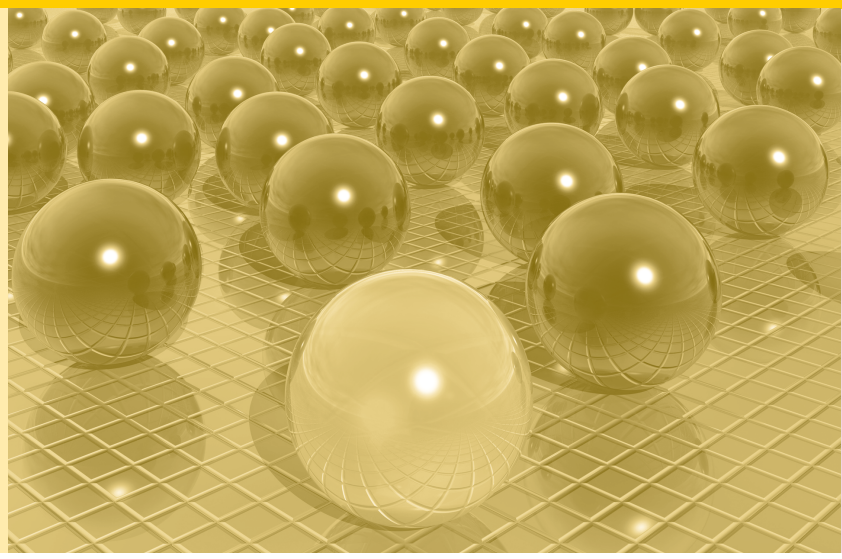
FDZ-Arbeitspapier
Nr. 33



Remote Access.
Eine Welt ohne Mikrodaten ??

Gerd Ronning, Philipp Bleninger, Jörg Drechsler,
Christopher Gürke

2011

STATISTISCHE ÄMTER
DES BUNDES UND DER LÄNDER
FORSCHUNGSDATENZENTREN

Impressum

Herausgeber: Statistische Ämter des Bundes und der Länder
Herstellung: Statistisches Bundesamt, Forschungsdatenzentrum
65180 Wiesbaden
Telefon 0611 75-4220 • Telefax 0611 75-3915
Internet: www.forschungsdatenzentrum.de
E-Mail: forschungsdatenzentrum@destatis.de

Fachliche Informationen
zu dieser Veröffentlichung:

Statistisches Bundesamt
Forschungsdatenzentrum

Tel.: 0611 75-4220
Fax: 0611 72-3915
forschungsdatenzentrum@destatis.de

Informationen zum Datenangebot:

Statistisches Bundesamt
Forschungsdatenzentrum

Tel.: 0611 75-4220
Fax: 0611 72-3915
forschungsdatenzentrum@destatis.de

Forschungsdatenzentrum der
Statistischen Ämter der Länder
– Geschäftsstelle –
Tel.: 0211 9449-2876
Fax: 0211 9449-8087
forschungsdatenzentrum@it.nrw.de

Erscheinungsfolge: unregelmäßig
Erschienen im Februar 2011

Diese Publikation wird kostenlos als **PDF-Datei** zum Download unter www.forschungsdatenzentrum.de angeboten.

Bei den enthaltenen statistischen Angaben handelt es sich um eigene Arbeitsergebnisse der genannten Autoren im Zusammenhang mit der Nutzung der Forschungsdatenzentren. Es handelt sich hierbei ausdrücklich nicht um Ergebnisse der Statistischen Ämter des Bundes und der Länder.

FDZ-Arbeitspapier
Nr. 33

Remote Access.
Eine Welt ohne Mikrodaten ??

Gerd Ronning, Philipp Bleninger, Jörg Drechsler,
Christopher Gürke

2011

STATISTISCHE ÄMTER
DES BUNDES UND DER LÄNDER
FORSCHUNGSDATENZENTREN

Remote Access.[1]
Eine Welt ohne Mikrodaten ??

GERD RONNING[2], PHILPP BLENINGER[3], JÖRG DRECHSLER[4]
und
CHRISTOPHER GÜRKE[5]

(Stand: 26. Januar 2011 - Version 18)

Key Words: Remote execution, confidential tabular data, saturated models, disclosure risk, artificial outlier, strategic dummy variable, cluster analysis, factor analysis, MDS.

Zusammenfassung

Die statistische Analyse mittels Fernrechnen (Remote Access bzw. Remote Execution) bietet die Möglichkeit, die Mikrodaten in originaler Form zu verwenden. Allerdings ist dabei zu beachten, dass bestimmte statistische Prozeduren, etwa die Verwendung bestimmter Schein-Variablen (dummy variables, fixed effects) durchaus ein Enthüllungsrisiko ("Inferenz-Enthüllung") darstellen und deshalb deren Ausführung durch den Server, an den die Do-Files gesandt werden, verhindert werden müssen. In dieser Arbeit werden die wichtigsten multivariaten Analysemethoden auf derartige Enthüllungsrisiken untersucht.

Abstract

Use of microdata is severely hampered in many areas of research. This is in particular true for data from statistical offices. One way to circumvent this problem is to anonymize

[1]Die Ergebnisse in diesem Papier stammen aus dem Projekt "Eine informationelle Infrastruktur für das ‚E-Science Age' Auf dem Weg zum ‚Remote-Access' – Verbesserung der kontrollierten Datenfernverarbeitung bei wirtschaftsstatistischen Daten durch Datenstrukturfiles und automatisierte Ergebniskontrolle", das vom Bundesministerium für Bildung und Forschung finanziell gefördert wird. Die empirischen Ergebnisse wurden teilweise bereits in einem anderen Manuskript verwendet: P. Bleninger, J. Drechsler und G. Ronning. "Remote data access and the risk of disclosure from linear regression: An empirical study" submitted for presentation at the conference 'PRIVACY IN STATISTICAL DATABASES 2010 (PSD 2010)', May 2010. Gerhard Tutz sind wir zu Dank für die Überlassung eines unveröffentlichten Manuskript mit Ergebnissen zu saturierten Modellen verpflichtet.

[2]Wirtschaftswissenschaftliche Fakultät, Universität Tübingen, D-72074 Tübingen. email: gerd.ronning@uni-tuebingen.de.

[3]IAB , Regensburger Str. 104, D-90478 Nürnberg, email: philipp.bleninger@iab.de

[4]IAB , Regensburger Str. 104, D-90478 Nürnberg, email: joerg.drechsler@iab.de

[5]Statistisches Bundesamt, Forschungsdatenzentrum, Gustav-Stresemann-Ring, 11, D-65189 Wiesbaden, email: christopher.guerke@destatis.de

the data such that both confidentiality is guaranteed and informational content of the data is not to much distorted by the anonymization procedure. However many researchers prefer the use of 'original' data. Therefore in recent years remote access/execution ('Fernrechnen') has become quite popular where the original micro data are used in the statistical analysis but are not available to the researchers. Clearly, this alternative takes more time since program files have to be sent to the statistical office. However, the euphoria for this approach has cooled down a bit since it has become apparent that here also problems of confidentiality exist. Most obvious is the fact that residuals cannot be provided. See, for example, Gomatam et al. (2005). However, there are very different kinds of 'disclosures' which are discussed in the paper. The paper also draws attention to the use of saturated models which bear the risk of reproducing confidential tabular data. Analysis of variance is the relevant tool in reproducing magnitude tables whereas the corresponding micro-econometric models can be used to reproduce frequency tables: Logit models give the results in case of a nominal variable and Poisson regression is the approach in case of count data.We also shortly discuss possible disclosure risk in the standard multivariate procedures (factor analysis, principal components, cluster analysis and multidimensional scaling).

