742	743	744	745	746	747	748	749	750	751	752	753	754	755	756	757	758	759	760	761	762	763	764	765	766	767	768	769	770	771	772	773	774	775	776	777	778	779	780	781	782	783	784	785	786	787	788	789	790	791	792	793	794	795	796	797	798	799	800	801	802	803	804	805	806	807	808	809	8010	8011	8012	8013	8014	8015	8016	8017	8018	8019	8020	8021	8022	8023	8024	8025	8026	8027	8028	8029	8030	8031	8032	8033	8034	8035	8036	8037	8038	8039	8040	8041	8042	8043	8044	8045	8046	8047	8048	8049	8050	8051	8052	8053	8054	8055	8056	8057	8058	8059	8060	8061	8062	8063	8064	8065	8066	8067	8068	8069	8070	8071	8072	8073	8074	8075	8076	8077	8078	8079	8080	8081	8082	8083	8084	8085	8086	8087	8088	8089	8090	8091	8092	8093	8094	8095	8096	8097	8098	8099	80100	80101	80102	80103	80104	80105	80106	80107	80108	80109	80110	80111	80112	80113	80114	80115	80116	80117	80118	80119	80120	80121	80122	80123	80124	80125	80126	80127	80128	80129	80130	80131	80132	80133	80134	80135	80136	80137	80138	80139	80140	80141	80142	80143	80144	80145	80146	80147	80148	80149	80150	80151	80152	80153	80154	80155	80156	80157	80158	80159	80160	80161	80162	80163	80164	80165	80166	80167	80168	80169	80170	80171	80172	80173	80174	80175	80176	80177	80178	80179	80180	80181	80182	80183	80184	80185	80186	80187	80188	80189	80190	80191	80192	80193	80194	80195	80196	80197	80198	80199	80200	80201	80202	80203	80204	80205	80206	80207	80208	80209	80210	80211	80212	80213	80214	80215	80216	80217	80218	80219	80220	80221	80222	80223	80224	80225	80226	80227	80228	80229	80230	80231	80232	80233	80234	80235	80236	80237	80238	80239	80240	80241	80242	80243	80244	80245	80246	80247	80248	80249	80250	80251	80252	80253	80254	80255	80256	80257	80258	80259	80260	80261	80262	80263	80264	80265	80266	80267	80268	80269	80270	80271	80272	80273	80274	80275	80276	80277	80278	80279	80280	80281	80282	80283	80284	80285	80286	80287	80288	80289	80290	80291	80292	80293	80294	80295	80296	80297	80298	80299	80300	80301	80302	80303	80304	80305	80306	80307	80308	80309	80310	80311	80312	80313	80314	80315	80316	80317	80318	80319	80320	80321	80322	80323	80324	80325	80326	80327	80328	80329	80330	80331	80332	80333	80334	80335	80336	80337	80338	80339	80340	80341	80342	80343	80344	80345	80346	80347	80348	80349	80350	80351	80352	80353	80354	80355	80356	80357	80358	80359	80360	80361	80362	80363	80364	80365	80366	80367	80368	80369	80370	80371	80372	80373	80374	80375	80376	80377	80378	80379	80380	80381	80382	80383	80384	80385	80386	80387	80388	80389	80390	80391	80392	80393	80394	80395	80396	80397	80398	80399	80400	80401	80402	80403	80404	80405	80406	80407	80408	80409	80410	80411	80412	80413	80414	80415	80416	80417	80418	80419	80420	80421	80422	80423	80424	80425	80426	80427	80428	80429	80430	80431	80432	80433	80434	80435	80436	80437	80438	80439	80440	80441	80442	80443	80444	80445	80446	80447	80448	80449	80450	80451	80452	80453	80454	80455	80456	80457	80458	80459	80460	80461	80462	80463	80464	80465	80466	80467	80468	80469	80470	80471	80472	80473	80474	80475	80476	80477	80478	80

