

Daumenregel, Checklisten und Anwendungsempfehlungen für Ökonometriker und Datenanalysten

Mumdzhev, Milko

Preprint / Preprint

Arbeitspapier / working paper

Empfohlene Zitierung / Suggested Citation:

Mumdzhev, M. (2010). *Daumenregel, Checklisten und Anwendungsempfehlungen für Ökonometriker und Datenanalysten*. (Nürnberger Beiträge zur Sozial- und Wirtschaftsforschung, 02/10). Nürnberg. <https://nbn-resolving.org/urn:nbn:de:0168-ssoar-256397>

Nutzungsbedingungen:

Dieser Text wird unter einer CC BY-NC-ND Lizenz (Namensnennung-Nicht-kommerziell-Keine Bearbeitung) zur Verfügung gestellt. Nähere Auskünfte zu den CC-Lizenzen finden Sie hier:

<https://creativecommons.org/licenses/by-nc-nd/4.0/deed.de>

Terms of use:

This document is made available under a CC BY-NC-ND Licence (Attribution-Non Commercial-NoDerivatives). For more information see:

<https://creativecommons.org/licenses/by-nc-nd/4.0>

Daumenregel, Checklisten und Anwendungsempfehlungen für Ökonometriker und Datenanalysten

Milko Mumdzhev

Zusammenfassung:

Eine Auflistung von Anwendungsempfehlungen und Hinweisen zur multivariaten Datenanalyse wird vorgelegt. Insbesondere werden die Bereiche Data Screening, lineare Regressionsanalyse, univariate und multivariate Varianz-Kovarianzanalyse, Profilanalyse, Diskriminanzanalyse, logistische Regressionsanalyse, Verweildauermodelle, Clusteranalyse, kanonische Korrelationsanalyse, Faktorenanalyse, Strukturgleichungsmodelle, Mehrebenenanalyse, Kontingenzanalyse, multi-dimensionale Skalierung und Conjoint-Measurement-Analyse stichpunktartig behandelt.

Copyright © Milko Mumdzhev

Das Arbeitspapier darf ohne die ausdrückliche Genehmigung des Autors nicht reproduziert werden.

Gliederung

Einleitung

1. Daumenregeln

- *Basics*
- *Sample size*
- *Covariation*
- *Epidemiology*
- *Environmental studies*
- *Design, Conduct, and Analysis*
- *Words, Tables, and Graphs*
- *Consulting*

2. Checklisten und Anwendungsempfehlungen für multivariate Analysemethoden

- *Data Screening*
- *Lineare Regressionsanalyse*
- *Univariate Varianz-Kovarianzanalyse*
- *Multivariate Varianz-Kovarianzanalyse*
- *Profilanalyse*
- *Diskriminanzanalyse*
- *Logistische Regressionsanalyse*
- *Verweildauermodelle*
- *Clusteranalyse*
- *Kanonische Korrelationsanalyse*
- *Faktorenanalyse*
- *Strukturgleichungsmodelle*
- *Mehrebenenanalyse*
- *Kontingenzanalyse*
- *Multidimensionale Skalierung*
- *Conjoint-Measurement-Analyse*

Verwendete Literatur

Einleitung

Nachfolgend sind Daumenregeln, Checklisten und Anwendungsempfehlungen für Vorgehen und Verfahrensweisen bei statistischen Datenauswertungen in der Markt- und Sozialforschung zusammengestellt. Sie sollen eine Orientierung geben, und auf wichtige Punkte und Details zeigen, ohne ausführlich bei der jeweiligen Methode in die Breite gehen zu wollen. Man kann sie als diejenigen notwendigen Schritte und Stichpunkte ansehen, welche Datenanalysten nacheinander prüfen, um letztendlich die geforderten Ergebnisse zur jeweiligen Studie zu produzieren. Anzumerken ist, dass ein Mehr oder Weniger an Aufwand, Anforderung und (argumentativer) Präzision der Analyse abhängig ist vom Ziel und Ausmaß der Studie, des Auftraggebers und Klienten, vom genauen Vertrag und Kostenvoranschlag, von bestehenden Deadlines usw.

Die Expertise eines (externen) Datenanalysten könnte in folgenden Projektphasen mit jeweils anfallenden Aufgaben hinzugezogen werden (Cabrera & McDougall, 2002: S.20):

Data Collection
Pretest and manage data collection
Review response rates
Coordinate exchange of sample / data
Progress meeting with client
Survey Validation
Clean, format, label data
Respondent profile (demographics)
Create tables for report
Profile of respondents' company / product usage
Examine brand and question order effects
Summary of responses by survey sections, companies
Standard errors for tables and graphs
Select results for graphical presentation in report
Meeting with project leader

Survey Analysis
Scale analysis, decision trees for value / quality
Correlations and factor analysis: price, value, and questions
Correlations and factor analysis by section: client vs. others
Factor analysis for product questions: client vs. others
Logistic regression: client vs. competitor
Client meeting: discuss validation and analysis results
Modeling:
Set up data for modeling
Modeling to determine relationships between variables
Structural equation modeling
Client meeting: discuss modeling prior to write-up
Reporting Results:
Project meeting: discuss results and organize writing efforts
Creating / inserting tables and graphics in report document
Presentation to client: meeting scheduled for entire day

Aufbau der Arbeit

Im Hauptteil des Arbeitspapiers sind zu oben genannten und weiteren Punkten Daumenregeln und Anwendungsempfehlungen, Checklisten und Hinweise zusammengestellt.

Wichtige Grundlagen der deskriptiven Statistik und des Studiendesigns werden ebenso (kurz und knapp) genannt wie der Ablauf einer Diskriminanzanalyse oder der Aufbau eines Pfaddiagramms in (linearen) Strukturgleichungsmodellen. Bei den Aufzählungen handelt es sich um Zusammenfassungen oder Exzerpte aus den entsprechenden Kapitel der zitierten Literatur, deren Lektüre und Handhabung dadurch nicht ersetzt werden; (erste) Orientierungen oder vertraglich zu fordernde Leistungen werden stichpunktartig vorschlagen.

Im ersten Teil werden Daumenregeln aufgezählt, die alle Bereiche des (Markt-)Forschungsprozesses betreffen, von der Datengewinnung bis zur Auswertung; im zweiten Teil werden Checklisten und Anwendungsempfehlungen für diverse Typen von Datenanalysen zitiert.

1. Daumenregeln

Dieser Abschnitt enthält Daumenregeln zu den Grundlagen der Statistik, Hinweise zum Studiendesign, zur Bestimmung von Stichprobengrößen und Zusammenhangsmaßen, zählt Besonderheiten von epidemiologischen Studien oder Datenpräsentationen an sich auf. Hintergründe und Erläuterungen mit Beispielen, etwaige Streitpunkte und vertiefende Literatur zu den Punkten wird von van Belle (2002) selbst geliefert.

Basics (van Belle, 2002: S.1-28)

1. Make a sharp distinction between observational and experimental studies (*Zufallsauswahl, Repräsentativität der Stichprobe*).
2. Always look for a physical model underlying the data being analyzed. Assume that a statistical model such as a linear regression model is a good first start only (*Ist Linearität gegeben? Wie sehen die Alternativen aus? Kann man zu Strukturgleichungen übergehen?*).
3. Be sure to understand the components and purpose of an omnibus quantity (*F-tests, p-Werte, Mahalanobisdistanzen etc. Wie und welche Datenreduktion findet statt?*).
4. Classical hypothesis tests assume that the observations a) are independent, b) all come from populations with the same variance, and, for parametric tests, c) follow a normal distribution. The most important (in terms of maintaining an assumed Type I error level: the probability of rejecting the null hypothesis when it is true) ist the first point, then the second, then the third (*Häufigkeitsanalysen, Plots und Histogramme können Aufschluss geben*).
5. One formal statement of Ockham's razor is: do not multiply hypotheses beyond necessity (*Das einfachste Modell von vielen ist vorzuziehen, sofern alle anderen Bedingungen konstant gehalten werden*).
6. Do not multiply probabilities beyond necessity (*Die Multiplikation von Wahrscheinlichkeiten (<1) ergibt entsprechend kleine Werte*).
7. Make sure the sample space is known for a statement of risk (*Wahrscheinlichkeitsinterpretationen beachten!*).
8. The use of one-sided p-values is discouraged. Ordinarily, use two-sided p-values.
9. The width of a confidence interval, involving estimation of variability and sample size, decreases rapidly until 12 observations are reached and then decreases less rapidly.
10. Always know what the unit of a variable is (*z.B. logarithmiertes Einkommen; Gradskalen*).