It is clear from the many examples given in the paper that the remote access/execution option will ask for a large amount of statistical expertise in the statistical office in order to check for disclosure risk. Additionally, there will be a tendency not to provide statistical results to the researcher if critical variables such as region or sector are demanded as regressors in the program file. Perhaps a much cruder classification of regions and sectors will be allowed which in a way is the situation used in providing anonymized data.

Inhaltsverzeichnis

1 Einleitung

Da die Weitergabe von anonymisierten Mikrodaten auch deren Aussagegehalt beeinträchtigen, ist es besser, gar nicht die Mikrodaten herauszugeben, sondern nur Ergebnisse, die auf diesen Mikrodaten basieren. Neben der Weitergabe von aggregierten Kennzahlen und Tabellen ist vor allem die Weitergabe von Ergebnissen aus statistischen Analysen (etwa Regressionsanalyse oder Logitanalyse) von Bedeutung.[6] Dieser zweite Ansatz wird als "Remote Access" bezeichnet.[7] Dabei wird in dieser Arbeit unter Remote Access die Situation verstanden, dass ausschließlich Ergebnisse der statistischen Analyse, nicht aber die zugrundeliegenden Mikrodaten dem Nutzer zugänglich gemacht werden.[8]

In dieser "Welt ohne Mikrodaten" gibt es jedoch weiterhin das Problem der Reidentifikation, vor allem dann, wenn die Datennutzer gezielt statistische Analysen darauf ausrichten. Ein prominentes Beispiel ist die Anforderung des minimalen und vor allem des maximalen Wertes für ein bestimmtes Merkmal.[9] Eine systematische Diskussion dazu findet sich in den Abschnitten 3 und 4. Aber oft genug kann schon eine enge Beziehung sprich hohe Korrelation zwischen zwei (stetigen) Merkmalen – eins vertraulich und eins öffentlich zugänglich – auch ein Enthüllungsrisiko (disclusure risk) bedeuten, da zumindest eine sehr gute Schätzung des vertraulichen Merkmals möglich ist. Gomatam et al. (2005) bezeichnen dies als "**Inferenz-Enthüllung**" (inferential disclosure). Daneben betrachten sie noch die Enthüllung der Identität (identity disclosure), beispielsweise durch die Adresse oder das Geburtsdatum, und die Enthüllung eines Attributes (attribute disclosure), beispielsweise das Einkommen einer Person.

2 Ein formales Modell für Remote Access

2.1 Vorbemerkung

Um eine sinnvolle Remote-Access-Strategie (RA-Strategie)[10] zu bestimmen, sind zunächst die Bestandteile/Komponenten des Entscheidungsproblems zu beschreiben. Dies soll im folgenden Unterabschnitt geschehen. Ziel ist es, für einzelne Strategien das Enthüllungs-Risiko wie auch den Daten-Nutzen zu bewerten, um daraus die optimale Strategie zu bestimmen. Dieser Abschnitt basiert weitgehend auf Ergebnissen aus der Arbeit von S. Gomatam und Koautoren (Gomatam et al. 2005).

[6] Gomatam et al. (2005) erwähnen als dritte Möglichkeit den Ansatz von Rubin, bei dem nur synthetische Daten zur Verfügung gestellt werden. Unseres Erachtens ist dieser Ansatz aber eher den Anonymsierungsansätzen zuzurechen.

[7] Siehe auch Keller-McNulty und Unger (1998), Duncan und Mukherjee (2000) sowie Schouten und Cigrang (2000).

[8] Dies wird in neuester Zeit als "Remote Execution" bezeichnet. Mit Verwendung dieses Begriffs soll vor allem auf das Risiko aufmerksam gemacht werden, dass unter Remote Access möglicherweise Mikrodaten auf dem Bildschirm erscheinen und abfotografiert werden können, beispielsweise mit einem Mobiltelefon.

[9] In der Sitzung der Arbeitsgruppe im Juli 2009 stellte Herr Gürke eine Arbeit von Reznek und Riggs (2005) vor. In dieser wurde diese Problematik bei Verwendung von Regressionen mit speziellen Dummy-Variablen untersucht. Siehe dazu auch Abschnitt 3.2.1.

[10] Gomatam et al. (2005, S. 165) sprechen von SDL-Strategien, wobei SDL für **S**tatistical **D**isclosure **L**imitation steht.

2.2 Das Modell

Es sei $\mathcal{D}$ der zu analysierende Datensatz. Der zu installierende Server ist ein Software-System, das Funktionen $F(\mathcal{D})$ dieses Datensatzes als Output hat. Das können beispielsweise statistische Kennzahlen wie Mittelwert und Varianz für ein bestimmtes Merkmal oder auch die Regressionskoeffizienten einer ökonometrischen Analyse sein. Der Server erhält vom Datennutzer eine Anfrage (englisch query) Q für einen bestimmten Output $F(\mathcal{D})$. Der Server hat dann zu entscheiden, ob er dieser Anfrage entspricht oder nicht. Man könnte sich auch vorstellen, dass der Server eine alternative Analyse vorschlägt. Bei der Regressionsanalyse mit einem "kritischen" Dummy könnte dies beispielsweise eine Analyse ohne diesen Dummy sein.

Die formale Analyse benötigt ferner die folgenden Komponenten:

- Die Menge $\mathcal{Q}$ der Anfragen, die der Rechner bearbeiten kann, d.h. es wird unterstellt, dass der Rechner diese Mengen an Anfragen gespeichert hat und sie als risikoreich bzw. risikolos bewertet (hat).
- Der Aktionsraum $\mathcal{A}$ der möglichen Aktionen A, wobei diese Menge eine Untermenge des Anfragenraums ist, d.h. $\mathcal{A} \preceq \mathcal{Q}$. Falls $A = Q$ gilt, wird die Anfrage Q vom Server beantwortet.[11]
- Das Maß $R(Q_1, Q_2, \ldots, Q_m)$ mißt das Enthüllungsrisiko für die Bereitstellung von $(F_1(\mathcal{D}), F_2(\mathcal{D}), \ldots, F_m(\mathcal{D}))$ für das Bündel von Anfragen $Q_1, Q_2, \ldots, Q_m$.
- Entsprechend mißt $U(Q_1, Q_2, \ldots, Q_m)$ den Forschungsnutzen bei Bereitstellung von $(F_1(\mathcal{D}), F_2(\mathcal{D}), \ldots, F_m(\mathcal{D}))$ für das Bündel von Anfragen $Q_1, Q_2, \ldots, Q_m$.