Example

Visible: 23 of 23 Variables

	RespondentID	DWELCLAS	INCOME	CHLD013	HL8	HEMPLOY	HSIZE	AGE	SEX	NCARS	AREA	BODROM	PARK
1	798161661	5	7500	1	2	2	3	32	1	2	100	2	1
2	798199409	6	4750	0	1	1	1	31	1	1	90	2	1
3	798374102	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	35	0	1	90	2	0
6	798235878	6	1500						1	1	100	2	0
7	797907742	4	12500						1	2	170	5	2
8	797871767	2	1500						1	3	180	3	0
9	797821230	6	1500						1	4	80	2	0
10	797532006	5	1500						1	5	50	1	1
11	797464962	6	1500						1	6	90	3	1
12	797194471	5	1500						2	2	120	4	3
13	797131794	5	4750						2	3	105	2	1
14	797114022	4	12500						2	4	120	3	0
15	796904634	5	7500						2	5	115	3	0
16	796630963	5	4750						2	6	200	3	0
17	796621883	3	4750						2	7	90	3	0
18	796416844	4	4750						2	8	110	2	3
19	796389423	5	2750						2	9	90	3	2
20	800591423	4	2750						0	0	50	1	0
21	799816252	5	1500						0	0	80	3	0
22	799479906	5	2750						0	1	62	1	1
23	799411703	6	1500	0	1	1	1	33	1	2	100	2	0
24	799305476	5	7500	0	2	1	2	48	0	3	120	4	0
25	798329993	6	1500	0	4	2	4	24	0	3	110	3	0
26	798276799	3	2750	0	4	1	5	23	1	2	90	4	0
27	796079788	6	12500	0	2	2	2	58	1	3	140	4	0

Help Cancel Continue Maximum Iterations for Convergence: 25

Data View Variable View IBM SPSS Statistics Processor is ready Unicode ON

Example

SPSS Statistics File Edit View Data Transform Analyze Direct Marketing Graphs Utilities Add-ons Window Help 100% 849 Tue 00:05 Q

Visible: 23 of 23 Variables

1:	RespondentID	DWELCLAS	INCOME	CHILD13	H18	HEMPLOY	HSIZE	AGE	ISDX	NCARS	AREA	BEDROOM	PARK
1	799061645	5	7500	1	2	2	3	32	1	2	100	2	1
2	7980399409	6	4750	0	1	1	1	33	0	1	90	2	0
3	7980740392	6	4750	2	2	2	4	42	0	2	220	4	2
4	7982752277	5	7500	0	3	2	4	32	1	3	120	3	0
5	7982642110	6	2750	1	1	1	2	33	0	1	90	2	0
6	798235878	6	3500	0	3	2	3	47	1	1	100	2	0
7	797967742	4	12500	0	3	0	3	62	1	2	178	5	2
8	797871267	2	1500							3	180	3	0
9	797821210	6	1500							1	80	2	0
10	797512006	5	1500							1	50	1	0
11	797464902	6	1500							1	90	3	0
12	797134471	5	1500							2	90	3	0
13	797135794	5	4750							2	120	4	3
14	797134022	4	12500							2	185	2	0
15	796804818	5	7500							2	120	3	0
16	796630965	5	4750							3	125	3	0
17	796623885	7	4750							2	200	3	0
18	796435844	4	4750							2	90	3	0
19	796389413	5	2750							2	120	2	0
20	800051415	4	2750							2	90	3	0
21	799810212	5	1500	0	*	*	*	*	0	0	50	1	0
22	799473906	5	2750	0	1	1	1	53	0	0	80	3	0
23	799411700	6	1500	0	1	1	1	33	1	1	62	1	1
24	799005476	5	7500	0	2	1	2	48	0	2	100	2	0
25	798021993	6	1500	0	4	2	4	24	0	3	120	3	0
26	798276799	3	2750	0	4	1	1	23	1	2	90	4	0
27	798079788	6	12500	0	2	2	2	58	0	1	140	4	0

Data View Variable View IBM SPSS Statistics Processor is ready Unicode:ON



FEUP

Example

SPSS Statistics File Edit View Data Transform Analyze Direct Marketing Graphs Utilities Add-ons Window Help 100% 89% Tue 00:06

Visible: 23 of 23 Variables

1:	RespondentID	DWELCLAS	INCOME	CHILD13	H18	HEMPLOY	HSIZE	IACE	ISEX	NCARS	AREA	BEDROOM	PARK
1	79906660	5	7500	1	2	2	3	32	1	2	100	2	1
2	798399409	6	4750	0	3	1	1	31	1	1	90	2	1
3	798374392	6	4750	2	2	2	4	42	0	2	229	4	2
4	798275277	5	7500	0	3	2	4	52	1	3	129	3	0
5	798264210	6	2750	1	3	1	2	33	0	1	90	2	0
6	798235478	6	1500	0	3	2	3	47	1	1	100	2	0
7	797967742	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	797512006	5	1500							2	90	1	1
11	797484902	6	1500							1	90	3	1
12	797394475	5	1500							2	90	3	0
13	797335294	5	4750							2	229	4	3
14	797114622	4	12500							2	105	2	1
15	796904634	5	7500							2	129	3	0
16	796830965	5	4750							3	125	3	0
17	796623885	7	4750							2	200	3	0
18	796416844	6	4750							2	90	3	0
19	796389473	5	2750							2	119	2	2
20	800596415	4	2750							2	90	3	2
21	799815212	5	1500	0	2	2	2	34	0	0	50	1	0
22	799475196	5	2750	0	1	1	1	51	0	0	80	3	0
23	799411703	6	1500	0	3	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	119	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