11. Know what properties a transformation preserves and does not preserve.
12. Do not let the scale of measurement rigidly determine the method of analysis (*Welche Methoden eignen sich allgemein, um die Analyse unterschiedlich skaliertes Merkmale durchzuführen?*).
13. Think of bootstrapping instead of the delta method in estimating complex relationships.
14. The standard error of a small sample of observations (15 or fewer) can be estimated by dividing the range by the number of observations in the sample (*Range als die Spannweite, nämlich größte Beobachtung minus kleinste Beobachtung*).

Sample size (van Belle, 2002: S.29-51)

1. The basic sample size formula (for 2 groups, with a 2-sided alternative, normal distribution with homogeneous variances and equal sample sizes) is $n=16/\Delta^2$, where $\Delta = (\text{Difference of the means under the null and alternative hypotheses (treatment difference, } \mu_0-\mu_1)) / \sigma \text{ (SD)}$.
2. The finite population correction (*für den Standardfehler*) can be ignored in initial discussions of survey sample size questions.
3. The sample size formula becomes: $n=8 ((\sigma_0/\mu_0)^2 / ((\mu_0-\mu_1)/\mu_0)^2)(1+((1-(\mu_0-\mu_1)/\mu_0)^2))$.
4. Do not formulate objectives for a study solely in terms of effect size (Δ).
5. Confidence intervals associated with statistics can overlap as much as 29 % and the statistics can still be significantly different.
6. To get equal precision with a two-sample situation with n observations per sample given $(n_0(n_0 < n))$ in the first sample and kn_0 observations in the second sample, choose $k=n/(2n_0-n)$.
7. To minimize the total cost of a study, choose the ratio of the sample sizes according to $n_1/n_0 = \sqrt{c_0/c_1}$. Pick sample sizes (n) inversely proportional to square root of the cost (c) of the observations. If costs are not too different, then equal sample sizes are suggested (because the square root of the ratio will be closer to 1).
8. Given no observed events in n trials, a 95% upper bound on the rate of occurrence (*in %*) is $3/n$.
9. Ordinarily the sample size calculation should be based on the statistics used in the analysis of the data.

Covariation (van Belle, 2002: S.53-74)

1. Before choosing a measure of covariation determine the source of the data (sampling scheme), the nature of the variables, and the symmetry status of the measure (z.B. *Korrelation*).
2. Do not summarize regression sampling schemes with correlations.
3. Do not correlate rates or ratios indiscriminately (*Verhältnismaße mit gemeinsamen Termen sind korreiliert*).
4. To determine the appropriate sample size to estimate a population correlation r , use this formula: $n=16/\Delta^2$, $\Delta = 0,5 \ln ((1+r)/(1-r))$.
5. Pairing should be investigated carefully. Do not pair unless the correlation between the pairs is greater than 0,5.
6. Do not use the coefficient of correlation as a measure of agreement when location and scale are relevant (*Invarianz des Korrelationskoeffizienten!*).
7. Test reliability in the social sciences can be assessed by means of agreement or the correlation coefficient.
8. Suppose the investigator has n pairs of observations available, with a specified range of values for the values of X . To get the smallest standard error for the estimate of the slope β in a simple linear regression model, place $n/2$ observations at the smallest value of X and the other $n/2$ observations at the largest value of X .
9. In assessing change, the spacing of the observations is much more important than the number of observations.

Epidemiology (van Belle, 2002: S.75-102)

1. Start with the Poisson distribution to model disease incidence or prevalence.
2. Under the rare disease assumption the odds ratio approximates the relative risk.
3. The estimate of sample size per group in a cohort study, based on the logarithm of the relative risk, R , is $n=(8(R+1)/R)/(\pi_0 (\ln R)^2)$, with power $1-\beta=0,80$, $\alpha =0,05$, and a two-sided alternative (π_0 is the probability of disease in the unexposed group, R is the relative risk for the exposed one).
4. In terms of variance, not much is gained beyond taking four or five controls per case.
5. In logistic regression situations about 10 events per variable are necessary in order to get reasonably stable estimates of the regression coefficients.

6. The exponential distribution is the basic distribution for modeling survival times when the rate is constant over time.
7. For comparing two populations with hazard functions λ_0 and λ_1 , use sample sizes of $n = 16/(\ln(\lambda_1/\lambda_0))^2$, where n is the number per group.
8. Be aware of surrogates. Accept substitutes warily (*z.B. das Ende der Versuchsreihe im Labor*).
9. In rare diseases the prevalence dominates the predictive value of a positive test.
10. Do not dichotomize unless absolutely necessary.
11. Select an additive or multiplicative model according to the following order: theoretical justification, practical implication, and computer implementation.

Environmental Studies (van Belle, 2002: S.103-127)

1. Think lognormal (in terms of the lognormal distribution) for measurement data in environmental studies.
2. Differences tend to be more symmetrical.
3. Beware of pseudoreplication (*z.B. 24 Stunden eines Tags vs. jeweils eine Stunde an 24 Tagen*).
4. Always consider alternatives to simple random sampling for a potential increase in efficiency, lower costs, and validity.
5. In assessing the importance of an effect, consider the size of the population to which it applies.
6. Models estimating small effects in large populations are particularly sensitive to assumptions. Extensive sensitivity studies are needed in such cases to validate the model.
7. In using a data base, first look at the metadata, then look at the data.
8. It is not always easy to determine the statistical basis for a standard. Be sure to understand the statistical underpinning of the standard.
9. How a pollutant is measured plays a key role in identification, regulation, enforcement, and remediation.
10. Ordinarily, a parametric analysis is the most informative. A nonparametric analysis is useful for an initial, quick look at data.
11. Distinguish among confidence, prediction, and tolerance intervals. Confidence intervals are statements about population means or other parameters. Prediction intervals address future (single or multiple) observations. Tolerance intervals describe the location of a specific proportion of a population, with specified confidence.

12. Consider these five areas of risk assessment (Hazard identification, Dose-response evaluation, Exposure assessment, Risk Characterization, Risk Management). The first four activity areas are heavily dependent on statistical input. The last stage involves regulation and policy, based on the information obtained in the first four steps.
13. In the chain from pollution generation to health effects, the assessment of human exposure is the most fragile and most complicated link. Examine the validity of the exposure assessment before doing risk assessment.
14. Calibration involves inverse regression, and the error associated with this regression must be assessed.

Design, Conduct, and Analysis (van Belle, 2002: S.129-151)

1. Randomization puts systematic sources of variability into the error term.
2. By randomly assigning treatment to experimental units within blocks, systematic effects between blocks can be eliminated with a resultant decrease of within-treatment variability and increased precision.
3. The effects of two or more factors can be assessed simultaneously by means of factorial designs in which treatment combinations are applied to the observational units. A benefit of such a design is that there may be decreased cost.
4. High-order interactions occur rarely, therefore it is not necessary to design experiments that incorporate tests for high-order interactions.
5. Aim for balance in the design of a study.
6. Analysis should follow design.
7. For every analysis there is an appropriate graphical display.
8. Ordinarily, plan to do a hierarchical analysis of treatment effects by including all lower order effects associated with the higher order effects.
9. Best rule for missing data problems: do not have any missing data. Well, that is unrealistic. Therefore, expect missing data and develop strategies to account for them. Do this before starting the study. The strategy should state explicitly how the type of missingness will be examined, how it will be handled, and how the sensitivity of the results to the missing data will be assessed.

Words, Tables, and Graphs (van Belle, 2002: S.153-173)

1. Use sentence structure for displaying 2-5 numbers, tables for displaying more numerical information, and graphs for complex relationships.
2. Tables:
 - Arrange the rows and columns in a table in a meaningful way in order to display as much structure as possible.
 - Limit the number of significant digits.
 - Make the table as self-contained as possible.
 - Use white space and lines to organize rows and columns.
 - Use the table heading to convey crucial information. Do not stint. The more informative the heading, the better the table.
3. When possible always graph the data.
4. Always think of an alternative to a bar graph.
5. In the case of longitudinal data identify both cross-sectional and longitudinal patterns.
6. Three key aspects of presenting high dimensional data are: rendering, manipulation, and linking. Rendering determines what is to be plotted, manipulation determines the structure of the relationships, and linking determines what information will be shared between plots or sections of the graph.

Consulting (van Belle, 2002: S.175-194)

1. Structure a consulting session to have beginning, a middle, and an end. The session begins with a problem context and definition, moves to resolution and solution, and concludes with a summary and allocation of responsibilities.
2. Pace the consultation by asking focused, helpful questions.
3. Make valid distinctions, not distinctions valid.