Die Menge $\mathcal{A}$ der möglichen Aktionen ergibt sich typischerweise durch Maximierung von $U(\mathcal{A})$ bezüglich der Menge $\mathcal{A}$ bei Beachtung einer Obergrenze für das Risiko $R(\mathcal{A})$ oder auch durch Auswahl einer Menge $\mathcal{A}$, deren Elemente sämtlich höheren Nutzen und geringeres Risiko haben als die Elemente in der Komplementärmenge der möglichen Aktionen. Dies wird später in Abschnitt 2.7 genauer erläutert.

2.2.1 Statischer und dynamischer Fall

Das Problem an dieser Modellierung ist die **Interaktion zwischen verschiedenen Anfragen**, denn das Enthüllungsrisiko für Q_1 und Q_2 gemeinsam kann durchaus höher sein als für einzelne Anfragen Q_1 und Q_2 (von verschiedenen Personen) . Solange die Analyse **statisch** ist, d.h. ein einziges Mal zu einem bestimmten Zeitpunkt eine Menge von Anfragen an den Server gerichtet wird, ist die mögliche Interaktion vermutlich technisch beherrschbar, soll heißen, das dadurch erhöhte Risiko wird vom Server bewertet.

Deutlich weniger einfach ist das Problem in einer **dynamischen** Analyse zu lösen, bei der Anfragen eines bestimmten Datennutzers **sequentiell** eingehen, was sicher den Regelfall darstellen wird! Denn nun muß für jeden Nutzer notiert werden, welche Anfragen bereits beantwortet wurden und welche Interaktionen die neuen Anfragen mit den alten haben.

[11] Gomatam et al. (2005) sprechen von einem Antwort-Raum $\mathcal{A}$.

Bezüglich der Realisierbarkeit eines solchen dynamischen Servers schreiben Gomatam et al. (2005, S. 166): "Whether dynamic servers are possible remains an open question."

2.2.2 Tabellen-Analyse

Allerdings ist diese sehr pessimistische Aussage zu relativieren, wenn man die spezielle Klasse der **Tabellen-Server** betrachtet, d.h. sich auf die Anfrage nach Tabellen beschränkt.[12] In diesem Fall ist $\mathcal{D}$ eine (sehr) große Kontingenztabelle(Häufigkeitstabelle, frequency table) bzw. eine entsprechende Tabelle mit Summen für die einzelnen Zellen ("Werte-Tabelle").[13]

Häufigkeitstabellen: Im Fall von Kontingenztabellen entspricht die Dimension der Tabelle $\mathcal{D}$ der Anzahl (**nominaler!!**) Merkmale im Datensatz. Die (möglicherweise ebenfalls analysierten) stetigen Variablen sind dabei "geeignet" zu Klassen zusammengefaßt, wobei die Zusammenfassung aus Sicherheitsgründen eher grob gestaltet wird. In den einzelnen Zellen stehen die jeweiligen (absoluten) Häufigkeiten für ein binäres (bzw. allgemeiner nominal skaliertes) Merkmal.

Häufigkeitstabellen für Zähldaten: Falls in der Tabelle die Häufigkeiten für ein Merkmal, dessen Ausprägungen die Menge der natürlichen Zahlen sind, dargestellt sind, spricht man ebenfalls von Häufigkeitstabellen. Die sei hier um den Zusatz "Zähldaten" ergänzt. [14]

Werte-Tabellen: Im Fall der Tabellen, die Summen enthalten, beispielsweise die jeweils gezahlte Umsatzsteuer, ist die Dimension der Tabelle $\mathcal{D}$ höchstens gleich $r-1$, wenn r die Anzahl der Merkmale im Datensatz ist. Denn für das r-te Merkmal werden ja die jeweiligen Summen in den Zellen berichtet. Wenn es $r_s < r$ stetige Merkmale und $r_d = r - r_s$ diskrete Merkmale im Datensatz gibt, dann können r_s verschiedene Tabellen mit Summen gebildet werden.

Ferner ist $\mathcal{Q}$ die Menge aller Untertabellen, die sich durch Herausschneiden oder Aggregieren daraus bilden lassen. Allerdings ist selbst in diesem "überschaubaren" Fall die Antwort auf die Frage nach dem Enthüllungsriskiko alles andere als einfach. Siehe beispielsweise die Arbeiten von Sarah Gießing, etwa Gießing und Dittrich (2005). Die dabei untersuchten Primär- und Sekundärsperrungen schränken den Anfragen-Raum $\mathcal{Q}$ auf den Aktions-Raum $\mathcal{A}$ ein.

Eine ausführliche Analyse des Enthüllungs-Risikos für Tabellen durch Remote Access findet sich in Abschnitt 3.4. Den Fall Werte-Tabellen behandelt Abschnitt 3.4.1. Auf das Risiko der Inferenzenthüllung bei der Verwendung von Häufigkeitstabellen in der Logit- und Probit-Analyse geht Abschnitt 3.4.2 ein. Die Enthüllung von Häufigkeitstabellen für Zähldaten wird in Abschnitt 3.4.3 behandelt.

[12] Gomatam et al. (2005) weisen in diesem Zusammenhang auf die Arbeiten von Dobra et al. (2002), Dobra, Karr und Sanil (2003) sowie Karr Dobra und Sanil (2003) hin.

[13] Gomatam et al. (2005) sprechen auch im zweiten Fall von einer Kontingenztabelle, obwohl der Begriff eigentlich nur verwendet wird, wenn in der Tabelle Häufigkeiten betrachtet werden. Im 'Handbook on Statistical Disclosure Control' wird zwischen 'frequency tables' und 'magnitude tables' unterschieden. Siehe Hundepool et al. (2009).

[14] Die Unterschiedung zwischen den beiden Tabellenarten wird in Abschnitt 3.4 relevant.

2.3 Modell-Server

Im Folgenden wird unterstellt, dass der Anfragen-Raum $\mathcal{Q}$ nur Anfragen für **statistische Modelle** enthält, die aus einer (oder mehreren) abhängigen Variablen (Gomotam et al. (2005) sprechen von Prediktor-Variablen) sowie einem Set von Einflußvariablen bestehen. Dabei können die Einflußvariablen sowohl stetig als auch diskret sein. Der Server, der Anfragen für diese Modelle beantwortet, wird als **Modell-Server** bezeichnet.

Die Antworten zu den Anfragen enthalten die Koeffizientenschätzungen, deren Standardabweichungen bzw. die geschätzte Kovarianz-Matrix des Koeffizientenvektors, Maße für die Güte der Anpassung sowie Ergebnisse für bestimmte Spezifikations-Tests. Es soll unterstellt werden, dass für alle Merkmale in $\mathcal{D}$ die arithmetischen Mittel und empirischen Varianzen verfügbar sprich abrufbar sind.