Data View Variable View IBM SPSS Statistics Processor is ready Unicode ON

Example

SPSS Statistics File Edit View Data Transform Analyze Direct Marketing Graphs Utilities Add-ons Window Help 100% Tue 00:07 Visible: 23 of 23 Variables

	RespondentID	DWELCLAS	INCOME	CHILD13	H18	HEMPLOY	HSIZE	IACE	ISEX	NCARS	AREA	BEDROOM	PARK
1	799061641	5	7500	3	2	2	3	32	1	2	180	2	1
2	798399409	6	4750	6	1	1	1	33	1	1	90	2	1
3	798374392	6	4750	2	2	2	4	42	0	2	220	4	2
4	798275277	5	7500	6	3	2	4	52	1	3	120	3	0
5	798264210	6	2750	1	1	1	2	33	0	1	90	2	0
6	798235478	6	3300	0	3	2	3	47	1	1	180	2	0
7	797967742	4	12500	0	3	0	3	62	1	2	178	5	2
8	797871267	2	1500							3	180	3	0
9	797821210	6	1500							1	80	2	0
10	797512006	5	2100							1	50	1	0
11	797464902	6	2100							1	90	3	0
12	797364475	5	2100							2	90	3	0
13	797335794	5	4750							2	120	4	3
14	797314612	4	12500							2	185	2	1
15	796804614	5	7500							2	120	3	0
16	796830765	5	4750							3	121	3	0
17	796421885	7	4750							2	200	3	0
18	796436844	4	4750							2	90	3	0
19	796389413	5	2750							2	130	2	2
20	800591415	4	2750							2	90	3	2
21	799830212	5	2100	0	4	4	4	24	0	0	50	1	0
22	799471806	5	2750	0	1	1	1	11	0	0	80	3	0
23	799411703	6	2100	0	1	1	1	33	1	1	62	3	1
24	799305478	5	7500	0	2	1	2	48	0	2	180	2	0
25	796325990	6	3300	0	4	2	4	24	0	3	130	3	0
26	796276799	3	2750	0	4	1	5	23	1	2	90	4	0
27	796279788	6	12500	0	2	2	2	18	1	1	140	4	0

Data View Variable View IBM SPSS Statistics Processor is ready Unicode:ON

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

SPSS Statistics File Edit View Data Transform Insert Format Analyze Direct Marketing Graphs Utilities Add-ons Window Help Tue 00:21

Output Log Factor Analysis Title Notes Active Dataset Correlation Matrix KMO and Bartlett Anti-image Matrix Communities Total Variance El Screen Plot Component Matrix Pattern Matrix Structure Matrix Component Corr

```

FACTOR
/VARIABLES TCBD DISTTC TWCBR TDWAK POPDENS EDUNINDEX AREA
/METHOD=PRINCIPAL
/MISSING LISTWISE
/ANALYSIS TCBD DISTTC TWCBR TDWAK POPDENS EDUNINDEX AREA
/PRINT INITIAL CORRELATION SIG DET KMO AIC EXTRACTION ROTATION
/FORMAT SORT REVERSE(.4)
/PLOT EIGEN
/CRITERIA MINEIGEN(1) ITERATE(25)
/EXTRACTION PC
/CRITERIA ITERATE(25) DELTA(0)
/ROTATION OBLIMIN
/SAVE REG(ALL)
/METHOD=CORRELATION.

```

Factor Analysis

(Dataselet1) /Users/

Correlation	TCBD	DISTTC	TWCBR	TDWAK	POPDENS	EDUNINDEX	AREA
TCBD							
DISTTC	.000		.000	.000	.000	.000	.062
TWCBR	.000	.000		.199	.001	.000	.097
TDWAK	.000	.000	.199		.000	.000	.036
POPDENS	.000	.000	.001	.000		.185	.000
EDUNINDEX	.000	.000	.000	.000	.185		.336
AREA	.000	.062	.097	.016	.000	.336	

a. Determinant = .236

KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.	.674
--	------

IBM SPSS Statistics Processor is ready | Unicode:ON H: 504, W: 629 pt.