Basic distinctions to be made in a Consulting Session (van Belle, 2002: S.179):

Intent of study	Exploratory	Confirmatory
Source of data	Random	Convenience
Sampling unit	Simple	Multiple
Randomization	Single	Multiple
Model	Fixed	Random or mixed
Type of Study	Intrinsic	Comparative
Type of Study	Cross-sectional	Longitudinal
Epidemiologic study	Cohort	Case-Control
Categories	Nominal	Some kind of ordering
Numerator / denominator	Rate	Proportion, Ratio

4. Know yourself, and know the investigator.
5. Tailor advice to the investigator's level of understanding. Avoid jargon, equations, and formulas.
6. When communicating with nonstatisticians, use natural units as much as possible.
7. When the consulting session is done, be clear as to who does what, and when.
8. Ethic precedes, guides, and follows consultation.
9. Be proactive. Being proactive means taking the lead in scientific discussions, coordinating separate research efforts, and thinking broadly about the scientific enterprise.
10. Use the web for reference, resource, and education, for consultants and investigators. But use it with discernment.
11. Listen to, and heed the advice of experts in the field.

2. Checklisten und Anwendungsempfehlungen für multivariate Datenanalysen

Im zweiten Teil der Arbeit werden folgende Kapitel stichpunktartig abgehandelt:

- Data Screening
- Lineare Regressionsanalyse
- Univariate Varianz-Kovarianzanalyse
- Multivariate Varianz-Kovarianzanalyse
- Profilanalyse
- Diskriminanzanalyse
- Logistische Regressionsanalyse
- Überlebensanalyse
- Clusteranalyse
- Kanonische Korrelationsanalyse
- Faktorenanalyse
- Stukturgleichungsmodelle
- Mehrebenenanalyse
- Kontingenzanalyse
- Multidimensionale Skalierung
- Conjoint-Measurement-Analyse.

Dabei sind wiederholte Nennungen bestimmter Vorgehensweisen vor der eigentlichen Analyseprozedur zum Zwecke der (vereinfachten) Lektüre gemacht worden, z.B. Missing Values- oder Ausreißerhandhabung, Prüfen der Modellannahmen (z.B. Normalverteiltheit der Variablen und gegebenenfalls deren Transformation, geringe Ausprägung von Multikollinearität). Die Hinweise und Checklisten von Tabachnick & Fidell (2007) sind entsprechenden Anwendungsempfehlungen von Backhaus et al. (1994 bzw. 2003) hintangestellt; beide überlappen sich zum Teil oder ergänzen sich geringfügig. Letztere erläutern die jeweilige Methode. Sofern SPSS-Befehle und / oder Optionen genannt werden, sind etwaige Syntaxen, Ausgaben und Ergebniskommentare ausgelassen worden.

Mögliche Fehler bei der Dateneingabe, z.B. in MS ACCESS oder SPSS Data Entry sind tabelliert (Cabrera & McDougall, 2002: S.71):

Potential Error	Possible Cause
Some categorical classes contain very few observations	Data Entry Error Artifact of the database
A categorical variable contains more levels than it should	Changes in coding used Column shift during read-in
Variables that contain a high proportion of missing values	Wrong format used in read-in Column shift during read-in Programming error
Values associated with user-defined variables do not make sense	Data entry error; Programming error Illegal operation performed
Values that lie outside the expected or allowable range of a variable	Outlier: data entry error Read-in or programming error Poor quality data

Vor jeder weiteren Analyse ist die Sichtung und Bereinigung des Datensatzes selbstverständlich.

Checklist for Screening Data (Tabachnick & Fidell, 2007: S.91)

1. Inspect univariate descriptive statistics for accuracy of input

- out of range values
- plausible means and standard deviations
- univariate outliers

Das Ausgeben von Häufigkeiten, deskriptiven (summary) Statistiken inklusive Mittelwerte, Abweichungen, Schiefe- und Kurtosismaßzahlen mit Standardfehlern, Minimum und Maximum plus Histogramme (mit der Normalverteilungskurve) ist die übliche Praxis (Tabachnick & Fidell, 2007: S.93).

2. Evaluate amount and distribution of missing data: deal with problem

Einer Missing Value Analysis kann listenweiser Ausschluss der jeweiligen Variablen oder Imputation folgen (Tabachnick & Fidell, 2007: S.62 ff.).

3. Check pairwise plots for nonlinearity and heteroskedasticity

Plotten von (je zwei) Variablen gegeneinander kann das typische ovale Zigarrenmuster bei Homoskedastie anzeigen, oder eine Auffächerung des Musters als Heteroskedastie erkennen helfen. Linearität kann daneben in Plots von (standardisierten) Residuen gegen vorhergesagte Werte sichtbar gemacht werden (Tabachnick & Fidell, 2007: S.83 ff.).

4. Identify and deal with nonnormal variables and univariate outliers

-check skewness and kurtosis, probability plots

-transform variables (if desirable)

-check results of transformation

Numerische Tests auf Vorliegen von Normalverteilung, Histogramme oder Plots der Variablen gegen die erwarteten Normalverteilung nebst anderen sind üblich, um Normalverteilung (als Winkelhalbierende) zu erkennen. Transformationen wie die Quadratwurzel, der Logarithmus (typisches Beispiel: Einkommen) oder die Inverse der Variable, um Schiefe oder Kurtosismaße den der Normalverteilung anzugleichen sind gängig und einfach zu implementieren, erschweren unter Umständen die Interpretation (Tabachnick & Fidell, 2007: S.78 ff.). Daneben ist die Betrachtung der z-Scores von Variablen oder der entsprechenden Boxplots für das Auffinden und Erkennen von univariaten Ausreißern hilfreich.

5. Identify and deal with multivariate outliers

-variables causing multivariate outliers

-description of multivariate outliers

Die Anzeige von (extremen) Ausreißern kann z.B. über Mahalanobisdistanz- oder Cooksmaßzahlen geschehen. Es mag nützlich sein, diese Fälle, vielleicht die entsprechende Variable zu löschen, oder sie nur zu transformieren, und so Ausreißer wie schwere Ränder zu behandeln (Tabachnick & Fidell, 2007: S.72 ff.).

6. Evaluate variables for multicollinearity and singularity

Kollinearitätsprobleme können beispielsweise durch Konditionsindizes über 30 und Varianzproportionen über 0,5 für mindestens 2 Variablen angezeigt sein (in SPSS über den Regressionsbefehl plus Optionen angezeigt). Ignorieren als die einfachste Handhabung, Summieren der entsprechenden Variablen, Löschen und andere Optionen stünden hierbei zur Verfügung (Tabachnick & Fidell, 2007: S.91).

Vorgehensweise Regressionsanalyse (Backhaus et al, 2003: S.52 ff.)

- Formulierung des Modells
- Schätzung der Regressionsfunktion
- Prüfung der Regressionsfunktion
 - Bestimmtheitsmaße, F-Statistik, Standardfehler
- Prüfung der Regressionskoeffizienten
 - T-Werte, Beta-Werte
- Prüfung der Modellprämissen
 - Linearität, Erwartungswert der Störgröße ist gleich Null, (falsche) Auswahl an Regressoren, keine Heteroskedastizität, keine Autokorrelation der Residuen, keine Multikollinearität, Normalverteilung der Residuen

Anwendungsempfehlungen Regressionsanalyse (Backhaus et al, 1994: S.48 f.)

1. Das Problem, das es zu untersuchen gilt, muss genau definiert werden. Welche Größe soll erklärt werden; sie muss metrisch skaliert sein.
2. Sachkenntnis und Überlegungen sind notwendig, um mögliche Einflußgrößen, welche auf die zu erklärende Variable einwirken, zu erkennen und zu definieren.
3. Die Stichprobe muss genügend groß sein; die Anzahl der Beobachtungen sollte mindestens doppelt so groß sein wie die der Prädiktoren.
4. Vor Beginn der Rechnung: hypothetische Modelle mit vorhandenen Variablen aufstellen; Art und Stärke der berücksichtigten Variablen muss überlegt werden.
5. Nach Schätzung der Regression: Bestimmtheitsmaß auf Signifikanz prüfen; wenn nicht signifikant, muss der gesamte Regressionsansatz verworfen werden.
6. Anschließend sind die einzelnen Regressionskoeffizienten sachlogisch und statistisch (Vorzeichen und Signifikanz) zu prüfen.
7. Einhaltung der Prämissen des linearen Regressionsmodells prüfen.
8. Eventuell sind Variablen zu entfernen, oder neue hinzuzufügen. Die Modellbildung ist ein iterativer Prozess, bei dem auf Basis von empirischen Ergebnissen neue Hypothesen formuliert, geprüft, verworfen werden etc.
9. Wenn alle Prämissen erfüllt sind, erfolgt die Überprüfung an der Realität.

Checklist for Standard Multiple Regression (Tabachnick & Fidell, 2007: S.172)

1. Issues

-ratio of cases to independent variables and missing data

Bei der klassischen Regression kann ein entsprechendes Fall-Prädiktoren-Verhältnis von 20 zu 1 sinnvoll sein (Tabachnick & Fidell, 2007: S.123).