2.4 Rezepte gegen Identitäts- und Attribut-Enthüllung

Als erstes sollten die Identifikationsmerkmale entfernt werden (nominale Anonymsierung). Ferner sollte bei diskreten und insbesondere bei dichotomen Merkmalen mindestens 3 Fälle pro Ausprägung vorhanden sein. Drittens sollte die Klassifizierung von stetigen Variablen, die durch Indikatorvariable (Dummy-Variable) erreicht wird, untersagt werden. Viertens sollte die Erzeugung von Hebelwirkungen (leverages) nicht erlaubt sein, indem alle Transformationen von der Art

$$z \quad = \quad g\left(x - h(x_m)\right)$$

verboten sind, außer im Fall $h(x) = 0$, wie beispielsweise bei der Logarithmierung einer Variablen. Siehe dazu auch den Abschnitt 3.3.

2.5 Enthüllungs-Risiko und Daten-Nutzen für Modell-Server

Die Definition des Enthüllungs-Risikos wie auch des Daten-Nutzens soll dazu dienen zu entscheiden, welche Anfragen aus $\mathcal{Q}$ schließlich beantwortet werden.

Bei dem **Enthüllungsrisiko** geht es darum, dass ein Angreifer sensible Information bezüglich bestimmter Merkmale oder bezüglich der Beziehung zu einem sensiblen Merkmal abschätzen oder sogar genau bestimmen kann. Dabei unterscheiden Gomatam et al. (2005 S. 169) zwischen zwei Risiken:

- Abschätzung für Daten innerhalb von $\mathcal{D}$ (in-sample prediction risk)
- Abschätzung für Daten außerhalb von $\mathcal{D}$ (out-of-sample prediction risk)

Für das zweite Risiko sei als Beispiel die Schätzung einer Beziehung auf der Basis von $\mathcal{D}$ genannt, die dazu benutzt wird, auch Werte außerhalb der Datenmenge $\mathcal{D}$ abzuschätzen oder zu bestimmen.

Als Datennutzen könnte man die Information aus der Aktionsmenge $\mathcal{A}$ ansehen, wobei dieser Nutzen in Beziehung gesetzt werden muß zum Nutzen der Anfrage-Menge $\mathcal{Q}$. Man beachte, dass damit implizit ein **subjektiver** Nutzen definiert wird, denn für jede Person kann die Anfragemenge anders sein.

2.6 Risiko und Nutzen im Rahmen einer linearen Regressionsanalyse

Im folgenden wird unterstellt, dass $\mathcal{Q}$ nur Anfragen für lineare Regressionsanalysen auf Basis der Datenmenge $\mathcal{D}$ enthält und dass diese Datenmenge eine einzelne sensible Variable (beispielsweise Forschungsausgaben für Hochtechnologie) enthält. Ziel des Servers ist es, allzu akkurate Vorhersagen über dieses sensible Merkmal zu verhindern.

Nach wie vor soll die Datenmenge durch die Matrix $\mathbf{X}$ gegeben sein, allerdings soll es jetzt $K+1$ Merkmale geben, und mit x_0 soll das sensible Merkmal bezeichnet werden. Außerdem soll ein **statischer** Modell-Server betrachtet werden.

Angenommen es gäbe $K = 2$ "andere" Merkmale. Dann lassen sich insgesamt (neben der trivialen Regression auf das Absolutglied allein) zwei Einfach-Regression und eine multiple Regression mit beiden Merkmalen als Regressoren konstruieren, insgesamt also $2^K = 4$ Anfragen. Im Fall von $K = 3$ Merkmalen gibt es (neben der trivialen Regression) drei Einfach-Regressionen, 3 verschiedene Regressionen mit zwei Regressoren und eine Regression mit allen drei Merkmalen als Einflußgrößen, insgesamt also $2^K = 8$ Anfragen. Da jede Anfrage entweder beantwortet wird oder verweigert wird, ergeben sich bei K Merkmalen

$$\tau \quad = \quad 2^{2^K} \text{ Elemente in } \mathcal{A}$$

Für $K = 2$ gilt $\tau = 2^4 = 16$ und für $K = 3$ erhält man $\tau = 2^8 = 256$.

Um die Aktionsmenge nicht zu groß werden zu lassen, kann man daran denken, sich auf eine Untermenge der Merkmale aus der Datenmenge $\mathcal{D}$ zu beschränken. Dazu unterteilen wir die Matrix $\mathbf{X}$ wie folgt:

$$\mathbf{X} \quad = \quad (\mathbf{X}_{\text{free}}, \mathbf{X}_{\text{supp}}) \qquad ,$$

wobei 'supp' für suppressed , also unterdrückt, steht. Daraus ergibt sich - bei zusätzlicher Berücksichtigung der sensiblen Variuablen x_0 - für die Matrix der Kreuzprodukte

$$\mathbf{S}_{\mathbf{x}_0 \cup \mathbf{X}} \quad = \quad \left(\begin{array}{c|cc} s_{00} & \mathbf{s}'_{\text{free}} & \mathbf{s}'_{\text{supp}} \\ \hline \mathbf{s}_{\text{free}} & \multicolumn{2}{c}{\multirow{2}{*}{$\mathbf{S}_{\mathbf{X}}$}} \\ \mathbf{s}_{\text{supp}} & & \end{array} \right)$$

Dabei enthält der Vektor $\mathbf{s}_{\text{free}}$ die Kreuzprodukte zwischen x_0 und den Merkmalen, die nicht unterdrückt wurden. Entsprechendes gilt für $\mathbf{s}_{\text{supp}}$. Wenn alle Elemente in dieser Matrix mit Ausnahme des Vektors $\mathbf{s}_{\text{supp}}$ gegeben sind, dann kann man zeigen (Gomatam et al. 2005 S. 169 und Appendix), dass dieser Vektor in einem Ellipsoid liegen muß, der durch bekannte Größen gegeben ist. Allerdings wird dabei unterstellt, dass $\mathbf{S}_{\mathbf{X}}$ bekannt ist, was aber bei Unterdrückung der Merkmale in $\mathbf{X}_{\text{supp}}$ etwas unplausibel ist.

Als **Risiko** wird die Möglichkeit angesehen, die sensible Variable x_0 aus der "maximalen" Regression, d.h. bei Verwendung **aller** Variablen aus $\mathbf{X}$ zu bestimmen. Insbesondere geht es darum, die Einheiten mit atypischen Werten für x_0, zu schätzen, etwa die Unternehmen mit den größten Ausgaben für Forschung. Je größer das Bestimmtheitsmaß für diese Regression ist, desto größer ist das Risiko. Falls bestimmte Merkmale unterdrückt werden, kann der Angreifer den Vektor $\mathbf{s}_{\text{supp}}$ simulieren[15] und sodann die "maximale" Regression

[15] Für Details siehe Appendix in Gomatam et al. 2005. Neben der Annahme einer Gleichverteilung können auch alternative Annahmen gemacht werden, die bestimmte Detailkenntnisse reflektieren (to reflect domain knowledge).