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



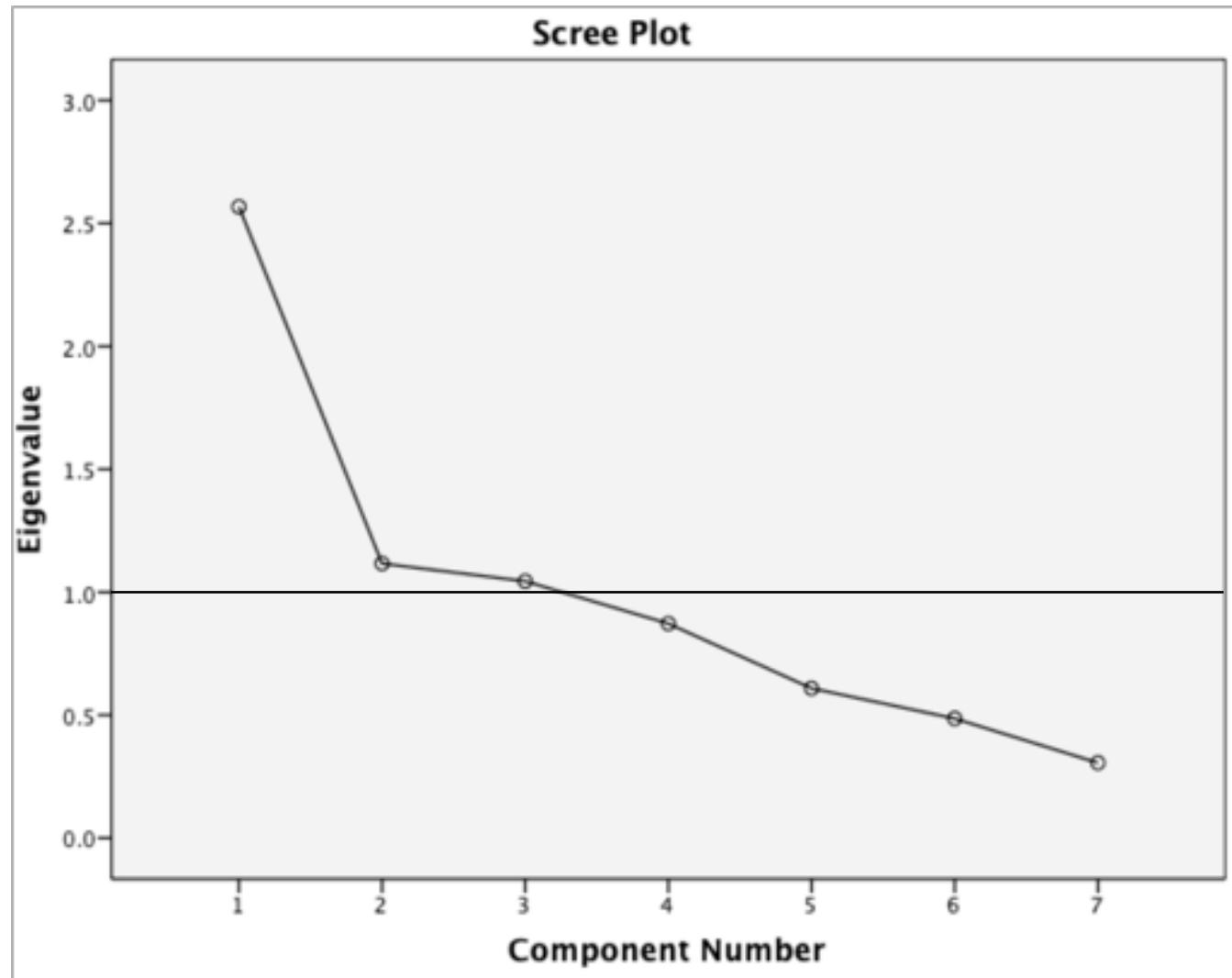
Total Variance Explained							
Component	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings ^a
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	
1	2.567	36.673	36.673	2.567	36.673	36.673	2.334
2	1.116	15.947	52.619	1.116	15.947	52.619	1.148
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



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?

	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