-normality, linearity, and homoscedasticity of residuals

Plots der Residuen gegen die vorhergesagten Werte können als (normalverteilte) Steuung um die x-Achse auf Zutreffen der Modellannahmen deuten, oder Nichtlinearität oder Heteroskedastie anzeigen (Tabachnick & Fidell, 2007: S.125 f.). Auch die Unabhängigkeit der Fehlerterme muss gegeben sein (Tabachnick & Fidell, 2007: S.128). Statistische Testverfahren wie der Durbin-Watson-Test u.a. bieten sich ebenso an wie grafische, um bestimmte Modellannahmen zu prüfen.

-outliers

-multicollinearity and singularity

-outliers in the solution

Die Anzeige von (extremen, multivariaten) Ausreißern kann z.B. über Mahalanobisdistanz- oder Cooksmaßzahlen geschehen. Es kann der Fall oder die entsprechende Variable gelöscht werden, man kann sie auch transformieren, um so Ausreißer im Sinne von schweren Rändern zu behandeln (Tabachnick & Fidell, 2007: S.72 ff.).

Kollinearitätsprobleme können beispielsweise durch Konditionsindizes über 30 und Varianzproportionen über 0,5 für mindestens 2 Variablen angezeigt sein (in SPSS über den Regressionsbefehl angezeigt). Ignorieren als die einfachste Handhabung, Summieren der entsprechenden Variablen und andere Optionen stünden hierbei zur Verfügung (Tabachnick & Fidell, 2007: S.91).

2. Major analyses

-multiple R^2 and its confidence limits, F ratio

-adjusted multiple R^2 , overall proportion of variance accounted for

-significance of regression coefficients

-squared semipartial correlations

Die entsprechende Ausgabe der Regressionsprozedur liefert die geforderten Werte, Richtungen, Stärken, Signifikanzen der Koeffizienten usw.

3. Additional analyses

- post hoc significance of correlations
- unstandardized (B) weights, confidence limits
- standardized (β) weights
- unique versus shared variability
- suppressor variables
- prediction equation

Suppressorvariablen erhöhen das R im Modell, indem sie die Wirkung der unabhängigen Variablen verstärken, und unnötige Varianz unterdrücken (Tabachnick & Fidell, 2007: S.154).

Checklist for Sequential Regression Analysis (Tabachnick & Fidell, 2007: S.177)

1. Issues

- ratio of cases to independent variables and missing data

Bei der stepwise-Regression kann ein entsprechendes Verhältnis von 40 zu 1 sinnvoll sein (Tabachnick & Fidell, 2007: S.123).

- normality, linearity, and homoscedasticity of residuals

Plots der Residuen gegen die vorhergesagten Werte können als (normalverteilte) Steuerung um die x-Achse auf Zutreffen der Modellannahmen deuten, oder Nichtlinearität oder Heteroskedastie anzeigen (Tabachnick & Fidell, 2007: S.125 f.). Auch die Unabhängigkeit der Fehlerterme muss gegeben sein (Tabachnick & Fidell, 2007: S.128). Statistische Testverfahren wie der Durbin-Watson-Test u.a. bieten sich ebenso an wie grafische, um bestimmte Modellannahmen zu prüfen.

- outliers

- multicollinearity and singularity

- outliers in the solution

Die Anzeige von (extremen, multivariaten) Ausreißern kann z.B. über Mahalanobisdistanz- oder Cooksmaßzahlen geschehen. Von Fall zu Fall mag es nützlich sein, die entsprechende Variable zu löschen, oder sie zu transformieren, um so Ausreißer im Sinne von schweren Rändern zu behandeln (Tabachnick & Fidell, 2007: S.72 ff.).

Kollinearitätsprobleme können beispielsweise durch Konditionsindizes über 30 und Varianzproportionen über 0,5 für mindestens 2 Variablen angezeigt sein (in SPSS über den Regressionsbefehl angezeigt). Ignorieren als die einfachste Handhabung, Summieren der

entsprechenden Variablen und andere Optionen stünden hierbei zur Verfügung (Tabachnick & Fidell, 2007: S.91).

2. Major analyses

- multiple R^2 and its confidence limits, F ratio
- adjusted multiple R^2 , overall proportion of variance accounted for
- squared semipartial correlations
- significance of regression coefficients
- incremental F

Die entsprechende Ausgabe der Regressionsprozedur liefert die geforderten Werte, Richtungen, Stärken, Signifikanzen der Koeffizienten usw. Natürlich sind schrittweise (stepwise) Betrachtungen hier angebracht.

3. Additional analyses

- unstandardized (B) weights, confidence limits
- standardized (β) weights
- prediction equation from stepwise analysis
- post hoc significance of correlations
- suppressor variables
- cross-validation (stepwise)

Suppressorvariablen erhöhen das R im Modell, indem sie die Wirkung der unabhängigen Variablen verstärken, und unnötige Varianz unterdrücken (Tabachnick & Fidell, 2007: S.154).

Anwendungsempfehlungen Varianzanalyse (Backhaus et al, 1994: S.84 f.)

1. Unabhängige Variablen: jedes Skalenniveau (z.B. als Kovariate); abhängige Variablen: metrisch.
2. Faktoren müssen sich eindeutig voneinander unterscheiden.
3. Varianzhomogenität: mögliche, nicht bedachte Einflussgrößen wirken sich bis auf zufällige Schwankungen, die sich ausgleichen, in allen Stichprobenzellen gleich aus.
4. Teilstichproben sollten die gleiche Struktur wie die Grundgesamtheit aufweisen.
5. Werte in der Grundgesamtheit sind normalverteilt (Fallzahl!).
6. Additivität der Einflussgrößen: Faktoreinfluss und Störvariableneinfluss sind unabhängig voneinander (Zufallsauswahl bei der Stichprobe).

7. Selbst bei Verletzung der Prämissen: VA ist bei gleichen Zellenbesetzungen relativ robust; dient zur Testung des Vorliegens eines Zusammenhangs, nicht über seine Stärke auszusagen.

Checklist for Analysis of Covariance (Tabachnick & Fidell, 2007: S.238)

1. Issues

-unequal sample size and missing data

Es kann z.B. sein, dass ungleiche Zelhäufigkeiten eine Bedeutung der Zellen indizieren (Tabachnick & Fidell, 2007: S.217 f.). Eine entsprechende Gewichtung muss somit nicht immer angebracht sein.

-within-cells outliers

Die Handhabung von Ausreißern, die womöglich Heterogenität der Regression induzieren, ist weiter oben in diesem Paper angedeutet.

-normality

Bei einer entsprechend großen Stichprobe kann Normalverteilung antizipiert werden (Tabachnick & Fidell, 2007: S.202 bzw. S.224).

-homogeneity of variance

Es wird angenommen, dass die Varianzen / Kovarianzen der Zellen eine separate Schätzung der selben Populationsvarianz darstellen (Tabachnick & Fidell, 2007: S.202).

-within-cells linearity

Linearitätsbeziehungen werden zwischen den Kovariaten und der abhängigen Variable sowie zwischen den Kovariaten angenommen. Transformationen der Variablen können sinnvoll sein, jedoch die Interpretation erschweren (Tabachnick & Fidell, 2007: S.202).

-homogeneity of regression

Man nimmt hier an, dass innerhalb der Zellen die Regressionskoeffizienten aus der selben Grundgesamtheit der Koeffizienten geschätzt werden, und somit in einem Plot der Kovariaten und der abhängigen Variable durch parallele Geraden angedeutet werden (S.202 f.).

-reliability of covariates

Kovariate werden als ohne Fehler gemessen angenommen (Beispiel: Alter). Reliabilitäten zu 0,8 als Grenze für Kovariate in nicht-experimentellen Designs zu wählen, kann eine Daumenregel sein (Tabachnick & Fidell, 2007: S.203).

2. Major analyses

Daumenregel, Checklisten und Anwendungsempfehlungen für Ökonometriker und Datenanalysten

- main effect(s) or planned comparison. If significant: adjusted marginal means and standard deviations or standard errors or confidence intervals
- interactions or planned comparisons. If significant: adjusted cell means and standard deviations or standard errors or confidence intervals (in table or interaction graph)
- effect sizes with confidence intervals for all effects

Die (entsprechend geforderte) Ausgabe der Prozedur liefert die entsprechenden Werte. Diagramme erleichtern die Interpretation und Präsentation. GLM- oder (M)AN(C)OVA-Befehle plus Optionen führen die Berechnungen durch.

3. Additional analyses

- evaluation of covariate effects
- evaluation of intercorrelations
- post hoc comparisons (if appropriate)
- unadjusted marginal and/or cell means (if significant main effect and/or interaction) if nonexperimental application

Signifikanztestungen bzw. Kovariate können in dem Sinne interpretiert werden, dass sie die abhängige Variable in einem bestimmten Maße "adjustieren". Kovariate, die nicht dazu beitragen, können in einem weiteren Schritt herausgenommen werden usw. (Tabachnick & Fidell, 2007: S.211 f.).