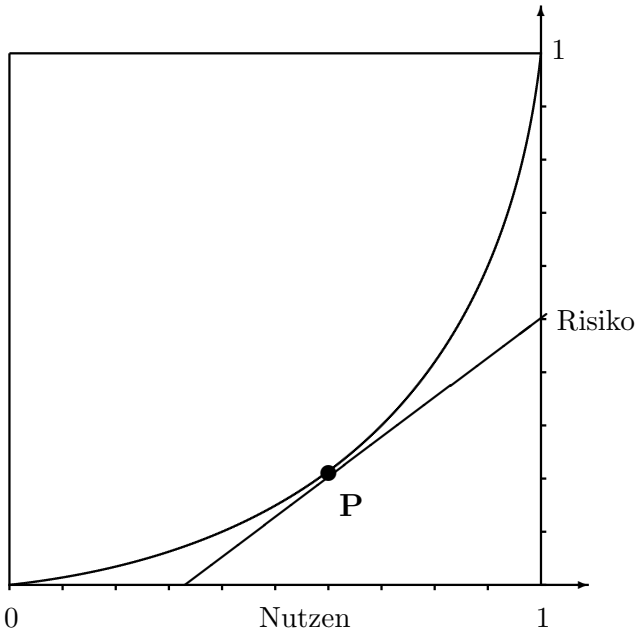


Abbildung 2/1: Risiko-Nutzen-Diagramm A

bestimmen. In diesem Fall wird das Risikomaß als Mittel der Bestimmtheitsmaße über alle Simulationen berechnet.

Als **Nutzenmaß** ließe sich das Bestimmtheitsmaß der **zulässigen maximalen** Regression ansehen, also der Regression von x_0 auf die verfügbaren Merkmale in $\mathbf{X}_{\text{free}}$. Damit ist natürlich das Nutzenmaß gleich dem obigen Risikomaß, wenn kein Merkmal unterdrückt wird!! Gomatam et al. (2005 S. 171) schlagen außerdem ein alternatives Maß vor, bei dem die "Wichtigkeit" der einzelnen Variablen durch eine Gewichtung berücksichtigt wird.

2.7 Die Simulationsstudie von Gomatam et al.

Die vorgeschlagenen Maße werden in einer **Simulationsstudie** illustriert. Der Datensatz enthält 9 Regresorvariablen sowie die sensible Variable x_0, die Anzahl Beobachtungen beträgt $n = 200$. Dabei werden in einem ersten Datensatz drei Regressoren hoch miteinander sowie mit x_0 korreliert, in einem zweiten Datensatz sind alle Merkmale nur mäßig miteinander korreliert. Unklar ist, in welcher Form die Information, dass die obersten 5 Prozent der sensiblen Variablen x_0 als Zielmenge (target set) betrachtet werden, die es besonders zu schützen gilt, in die Analyse eingeht.

Die Ergebnisse lassen sich – schematisch und nur teilweise !! – durch die beiden Abbildungen 2/1 und 2/2 charakterisieren. In beiden nimmt der Nutzen zu, wenn das Risiko zunimmt. Dabei ergibt sich eine konvexe Gestalt des Zusammenhangs, die nicht näher kommentiert wird. Die Kurve (oder besser Kurvenschar - siehe den Originalartikel!) repräsentiert für jeden Punkt eine bestimmte Regression.[16] Je nachdem, ob das Dominanzprinzip oder das Maximin-Prinzip verwendet wird, kommt Abbildung 2/1 oder Abbildung

[16]Dies wird bei Gomatam et al. (2005) durch farbige Graphiken deutlich gemacht. Siehe die Abbildungen 1 bis 4 dort.

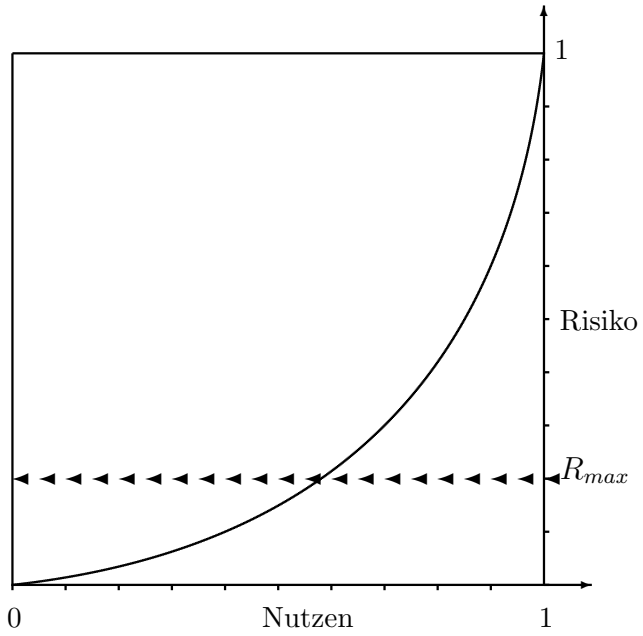


Abbildung 2/2: Risiko-Nutzen-Diagramm B

2/2 zur Anwendung. Im ersten Fall wird ein Punkt gewählt, der möglichst großen Nutzen und möglichst kleines Risiko besitzt. Wie die Steigung der "Trade-off Funktion" bzw. Indifferenzgeraden dabei bestimmt wird, wird nicht weiter erläutert.[17] Im zweiten Fall wird für ein gegebenes maximales Risiko R_{max} der Nutzen der "optimalen" Strategie durch die Punkte im Schnittpunkt bestimmt.

Erstaunlich an den Ergebnissen ist, dass von den drei hoch miteinander korrelierten Regressor-Variablen nur eine unterdrückt wird. Plausibel wäre die Unterdrückung von zwei der drei Variablen. Auch dies wird nicht weiter kommentiert. Außerdem wird berichtet, dass im Fall der hohen Korrelation (Datensatz I) das Risiko dann hoch ist, wenn eine der drei miteinander (und mit der abhängigen Variablen!!) korrelierten Regressoren involviert ist. Im Datensatz II, in dem kein Regressor große Vorhersagekraft hat, ergibt sich kein so klares Bild, es wirkt ein "Dimensionseffekt", soll heißen, sowohl Risiko als auch Nutzen steigen mit größerer Anzahl von Regressoren.

2.8 Eine (subjektive) Bewertung von Gomatam et al.

- Die Arbeit von Gomatam et al.(2005) geht davon aus, dass die Ergebnisse, die auf Originaldaten basieren, dem Datennutzer zur Verfügung gestellt werden, sofern das Enthüllungsrisiko dies zuläßt. Im Gegensatz dazu schlägt beispielsweise Heitzig (2005) vor, die Daten leicht zu verfremden (er nennt dies Jackknife-Methode) und die aus den verschiedenen Läufen resultierenden Ergebnisse als Intervall oder auch als Menge der Ergebnisse in diesem Intervall zur Verfügung zu stellen. Meines Erachtens wird es eine wesentliche Entscheidung im Projekt infinitE sein, welchen Weg man gehen soll.

[17] Siehe die allgemeinen Bemerkungen dazu bei Gomatam et al. (2005, S. 171/2).

- Der Hinweis auf die Problematik bei sequentiellen Anfragen ist wichtig. Die Anforderungen an den Server erhöhen sich dadurch dramatisch.
- Die Idee, Nutzen und Risiko gemeinsam zu betrachten, klingt gut, ist aber m.E. im Artikel noch wenig überzeugend ausgeführt.