Checklist for Multivariate Analysis of Variance (Tabachnick & Fidell, 2007: S.292)

1. Issues

- unequal sample sizes and missing data

Eine Daumenregel besagt, dass man mehr Fälle als abhängige Variablen in jeder Zelle haben sollte (Tabachnick & Fidell, 2007: S.250).

- normality of sampling distributions

Multivariate Normalverteilung wird angenommen (z.B. für Signifikanztests); Nichtnormalverteilung kann unter bestimmten Bedingungen (z.B. bei entsprechender Zellgröße), auch ohne Datenmanipulation beibehalten werden, sofern robuste Ergebnisse erwartet werden können (Tabachnick & Fidell, 2007: S.251).

- outliers

Tests, um uni- und multivariate Ausreißer zu identifizieren, sollten separat für jede Zelle durchgeführt werden (Tabachnick & Fidell, 2007: S.251).

-homogeneity of variance-covariance-matrices

Diese Bedingung ist die Verallgemeinerung der Bedingung der Varianzhomogenität für einzelne abhängige Variablen (Tabachnick & Fidell, 2007: S.251).

-linearity

Vorliegen von Nichtlinearität wirkt sich auf die Testergebnisse aus (Tabachnick & Fidell, 2007: S.252).

-in stepdown analysis, when dependent variables act as covariates

-homogeneity of regression

Sofern Heterogenität in einem Schritt vorliegt, wird die restliche Analyse nicht zu interpretieren sein (Tabachnick & Fidell, 2007: S.251). Entsprechende Testungen werden für MANOVA und MANCOVA gerechnet und ausgegeben.

-reliability of dependent variables

-multicollinearity and singularity

Entsprechende redundante Variablen zu identifizieren, zu löschen u.a. wurde weiter oben für die Regression angedeutet.

2. major analyses: planned comparisons or omnibus F, when significant. Importance of dependent variables

-within cells correlations, stepdown F, univariate F

-effect sizes with confidence interval for significant stepdown F

-means or adjusted marginal and/or cell means for significant F, with standard deviations, standard errors, or confidence intervals

Entsprechende Ausgaben der Prozeduren (z.B. MANOVA) enthalten die (geforderten) Statistiken, Fehler, Signifikanzen, Testergebnisse etc.

3. multivariate effect size(s) with confidence interval(s) for planned comparison or omnibus F

4. additional analyses

-post hoc comparisons

-interpretation of independent variables - covariates interaction (if homogeneity of regression violated)

Checklist for Multivariate Analysis of Covariance (Tabachnick & Fidell, 2007: S.303)

1. Issues

- unequal sample sizes and missing data
- normality of sampling distributions
- outliers
- homogeneity of variance-covariance-matrices
- linearity
- homogeneity of regression
 - covariates
 - dependent variables for stepdown analysis
- reliability of covariates (and dependent variables for stepdown)
- multicollinearity and singularity

Entsprechende Hinweise zu diesem Block sind der MANOVA-Checkliste weiter oben in diesem Paper zu entnehmen.

2. Major analyses: planned comparisons or omnibus F; when significant: Importance of dependent variables

- within-cell correlations, stepdown F, univariate F
- effect size with its confidence interval for significant F, and standard deviations or standard errors or confidence intervals

Entsprechende Ausgaben der Prozeduren enthalten die (geforderten) Statistiken, Fehler, Signifikanzen etc.

3. multivariate effect size(s) with confidence interval(s) for planned comparison or omnibus F

4. additional analyses

- assessment of covariates

Multiple Regressionen für jede abhängige Variable mit den Kovariaten als Prädiktoren können die Stärke der Kovariate, die abhängige Variable zu strukturieren, aufzeigen helfen (Tabachnick & Fidell, 2007: S.296).

- interpretation of independent variables-covariates interaction (if homogeneity of regression violated for stepdown analysis)
- post hoc comparisons

Checklist for Profile Analysis (Tabachnick & Fidell, 2007: S.357)

1. Issues

- unequal sample sizes and missing data
- normality of sampling distributions
- outliers
- homogeneity of variance-covariance-matrices
- linearity
- multicollinearity and singularity

Weitere Angaben sind der MANOVA-Checkliste zu entnehmen, da die Profilanalyse soweit eine spezielle Manova darstellt.

2. Major analysis

- test for parallelism. If significant: Figure showing profile for deviation from parallelism

Die Frage lautet hier: haben verschiedene Gruppen parallele Profile (Tabachnick & Fidell, 2007: S.318 ff.).

- test for differences among levels, if appropriate. If significant: Marginal means for groups and standard deviations or standard errors or confidence intervals

Frage: unterscheiden sich die Gruppen(mittelwerte) signifikant voneinander?

- test for deviation from flatness, if appropriate. If significant: means for measures and standard deviations or standard errors or confidence intervals

- effect sizes with confidence limits for all three tests

Ausgaben enthalten die entsprechenden Ergebnisse.

3. Additional analyses

- planned comparisons
- post hoc comparisons appropriate for significant effect(s)
 - comparisons among groups
 - comparisons among measures
 - comparisons among measures within groups
- power analysis for nonsignificant effects

Vorgehensweise bei Diskriminanzanalysen (Backhaus et al, 1994: S.94):

1. Definition der Gruppen
2. Formulierung der Diskriminanzfunktion
3. Schätzung der Diskriminanzfunktion
4. Prüfung der Diskriminanzfunktion
5. Prüfung der Merkmalsvariablen
6. Klassifizierung von neuen Elementen

Anwendungsempfehlungen bei Diskriminanzanalysen (Backhaus et al, 1994: S.147)

1. Die Stichprobe darf keine Elemente enthalten, die gleichzeitig zu mehr als nur einer Gruppe gehören.
2. Der Umfang der Stichproben sollte wenigstens doppelt so groß sein wie die Anzahl der Variablen.
3. Die Anzahl der Merkmalsvariablen sollte größer sein als die Anzahl der Gruppen.
4. Zunächst sollte die (SPSS-)Schätzung (Optimierung) nach dem Wilkskriterium erfolgen, en bloc oder schrittweise.
5. Wenn Unsicherheit bezüglich der auszuwählenden Variablen besteht, sollte das RAO-Kriterium angewendet werden.
6. Soll eine Unterscheidung der am schlechtesten trennbaren Gruppen erreicht werden, sind die Kriterien MAHAL, MAXMINF oder MINRESID anzuwenden.
7. Grafische Darstellungen erleichtern die Interpretation. Eine Beschränkung auf zwei Diskriminanzfunktionen ist im Mehrgruppenfall von Vorteil.
8. Die Gleichheit der Gruppenstreuungen ist zu prüfen. Gegebenfalls sind die individuellen Gruppenstreuungen zu berücksichtigen. Es entfällt damit die Anwendbarkeit von Klassifizierungsfunktionen.
9. Im Mehrgruppenfall sollten nicht alle mathematisch möglichen, sondern nur die signifikanten und wichtigsten Diskriminanzfunktionen für die Klassifizierungen verwendet werden.
10. Bei ungleichen Kosten einer Fehlklassifikation muss die Klassifizierung auf Basis des Wahrscheinlichkeitskonzepts vorgenommen werden.

Checklist for Direct Discriminant Analysis (Tabachnick & Fidell, 2007: S.425)

1. Issues

-unequal sample sizes and missing data

Unterschiedliche Gruppengrößen stellen im weitesten Sinne kein Problem dar. Insofern müsste aber darüber entschieden werden, ob die Wahrscheinlichkeit einer Gruppenzuordnung, die Tatsache wiedergeben soll, dass diese mehr oder weniger gut in der Stichprobe enthalten ist. Die Fallzahl der kleinsten Gruppe sollte größer sein als die Anzahl der Prädiktoren (Tabachnick & Fidell, 2007: S.381).

-normality of sampling distributions

Wenn z.B. die Fallzahl gering ist, und die Gruppen ungleich verteilt sind, kann eine Transformation der Variablen in Betracht kommen (Tabachnick & Fidell, 2007: S.382).

-outliers

Univariate und multivariate Ausreißer sollten für jede Gruppe separat identifiziert und eventuell gestrichen werden (Tabachnick & Fidell, 2007: S.382).

-linearity

Nichtlineare Beziehungen der unabhängigen Variablen innerhalb der Gruppen wirken sich auf Testverfahren aus (Tabachnick & Fidell, 2007: S.383).

-homogeneity of variance-covariance-matrices

Daumenregeln: Variablentransformation dann, wenn keine Varianzhomogenität vorliegt, die Gruppen klein und ungleich verteilt sind, und Inferenz das Ziel der Prozedur ist. Wenn es um Klassifikation geht und die Streuungen heterogen sind, sollte man separate Kovarianzmatrizen und/oder quadratische Diskriminanzanalyse verwenden, solange die Fallzahlen groß sind, und die Variablen normalverteilt sind, und nichtparametrische Klassifikationen vorziehen, wenn die Variablen nicht normalverteilt sind, und / oder die Fallzahlen gering sind (Tabachnick & Fidell, 2007: S.383).

-multicollinearity and singularity

2. Major analysis

-significance of discriminant functions. If significant:

-variance accounted for and confidence limits for each significant function

-plot(s) of discriminant functions

-structure matrix

-effect size and confidence limits for solution

-variables separating each group with effect sizes and confidence limits

Die Ausgaben enthalten u.a. erklärte Varianzen durch die Funktionen, Wilks Lambda, die Koeffizienten und Klassifikationstabellen.

3. Additional analysis

Die Interpretation der Ergebnisse (Ladungen, Diskriminanzfunktion an sich etc.) ist nicht immer einfach, und bedarf weiterer Denk-, Rechen- oder Zeichenschritte.

-group means and standard deviations for high-loading variables

-pooled within-group correlations among predictor variables

-classification results

Verzerrungen treten auf, wenn der Koeffizient, welcher den Fall einer Gruppe zuordnet, aus dem Fall selbst abgeleitet wird (Tabachnick & Fidell, 2007: S.405).

-jackknifed classification

Jeder Fall hat hier Koeffizienten, die aus den anderen Fällen errechnet werden (Tabachnick & Fidell, 2007: S.405).

-cross-validation

-change in Rao's V (or stepdown F) plus univariate F for predictors

Anwendungsempfehlungen bei logistischen Regressionsanalysen (Backhaus et al, 2003: S.470 f.)

1. Die Fallzahl sollte pro Gruppe (=eine Ausprägung der abhängigen Variable) nicht kleiner als 25 sein.
2. Eine größere Zahl an unabhängigen Variablen verlangt auch nach höheren Beobachtungszahlen pro Gruppe.
3. Die unabhängigen Variablen sollten weitgehend frei von Multikollinearität sein.
4. Es sollte keine Autokorrelation vorliegen, sprich: die Beobachtungen sollten unabhängig voneinander sein.
5. Der logistische Wahrscheinlichkeitsverlauf sollte für die Fragestellung auch auf seine Plausibilität geprüft werden.
6. Der LR-Test zur Beurteilung der Signifikanz des Gesamtmodells ist unabhängig von der Struktur des Datensatzes immer geeignet.

7. Generell wird eine Ausreißerdiagnostik auf Basis der Pearson-Residuen pro Beobachtung empfohlen.

Checklist for Standard Logistic Regression with Dichotomous Outcome (Tabachnick & Fidell, 2007: S.479)

1. Issues

-ratio of cases to variables and missing data

Ein zu geringes Verhältnis kann zu übergroßen Schätzern und Fehlern führen. Variablenherausnahme oder Zusammenfügen von Zellen (Kategorien) kann Abhilfe schaffen (Tabachnick & Fidell, 2007: S.442).

-adequacy of expected frequencies (if necessary)

Variablenherausnahme oder Zusammenfügen von Zellen (Kategorien) kann Abhilfe schaffen (Tabachnick & Fidell, 2007: S.442).

-outliers in the solution (if fit inadequate)

-multicollinearity

Die Herausnahme redundanter Variablen aus dem Modell kann eine Lösung sein (Tabachnick & Fidell, 2007: S.443).

-linearity in the logit

Es wird angenommen, dass die Logittransformation der abhängigen Variable und stetige Prädiktoren in einer linearen Beziehung zueinander stehen (Tabachnick & Fidell, 2007: S.443). Zur Testung kann man grafische (Diagramme) oder statistische Verfahren nutzen.

2. Major analysis

-evaluation of overall fit. If inadequate:

-significance tests for each predictor

-parameter estimates

-effect size for model

-evaluation of models without predictors

Ausgaben enthalten u.a. Maße zur Modellgüte / Anpassung, Koeffizienten, Klassifikationstabellen.

3. Additional analysis

-odds ratios; classification or prediction success table

-interpretation in terms of means and / or percentages

Checklist for Sequential Logistic Regression with Multiple Outcomes (Tabachnick & Fidell, 2007: S. 496)

1. Issues

- ratio of cases to variables and missing data
- adequacy of expected frequencies (if necessary)
- outliers in the solution (if fit inadequate)
- multicollinearity
- linearity in the logit

Für die multinomiale Variante der einfachen logistischen Regression gelten ähnliche Bedingungen und Einschränkungen wie für letztere.

2. Major analysis

- evaluation of overall fit at each step
 - significance tests for each predictor at each step of interest
 - parameter estimates at each step of interest
 - effect size at each step of interest
- evaluation of improvement in model at each step

In den Ausgaben sind enthalten: Koeffizienten, Standardfehler, Maße der Modellgüte etc.

3. Additional analyses

- odds ratios
- classification and / or prediction success table
- interpretation in terms of means and / or percentages
- evaluation of models without individual predictors

Checklist for Predicting Survival from Covariates, Including Treatment (Tabachnick & Fidell, 2007: S.557)

1. Issues

-adequacy of sample sizes and missing data

Eine Fallzahl von 60 wird empfohlen, wenn 5 oder weniger Kovariate geschätzt werden sollen (Tabachnick & Fidell, 2007: S.509).

-normality of distributions

Das Vorliegen von Normalverteiltheit, Linearität und Homoskedastie ist nicht zwingend für die Überlebensanalyse, wirkt sich aber auf Tests aus, oder verbessert die Modellgüte (Tabachnick & Fidell, 2007: S.510).

-absence of outliers

Identifikation und Löschung von Ausreißern wurde weiter oben angesprochen.

-differences between withdrawn and remaining cases

-changes in survival experience over time

Es wird angenommen, dass die Faktoren, die das Überleben zu Beginn der Studie beeinflussen, es auch bis zum Ende (über mehrere Jahre) tun (Tabachnick & Fidell, 2007: S.510).

-proportionality of hazards

Im pr. Cox-Hazardmodell wird angenommen, dass die Form der Survivalfunktion über die Zeit für alle Fälle und Gruppen die gleiche ist (Tabachnick & Fidell, 2007: S.510). Grafische oder statistische Verfahren dienen zur Prüfung.

-multicollinearity

Entsprechende Kovariate, die redundant sind, werden aus dem Modell gelöscht (Tabachnick & Fidell, 2007: S.510).

2. Major analyses

-test of treatment effect, if significant:

-parameter estimates, including odds ratios

-effect size and confidence limits

-survival function showing groups separately

-effects of covariates, for significant ones:

-direction of effect(s)

-parameter estimates, including odds ratios

-effect size and confidence limits

(*SURVIVAL-* bzw. *COXREG-*)Ausgaben enthalten Überlebenstabellen, Koeffizienten, Signifikanzen Plots der Gruppen etc.

3. Additional analyses

-contingencies among covariates

-survival function based on covariates alone

Anwendungsempfehlungen bei Clusteranalysen (Backhaus et al, 1994: S.312 ff.)

Konkretisierung der Problemstellung der Untersuchung	Was ist das Ziel der Untersuchung? Welche Hypothesen sollen getestet werden?
Bestimmung der zu identifizierenden Objekte	Wie lassen sich die Untersuchungsobjekte beschreiben? Wie viele Objekte sollen berücksichtigt werden?
Auswahl der Variablen	Sollen qualitative und / oder quantitative Merkmale herangezogen werden? Wie groß soll die Zahl der Variablen sein? Ist eine Standardisierung sinnvoll?
Festlegung eines Ähnlichkeitsmaßes	Welches Maß soll gewählt werden? Wie sind gemischte Variablen zu behandeln?
Auswahl eines Algorithmus zur Gruppierung	Soll ein hierarchisches oder partitionierendes Verfahren gewählt werden? Welche Auswirkungen hat ein Wechsel des Algorithmus?
Bestimmung der Gruppenzahl	Wie viele Gruppen sollen gebildet werden? z.B. nach dem Elbow-Kriterium Wie verändern sich die Ergebnisse bei verschiedener Gruppenzahl?
Durchführung des Gruppierungsvorganges	
Analyse und Interpretation der Ergebnisse	Wie unterscheiden sich die ermittelten Cluster? Lassen sich die Ergebnisse sinnvoll interpretieren?

Folgende Punkte sind von Bedeutung:

1. Anzahl der Objekte (Repräsentativität?)
2. Problem der Ausreißer (als Objekte mit vollkommen anders gelagerten Kombination der Merkmalausprägungen im Vergleich zu den anderen; das Single-Linkage-Verfahren kann solche auffinden)
3. Anzahl zu betrachtender Variablen (theoretisch relevant)
4. Gewichtung der Variablen (Vorschalten einer explorativen Faktorenanalyse bei korrelierten Ausgangsdaten und angestrebte Gleichgewichtung der Variablen, Verwendung der Mahalanobis-Distanz; Ausschluss von hochkorrelierten Variablen)
5. Vergleichbarkeit der Merkmale (konstant vs. variabel, unterschiedliche Skalenniveaus vs. Standardisierung)

Checklist for Canonical Correlation (Tabachnick & Fidell, 2007: S.601)

1. Issues

- missing data
- normality, linearity, homoscedasticity

Unter diesen Annahmen wird die Qualität der Analyse z.B. in Hinblick auf Inferenz gesteigert (Tabachnick & Fidell, 2007: S.570 f.).