3 Enthüllungsrisiken in statistischen Modellen bei Remote Access

In diesem Abschnitt werden statistische Modelle daraufhin untersucht, ob und inwieweit Enthüllungsrisiken trotz Remote Access bestehen. Es wird der Versuch unternommen, einen möglichst umfassenden Überblick zu geben, ohne dass allerdings der Anspruch erhoben wird, dass dieser Überblick - sowohl bezüglich der angesprochenen Themenbereiche als auch bezüglich der möglichen Risiko-Fälle in den angesprochenen Themenbereichen - erschöpfend ist. Ganz im Gegenteil ist zu erwarten, dass die Enthüllungs-Problematik bei Remote Access neugierige "Angreifer" immer wieder zu neuen Ideen anspornen wird. Andererseits hat das deutsche Statistik-Gesetz mit dem Begriff der "Faktischen Anonymität" die Vorstellung von einem wissenschaftsorientierten Datennutzer etabliert, der kein Interesse an gezielten Enthüllungen hat. Es ist meines Erachtens Utopie zu meinen, dass die Prüfung der Outputs im Rahmen des Remote Access eine hundertprozentige Sicherheit für Nicht-Enthüllung garantiert, zumal die Ressourcen für die Prüfung begrenzt sind bzw. bei adäquater finanzieller Entlohnung für die Wissenschaft zu teuer werden.

3.1 Enthüllungsrisiken bei Residual-Analyse

Da die Residuen e in einem beliebigen geschätzten Modell definitionsgemäß gleich der Differenz zwischen geschätzen Werten $\hat{y}$ und beobachteten Werten y der abhängigen Variablen sind,

$$e \quad = \quad \hat{y} - y \qquad ,$$

lassen sich die echten Werte der Variablen y dann ohne weiteres bestimmen, wenn neben den geschätzten Werten auch die Residuen zur Verfügung stehen. Auch im Rahmen von Modellen mit diskreten abhängigen Variablen, bei denen die abhängige Variablen nominal skaliert ist (Logit- und Probitanalyse), stellt die Reidualanalyse ein Enthüllungsrisiko dar: Da die Ausprägungen der abhängigen Variablen entweder den Wert 0 oder den Wert 1 aufweisen und die geschätzten Werte zwischen Null und Eins liegen, weist ein negativer (positiver) Wert des Residuums auf einen "Erfolg" ("Mißerfolg") für die entsprechende Beobachtung hin.[18]

Andererseits ist es oftmals für den Statistiker von Interesse, nicht nur die aus den Residuen bestimmten Werte von Spezifikationstests, etwa beim Test auf Kointegration, zu erhalten, sondern auch die Residuen selbst - graphisch - zu betrachten. Gomatam et al. (2005, S. 167) schlagen deshalb vor, dass anstelle der "echten" Residuen solche zur Verfügung gestellt werden, die aus Simulationen bestimmt sind. Mit Verweis auf Arbeiten von Reiter (2003) schlagen diese Autoren vor, Werte für die abhängige Variable y sowie für die Residuen e zu simulieren.

[18] Darauf weisen Reither und Kohnen (2005, S. 890) hin.

Eine Möglichkeit wäre, zunächst die (eindimensionale!!) Verteilung der echten Residuen durch eine Kerndichteschätzung zu approximieren, sodann aus dieser Verteilung die Residuen e_s zu simulieren und schließlich die Werte der abhängigen Variablen durch

$$y_s \quad = \quad f(\mathbf{x}, \hat{\boldsymbol{\beta}}) \; + \; e_s$$

darzustellen. Dabei bezeichnet $f(\mathbf{x}, \hat{\boldsymbol{\beta}})$ die geschätzte systematische Komponente des Modells. Allerdings ist dieser Weg nur für Querschnittsdaten gangbar; für Zeitreihendaten müssen die stochastischen Eigenschaften der Residuen erhalten bleiben. Denkbar wäre, in diesem Fall ein stochastisches Modell vom ARMA-Typ zu schätzen und die Residuen aus diesem geschätzten Modell zu bestimmen bzw. zu simulieren.

In Reiter und Kohnen (2005) wird die Residualanalyse in Logitmodellen auf Enthüllungsrisiken untersucht. Die Autoren schlagen eine Analyse anhand "gruppierter Residuen" (siehe Landwehr et al. (1984), Gelman et al. (2000) vor, um das Enthüllungsrisiko zu minimieren.

3.2 Enthüllung durch gezielt gesetzte Dummy-Variable

3.2.1 Formale Überlegungen

Alternativ kann man die Kenntnis über den Wert für eine bestimmte Einheit ausnutzen, um eine Dummy-Variable

$$\Im_{x=x_m} \quad = \begin{cases} 1 & \text{falls } x = x_m \\ 0 & \text{sonst} \end{cases}$$

zu konstruieren und diese als zusätzlichen Regressor in die Regressionsanalyse einzufügen, ggfs. auch in Interaktion mit anderen Regressoren. Dadurch wird der spezielle Einfluß der m-ten Einheit bestimmt.

Da der Wert im allgemeinen nur gerundet bekannt ist, ist die Konstruktion der Dummy-Variablen in der Form

$$\Im_{x \simeq x_m} \quad = \begin{cases} 1 & \text{falls } x - \delta < x_m < x + \delta \\ 0 & \text{sonst} \end{cases}$$

vermutlich wirkungsvoller.

Ein formaler Beweis, dass

$$\hat{y}_m \quad = \quad y_m$$

gilt, wird für den Fall, dass nur der Dummy (neben der Konstanten) als Regressor spezifiziert wird (Einfachregression), in Appendix A.3.1 präsentiert. In der Praxis wird ein Angreifer mit "Schurken-Mentalität"[19] vermutlich andere Regressoren als "Tarnung" zusätzlich verwenden. Auch in diesem Fall wird bei Verwendung des Dummys $\hat{y}_m \quad = \quad y_m$ gelten. Einen formalen Beweis liefert die Darstellung in Appendix A.3.2.

Dies führt zu der Frage, wie der Angreifer erkennt, ob er einen einzelnen Betrieb eindeutig identifiziert hat. Das Einfachste wäre, die Residuen auf Nullwerte zu überprüfen. Allerdings werden Residuen, wie oben bereits bemerkt, im allgemeinen nicht berichtet. Alternativ

[19] Siehe dazu die abschließenden Bemerkungen in Abschnitt 5.

könnte der Angreifer untersuchen, ob der Mittelwert der strategischen Dummy-Variablen gleich $1/n$ ist. Falls der Server jedoch auch Mittelwerte von binären Variablen, die ja stets besonders sensible Information enthalten, unterdrückt, könnte der Angreifer noch versuchen, die Varianz zu bestimmen, die im Fall einer einedtigen Identifikation durch $Var(\Im) = 1/n^2$ gegeben ist.