- outliers
- multicollinearity and singularity

Weiter oben, in nahezu jeder Checkliste, wurden Vorgehensweisen bei Ausreißer- oder Multikollinearitätsproblemen angedeutet.

2. Major analyses

- significance of canonical correlations
- correlations of variables and variates
- variance accounted for
 - by canonical correlations
 - by same-set canonical variates
 - by other-set canonical variates (redundancy)

Daumenregel, Checklisten und Anwendungsempfehlungen für Ökonometriker und Datenanalysten

Ausgaben enthalten Korrelationen, Varianzproportionen, Signifikanzen usw. Interpretation der Ergebnisse, wie auch allgemein der Prozedur, mag in Hinblick einer Präsentation besonders schwer fallen. Grafische Darstellungen, Pfaddiagramme u.ä. sollten hinzugenommen werden.

3. Additional analyses

-canonical coefficients

-canonical variates scores

Vorgehensweise bei Faktorenanalysen (Backhaus et al, 1994: S.198)

1. Variablenauswahl und Errechnung der Korrelationenmatrix
2. Extraktion der Faktoren
3. Bestimmung der Kommunalitäten
4. Zahl der Faktoren
5. Faktorinterpretation
6. Bestimmung der Faktorwerte

Empfehlungen zur Durchführung einer Faktorenanalyse (Backhaus et al, 1994: S.254)

Ausgangserhebung	Daten müssen metrisch skaliert sein (mindestens Intervallskala). Fallzahl sollte mindestens der dreifachen Variablenzahl entsprechen, mindestens aber der Zahl der Variablen.
Erstellen der Ausgangsdatenmatrix	
Errechnen der Korrelationsmatrix	
Kommunalitätenschätzung	Eigene Vorgaben Iterative Schätzung
Faktorextraktion	Hauptachsenanalyse (Wie lässt sich die Ursache bezeichnen, die für die hohen Ladungen der Variablen auf diesen Faktor verantwortlich ist) nimmt Werte kleiner 1 bei der Kommunalitätenschätzung.

	Hauptkomponentenanalyse (Wie lassen sich die auf einen Faktor hoch ladenden Variablen durch einen Sammelbegriff (Komponente) zusammenfassen) nimmt 1 als Startwert bei der Kommunalitätenschätzung (vgl. Backhaus et al, 1994:S.221 f.).
Bestimmung der Faktorenzahl	Kaiser-Kriterium; Screeplot
Rotation	Varimax-Kriterium
Interpretation	Höchstens Faktorladungen größer 0,5 verwenden.
Bestimmung der Faktorwerte	Regressionsschätzung

Checklist for Factor Analysis (Tabachnick & Fidell, 2007: S.670)

1. Limitations

-outliers among cases

-sample size and missing data

Daumenregel: man habe mindestens 300 Fälle für die Faktorenanalyse (Tabachnick & Fidell, 2007: S.613).

-factorability of R (as a correlation matrix)

Falls in R keine Korrelation über 0,3 liegt, sollte man über die Durchführung einer Faktorenanalyse nachdenken (Tabachnick & Fidell, 2007: S.614).

-normality and linearity of variables

Falls entsprechende Schiefe und Kurtosis vorliegen, oder Nichtlinearität vorgefunden wird, sollte man über Variablentransformation nachdenken (Tabachnick & Fidell, 2007: S.613).

-multicollinearity and singularity

In der Hauptkomponentenanalyse wird nicht die Inverse einer Matrix genommen. Für andere Typen von Faktorenanalysen wird Multikollinearität in dieser Hinsicht zum Problem, und sollte entsprechend gehandhabt werden (Tabachnick & Fidell, 2007: S.614).

-outliers among variables

Eine Variable, die so gut wie gar nicht mit allen anderen Variablen und zentralen Faktoren korreliert, ist in diesem Sinne ein Ausreißer, und sollte ignoriert werden (Tabachnick & Fidell, 2007: S.615).

2. Major analyses

- number of factors
- nature of factors
- type of rotation
- importance of factors

Man könnte die FA mit einer Kombination aus “principal components extraction” und “Varimaxrotation” beginnen. Die Ergebnisse kann man nutzen, die (korrekte) Anzahl von Faktoren zu bestimmen, oder Variablen, die man ausschließen sollte, zu identifizieren. Alternative Vorgehensweisen sollten durchgerechnet werden (Tabachnick & Fidell, 2007: S.642).

3. Additional analyses

- factor scores
- distinguishability and simplicity of factors
- complexity of variables
- internal consistency of factors
- outlying cases among the factors

Ausführungen zu jedem Punkt der Analyse (von der Nichtnormalverteilung bis zu den Interfaktorenkorrelationen) können in der Gesamtheit ein zerfahrenes Bild hinterlassen. Diagramme / Plots usw. erleichtern die (argumentierende) Präsentation.

Anwendungsempfehlungen bei Kausalanalysen mit LISREL (Backhaus et al, 1994: S.323 ff.)

Ablaufschritte (Backhaus et al, 1994: S.361 f.)

1. Hypothesenbildung
2. Erstellung eines Pfaddiagramms
3. Spezifikation der Modellstruktur
4. Identifikation der Modellstruktur
5. Parameterschätzungen
6. Beurteilung der Schätzergebnisse
7. Modifikation der Modellstruktur

- Das Strukturmodell bildet die theoretisch vermuteten Zusammenhänge zwischen den latenten Variablen ab. Dabei werden die endogenen Variablen durch die im Modell unterstellten kausalen Beziehungen erklärt, während die exogenen Variablen als erklärende Größen dienen.
- Das Messmodell der latenten exogenen Variablen enthält empirische Indikatoren, die zur Operationalisierung der exogenen Variablen dienen, und spiegelt die vermuteten Zusammenhänge zwischen diesen Indikatoren und den exogenen Variablen wider.
- Das Messmodell der latenten endogenen Variablen enthält empirische Indikatoren, die die endogenen Variablen operationalisieren, und gibt die entsprechenden Zusammenhänge wider.

Empfehlungen zur Erstellung eines Pfaddiagramms (Backhaus et al, 1994: S.367 ff.)

Allgemeine Konstruktionsregeln

1. Eine kausale Beziehung zwischen 2 Variablen wird immer durch einen geraden Pfeil dargestellt.
2. Ein Pfeil hat seinen Ursprung immer bei der unabhängigen und endet bei der abhängigen Variable.
3. Pfeife haben nur einen Ursprung und einen Endpunkt.
4. Je-desto-Hypothesen beschreiben kausale Beziehungen, wobei die zu Anfang genannte Größe immer die verursachende und die zuletzt genannte Größe immer die kausal abhängige ist.
5. Der Einfluss von Residual- bzw. Messfehlervariablen wird durch Pfeile, welche ihren Ursprung in diesen haben, dargestellt.
6. Nicht kausal interpretierbare Beziehungen werden durch gekrümmte Doppelpfeile dargestellt (nur zwischen latenten exogenen Variablen oder zwischen Messfehlervariablen zulässig).

Allgemeine Bezeichnungs- und Darstellungsweisen

7. Die Stärke kausaler oder nicht kausal interpretierbare Beziehungen wird durch griechische Kleinbuchstaben mit zwei Zahlenindices versehen.
8. Bei kausalen Beziehungen gibt der erste Index die abhängige Variable und der zweite die unabhängige Variable.
9. Direkt beobachtbare (Mess-)Variablen werden in Rechtecken dargestellt, latente Variablen durch Kreise angedeutet, und Messfehlervariablen ohne Umrandung gezeigt.

10. Beziehungen zwischen Indikatorvariablen und latenten Variablen können durch gestrichelte Pfeile dargestellt werden, um deutlich zu machen, dass sie wie eine kausale Beziehung behandelt werden (Indikatorvariablen als abhängige, latente als unabhängige Variablen).

Konstruktionsregeln im vollständigen Modell

11. Ein vollständiges Modell besteht aus zwei Messmodellen und einem Strukturmodell.
12. Das Pfaddiagramm ist wie folgt aufgebaut:
 - Links steht das Messmodell der latenten exogenen Variablen (x und ξ und deren Beziehungen).
 - In der Mitte ist das Strukturmodell (η, ξ und deren Beziehungen).
 - Rechts ist das Messmodell der latenten endogenen Variablen (y, η und deren Beziehungen).
13. Für die Kennzeichnung kausaler Beziehungen werden verwendet: β, γ, λ .
14. Kausal nicht interpretierbare Beziehungen zwischen exogenen Variablen werden durch das Symbol Φ gezeigt.

Empfehlungen zur mathematischen Formulierung des Pfaddiagramms

Als Gleichungssystem:

1. Für jede abhängige Variable (x, y, η) lässt sich genau eine Gleichung formulieren.
2. Abhängige Variable sind solche Variablen, auf die ein Pfeil hinzeigt.
3. Variablen auf die ein Pfeil hinzeigt, stehen links vom Gleichheitszeichen und solche, von denen ein Pfeil ausgeht, stehen rechts vom Gleichheitszeichen.
4. Die Pfeile des Pfaddiagramms werden mathematisch durch Pfadkoeffizienten repräsentiert, deren Größe die Stärke des jeweiligen Zusammenhangs angibt.
5. Werden abhängige Variablen (x, y, η) von mehreren unabhängigen Variablen beeinflusst, werden letztere linear additiv verknüpft.

In Matrizenschreibweise:

1. Ein vollständiges Modell besteht aus drei Matrizengleichungen: zwei für die Messmodelle und eine für das Strukturmodell.
2. Die Koeffizienten zwischen je zwei Variablengruppen werden in einer Matrix zusammengefasst, wobei alle Matrizen durch griechische Großbuchstaben entsprechen den Bezeichnungen der Koeffizienten gekennzeichnet werden.

3. Die Variablen selbst werden als Spaltenvektoren aufgefasst, und zur Kennzeichnung werden die griechischen Kleinbuchstaben beibehalten.

Annahmen (Backhaus et al, 1994: S.425 ff.)

1. ML-Schätzung und GLS-Methode setzen voraus, dass die beobachteten Variablen x und y einer Multinormalverteilung folgen. Wenn als Schätzverfahren ULS, WLS- oder DWLS gewählt werden, ist die Annahme nicht erforderlich.
2. Messmodelle gleichen Modellen der Faktorenanalyse / die Messfehlervariablen sind nicht mit den hypothetischen Konstrukten und auch nicht untereinander korreliert.
3. Für das Strukturmodell wird angenommen, dass die Residuen nicht mit den exogenen latenten Variablen korrelieren und deren Erwartungswerte Null sind.
4. Es besteht keine Korrelation zwischen Messfehlern und den Residuen der Strukturgleichungen oder anderen Konstrukten.
5. Es wird Linearität und Additivität der Konstrukte und Messhypothesen unterstellt.
6. Damit die Parameterwerte geschätzt werden können, muss die modelltheoretische Kovarianz-Matrix positiv definit sein. Das Modell muss identifizierbar sein.

Checklist for Structural Equations Modeling (Tabachnick & Fidell, 2007: S.769)

1. Issues

- sample size and missing data
- normality of sampling distributions
- outliers
- linearity
- adequacy of covariances
- identification
- path diagram-hypothesized model
- estimation method

Vorschläge zu diesem Block kann man weiter oben in diesem Paper finden.

2. Major analyses

- assessment of fit
- residuals

-model χ^2

-fit indices

-significance of specific parameters

-variance in a variable accounted for by a factor

3. Additional analyses

-lagrange multiplier test

-tests of specific parameters

-addition of parameters to improve fit

-wald test for dropping parameters

-correlation between hypothesized and final model or cross-validate model

-diagram-final model

Checklist for Multilevel Modeling (Tabachnick & Fidell, 2007: S. 849)

Die Erweiterung dieser Modelle auf mixed-effects-Modelle im Sinne von Longitudinaldaten mit unterschiedlichen Messzeitpunkten ist einleuchtend.

1. Issues

-adequacy of sample size and missing data

-normality of distributions at each level

-absence of outliers at each level

-absence of multicollinearity and singularity

-independence of errors (intraclass correlation)

2. Major analyses

-analysis with first-level predictors

-analysis with second-level predictors and significant first-level predictors, etc.

-determination of final model

-parameter estimates for final model

-comparison of final with an intercepts-only model

3. Additional analyses

-adding main effects and / or interactions

-additional exploratory analyses

Checklist for Hierarchical Multiway Frequency Analysis (Tabachnick & Fidell, 2007: S.906)

1. Issues

- adequacy of expected frequencies
- outliers in the solution

2. Major analysis

- model screening
- model selection
- evaluation of overall fit. If adequate:
 - significance tests for each model effect and their confidence intervals
 - parameter estimates

3. Additional analyses

- interpretation via proportions
- identifying extreme cells (if fit inadequate)

Allgemein bieten sich für Kontingenzanalysen in SPSS die Befehle crosstabs (inklusive eines Signifikanztests), genlog, loglinear, hiloglinear oder correspondence an. Das Vorgehen bei einer Korrespondenzanalyse teilen Backhaus et al. (2003) in folgende Bereiche ein (vgl. Backhaus et al. 2003: S.680 ff.):

- vorbereitende Schritte*
- Standardisierung der Daten*
- Extraktion der Dimensionen*
- Normalisierung der Koordinaten*
- Interpretation.*

Ablauf einer multidimensionalen Skalierung (Backhaus et al, 1994: S. 441 ff.)

1. Messung von Ähnlichkeiten

- Methode der Rangreihung
- die Ankerpunktmethode
- das Ratingverfahren

2. Wahl des Distanzmodells

- euklidische Metrik
- die City-Block-Metrik

-allgemein die Minkowski-Metrik

3. Ermittlung der Konfiguration
4. Zahl und Interpretation der Dimensionen
 - der Verdichtungskoeffizient
 - der STRESS-Wert
 - nach Interpretierbarkeit
5. Aggregation von Personen
 - Liegt Homogenität der Personen vor?

Anwendungsempfehlungen bei MDS (Backhaus et al, 1994: S.485 f.)

1. Die Zahl der Objekte sollte nicht zu klein sein (mehr als acht).
2. Die Erhebung der Ähnlichkeitsdaten wird durch Anwendung des Ratingverfahrens erleichtert. Für individuelle Analysen sind in der Regel Rangdaten erforderlich.
3. Bei der Wahl des Distanzmodells sollte die euklidische Metrik bevorzugt werden.
4. Es sollten nicht mehr als zwei oder drei Dimensionen vorgegeben werden.
5. Für aggregierte Analysen ist ein Verfahren mit Replikationen zu bevorzugen.
6. Zur Erleichterung der Interpretation sollten die Achsen geeignet rotiert werden.
7. Sachkenntnis ermöglicht eine sinnvolle Interpretation der Ergebnisse.

Ablaufschritte einer Conjoint-Measurement-Analyse (Backhaus et al, 1994: S.543)

1. Eigenschaften und Eigenschaftsausprägungen: Anzahl ist möglichst gering zu halten. Es sollen von einander unabhängige Eigenschaften sein, die für die Untersuchung relevant und beim Produktdesign konkret umsetzbar sind. Die einzelnen Ausprägungen stehen dabei in einer kompensatorischen Beziehung.
2. Erhebungsdesign: im Erhebungsdesign sollten nicht mehr als maximal 20 fiktive Produkte enthalten sein. Wird die Zahl überschritten, so sollte ein reduziertes Design unter Verwendung der Profilmethode erstellt werden.
3. Bewertung der Stimuli (Rangreihung)
4. Schätzung der Nutzenwerte: ein additives Nutzenmodell sollte zugrunde liegen; bei schlechten STRESS-Werten, die mangelnde Anpassung signalisieren, kann über eine veränderte Ausgangskonfiguration vielleicht eine Verbesserung der Lösung erreicht werden.

5. Aggregation der Nutzenwerte: die gemeinsame Conjoint-Analyse kann zu einer größeren Differenzierung der Teilnutzenwerte einzelner Eigenschaften / zu besser interpretierbaren Ergebnissen führen; wenn die Anzahl der Daten nicht zu groß ist, ist die gemeinsame Analyse der Aggregation der Einzelanalysen vorzuziehen.
6. Segmentation: eine Aggregation oder gemeinsame Analyse über alle Personen ist nur bei hinreichender Homogenität der individuellen Teilnutzenwerte gerechtfertigt. Mit einer Clusteranalyse sollte diese geprüft werden. Bei ausgeprägter Heterogenität sind segmentspezifische Analysen durchzuführen.

Verwendete Literatur:

Backhaus, Klaus et al., 1994: Multivariate Analysemethoden, Berlin: Springer Verlag.

Backhaus, Klaus et al., 2003: Multivariate Analysemethoden, Berlin: Springer Verlag.

Cabrera, Javier & McDougall, Andrew, 2002: Statistical Consulting. New York: Springer Verlag.

Tabachnick, Barbara G. & Fidell, Linda S., 2007: Using Multivariate Statistics. Boston: Allyn & Bacon.

Van Belle, Geralt, 2002: Statistical Rules of Thumb. New York: John Wiley & Sons.