3.2.2 Ein empirisches Beispiel

Das IAB-Betriebspanel für das Jahr 2007 mit einem Beobachtungsumfang von 12.814 Betrieben wurde benutzt, um die Enthüllung bestimmter Werte für einzelne Unternehmen durch Setzung eines "strategischen" Dummys empirisch zu untersuchen. Dabei wurde die unterstellte Information bezüglich der Betriebsgröße (= Anzahl Beschäftigte) verwendet, um den Umsatz eines bestimmten Betriebs in Erfahrung zu bringen: Zunächst wurde aus dem Betriebspanel die Untermenge der Betriebe bestimmt, die eine Größe (= Beschäftigung) oberhalb des α-Quantil ausweisen, wobei α zwischen 90 % und 99,9 % variiert wurde. Aus dieser Teilmenge wurden dann zufällig Betriebe ausgewählt, deren Daten als Angriffswissen verwendet wurde.

Tabelle 3.1 gibt in Spalte 2 den jeweiligen Umfang N der Teilmenge an. Sodann wurden, soweit mehr als 100 Betriebe zur Verfügung standen, $n = 100$ Betriebe zufällig nacheinander ausgewählt. Im anderen Fall wurden *alle* Betriebe ausgewählt.[20] Siehe dazu Spalte 3 in der Tabelle. Für den jeweils ausgewählten Betrieb m wurde der Dummy

$$\Im_{x=x_m} = \begin{cases} 1 & \text{falls } x = BG_m \\ 0 & \text{sonst} \end{cases}$$

spezifiziert und eine Einfach-Regression mit konstantem Term sowie diesem Dummy auf der Basis von N Beobachtungen berechnet. Falls die Betriebsgröße des m-ten Betriebs nur einmal im Datensatz vorkommt, und zwar in allen 100 Fällen, dann würde sich wegen der theoretischen Ergebnisse in Abschnitt 3.2.1 eine mittlere Abweichung

$$\Delta = \frac{1}{n_p} \sum_{i=i_m} \frac{\hat{y}_i - y_i}{y_i} \tag{3-1}$$

von Null ergeben, wobei n_p die Anzahl gepickter Unternehmen (siehe Spalte 2 in der Tabelle) bezeichnet und die Summation über die jeweils gepickten Unternehmen läuft, was durch "$i = i_m$" angedeutet ist.

von Null ergeben. Offensichtlich ist aber in einigen Fällen, vor allem für großes N, der vermutlich oft gerundete Wert der Betriebsgröße mehrfach vorhanden. In diesem Fall "pickt" der Indikator mehrere Betriebe mit unterschiedlichen großen Umsätzen und es gilt dann $\hat{y}_m \neq y_m$, was zu einer positiven mittleren Abweichung führt.[21] Die Ergebnisse für dieses Szenario sind in der Zeile "exakte BG" angegeben.

[20] Die weiter unten berichteten Regressionsergebnisse basieren dagegen auf Analysen, die den gesamten Datensatz verwenden.

[21] Siehe dazu den Appendix A.3.3, in dem gezeigt wird, dass in diesem Fall der geschätzte Wert der abhängigen Variablen gleich dem *arithmetischen Mittel* der y-Werte der gepickten Unternehmen ist, d.h. insbesondere gilt

$$\hat{y}_m = \frac{1}{n_{x=x_m}} \sum_{x=x_m} y_i$$

.

Tabelle 3.1: Enthüllungsrisiko einer strategischen Dummyvariable

Quantil (%)	N	gezogene Betriebe	Indikatoren	Re-identif.-risiko (%)	mittl. rel. Abweichung (Δ) **Einf.-R.**	**mult. R.**
90,0	1.282	100	exakte BG	32,0	25,0	
			approx. BG	3,0	72,0	64,99
			+ Bundesland	20,0	30,7	
			+ Rechtsform	39,0	20,7	
			+ Branche	90,0	0,8	
95,0	643	100	exakte BG	57,0	15,0	
			approx. BG	03,0	58,6	38,83
			+ Bundesland	29,0	28,7	
			+ Rechtsform	42,0	24,1	
			+ Branche	83,0	00,7	
99,0	129	100	exakte BG	97,0	-0,8	
			approx. BG	08,0	26,7	51,22
			+ Bundesland	69,0	-0,9	
			+ Rechtsform	79,0	0,6	
			+ Branche	94,0	0,8	
99,5	65	65	exakte BG	100,0	0,0*	
			approx. BG	16,9	16,1	95,87
			+ Bundesland	86,2	3,9	
			+ Rechtsform	92,3	1,6	
			+ Branche	96,9	0,9	
99,9	13	13	exakte BG	100,0	0,0*	
			approx. BG	30,8	-0,5	-1,25
			+ Bundesland	100,0	0,0*	
			+ Rechtsform	100,0	0,0*	
			+ Branche	100,0	0,0*	

Datenbasis: IAB Betriebspanel 2007, IAB Nürnberg
Erläuterungen:
BG = Betriebsgröße, Einf.R. = Einfachregression, mult.R = multiple Regression
approx. BG = Betriebsgröße im Intervall wahre BG ± 2,5 %
mittl. rel. Abweichung = Differenz zwischen geschätztem und wahren Umsatz, gemittelt über alle gezogenen Betriebe
0,0* = Relative Abweichung ist kleiner als 10^{-12}

Wie aus der mit "Re-identifikations-Risiko" überschriebenen Spalte hervorgeht, ist bei Wahl des 90%-Quantils in 32 % der Fälle durch den Dummy der Betrieb m eindeutig identifiziert worden. In 68 % der Fälle dagegen existieren mehrere Betriebe mit identischer Betriebsgröße! Wird die Teilmenge durch Wahl eines größeren α reduziert, so steigt des (Re-)Identifikations-Risiko über 57 % und 97 % auf schließlich 100 % (für $\alpha = 0,995$ und $\alpha = 0,999$). In diesen Fällen ist dann, wie bereits oben beschrieben, die Abweichung in allen n Fällen gleich Null und damit auch die resultierende mittlere Abweichung.

Um auch den Fall zu berücksichtigen, dass die Betriebsgröße nur approximativ bekannt ist, wurde außerdem der Dummy

$$\Im_{x \approx x_m} = \begin{cases} 1 & \text{falls } 0,975 \cdot BG_m \leq x \leq 1.025 \cdot BG_m \\ 0 & \text{sonst} \end{cases}$$

gebildet. Die Ergebnisse für die entsprechende Analyse sind in der Tabelle in der Zeile "approx. BG" dargestellt. Natürlich sinkt durch die "unschärfere" Identifikation der Anteil

der eindeutig identifizierten Betriebe. Selbst für das "oberste" Quantil, in dem nur noch $N = 13$ Unternehmen vorhanden sind, liegt die Häufigkeit dieser Identifikation bei rund 31 %! Entsprechend größer ist nun auch die mittlere relative Abweichung: Erst im obersten Quantil sinkt dieser Wert(absolut gesehen) auf unter 1 %.

Wie wichtig in der Analyse des Enthüllungsrisikos die *zusätzliche* Berücksichtigung der Information bezüglich Region, Rechtsform und Branche ist, zeigen die in der Tabelle 3.1 gezeigten Ergebnisse eindrücklich: Bereits die zusätzliche Information bezüglich des relevanten Bundeslandes bewirkt in allen Fällen eine deutliche Steigerung des (Re-)Identifikations-Risikos sowie Reduzierung der mittleren relativen Abweichung! Siehe die Ergebnisse in der Zeile "+ Bundesland". Diese Tendenz verstärkt sich, wenn außerdem *zusätzlich* noch die Rechtsform sowie die Branche bekannt sind und in der Regressionsanalyse berücksichtigt werden. Siehe dazu die beiden Zeilen "+ Rechtsform" und "+ Branche".[22]

Es wurde auch der realistischere Fall einer Regressionsanalyse, in der auch weitere Einflußgrößen berücksichtigt sind, untersucht (multiple Regression).Wie die formale Analyse in Appendix A.3.2 zeigt, wird auch in diesem Fall $\hat{y}_1 = y_1$ gelten, sofern die Dummyvariable ein einziges Unternehmen identifiziert. Wenn allerdings in einigen der n untersuchten Fälle *keine* eindeutige Identifikation von Betrieb m erfolgt, ist wie im Fall der Einfachregression eine mittlere relative Abweichung größer Null zu erwarten. Siehe dazu Appendix A.3.3, der dies formal darstellt. Für das empirische Beispiel belegen dies die Ergebnisse in der letzten Spalte von Tabelle 3.1: Es wurde nur der Fall der "approximativen" Dummy Variablen untersucht. Da in allen Fällen das Reidentifikations-Risiko (siehe fünfte Spalte) kleiner als 100 % ist, ist klar, dass sich eine mittlere Abweichung größer Null ergibt (die hier nicht ausgewiesen ist!). Der Appendix A.3.2 macht klar, dass die Richtung wie auch Ausmaß der Abweichung von der jeweiligen Datenkonstellation abhängt, was durch den nicht-monotonen Verlauf der mittleren Abweichung illustriert wird: Die Abweichung sinkt zunächst von rund 65 % auf 39 %, steigt dann wieder auf 51 % und 96 % und fällt dann schließlich auf 1 %.

In einem späteren Stadium der Untersuchung wurde das Enthüllungsrisiko von strategischen Dummyvariablen nochmals untersucht. Die Ergebnisse sind in der Tabelle 3.2 zusammengefaßt.[23] Die Tabelle unterscheidet sich von Tabelle 3.1 in folgender Hinsicht: Es wurde sowohl der gesamte Datensatz im Umfang von 12.814 Betrieben als auch Teilmengen davon untersucht. Dabei wurde stets die **gesamte Teilmenge** statt wie zuvor nur eine Stichprobe von 100 Betrieben als "Zielbetrieb" ausgewählt. Dadurch sinkt natürlich das Reidentifikations-Risiko. Man vergleiche - für vergleichbare Quantile - Spalte 5 von Tabelle 3.1 mit Spalte 4 von Tabelle 3.2. Außerdem wurde statt des durchschnittlichen relativen Fehlers (siehe (3-1)) der durchschnittliche relative absolute Fehler berechnet, der durch

$$\Delta_{\text{abs}} = \frac{1}{N} \sum_{j=1}^{N} \frac{|\hat{y}_{m=j} - y_{m=j}|}{y_{m=j}} \tag{3-2}$$

gegeben ist. Für dieses Maß, das zudem nun über *alle* Betriebe der Teilmenge summiert, kompensieren sich natürlich positive und negative Abweichungen nicht, so dass in vielen Fällen, vor allem bei Betrachtung der Gesamtmenge und Teilmengen, die auch viele kleinere Betriebe enthalten, relativ große "mittlere" Abweichungen resultieren. Die letzte Spalte,

[22] Die gewählte Reihenfolge der Berücksichtigung spielt natürlich eine Rolle. Die Ergebnisse signalisieren jedoch, dass die Berücksichtigung des Bundeslandes die stärkste Steigerung des (Re-)Identifikations-Risikos bedeuten. Insofern ist die gewählte Reihenfolge besonders informativ für die Analyse.

[23] Diese Tabelle stammt aus Bleninger et al. (2010). Sie wurde ins Deutsche übersetzt.

die jeweils einen Fehler von Null angibt, reflektiert die theoretisch bereits abgeleitete Tatsache, dass bei eindeutiger Identifikation eines Betriebes der relative Fehler und damit auch der relative absolute Fehler gleich Null ist. Ansonsten entsprechen die Ergebnisse dieser Tabelle "tendentiell" denen in der oben präsentierten Tabelle 3.1.

Tabelle 3.2: Enthüllungsrisiko bei Verwendung von strategischen Dummyvariablen (Zusatzergebnisse)

quantile	N	Indicatoren $\Im_k$	Re-identifikations-Risiko	Δ_{abs}	Fehler bei eindeutiger Identifikation
all	12814	exakte BG	0.034	13801.825	0
		approx. BG	0.0009	11025.450	0
		+ Bundesland	0.023	11739.574	0
		+ Rechtsform	0.116	13633.345	0
		+ Branche	0.658	1.478	0
0.5	6516	exakte BG	0.066	26985.871	0
		approx. BG	0.002	21526.008	0
		+ Bundesland	0.046	21945.190	0
		+ Rechtsform	0.200	26774.728	0
		+ Branche	0.846	0.323	0
0.75	3217	exakte BG	0.134	52023.417	0
		approx. BG	0.003	40965.983	0
		+ Bundesland	0.085	39651.427	0
		+ Rechtsform	0.228	48390.483	0
		+ Branche	0.868	0.147	0
0.9	1282	exakte BG	0.335	1.956	0
		approx. BG	0.009	4.296	0
		+ Bundesland	0.186	1.944	0
		+ Rechtsform	0.352	1.499	0
		+ Branche	0.895	0.070	0
0.99	129	exakte BG	0.969	0.011	0
		approx. BG	0.085	1.311	0
		+ Bundesland	0.682	0.136	0
		+ Rechtsform	0.806	0.055	0
		+ Branche	0.953	0.021	0
0.999	13	exakte BG	1	0	0
		approx. BG	0.310	0.093	0
		+ Bundesland	1	0	0
		+ Rechtsform	1	0	0
		+ Branche	1	0	0

3.3 Enthüllungen via Transformation

3.3.1 Formale Überlegungen

Ein Datenangreifer interessiert sich für den Wert des Merkmal y für ein bestimmtes Unternehmen m. Falls er sicher ist, dass dieses Unternehmen im Datensatz enthalten ist und er Kenntnis über den Wert x_m eines anderen Merkmals x besitzt, kann er diese Kenntnis ausnutzen, um den Wert y_m abzuschätzen, indem er eine Regression mit y als abhängiger Variablen und

$$z = \frac{1}{|x - x_m| + \varepsilon} \tag{3-3}$$

als Regressorvariablen bestimmt, wobei ε ein beliebig kleiner (positiver) Wert ist. Für $x = x_m$ wird der Regressorwert z extrem groß und entfaltet damit eine Hebelwirkung (leverage), die dazu führt, dass die Regressionsgerade ein kleines Residuum bzw. einen guten Fit in

$$z(x_m) = \frac{1}{\varepsilon}$$

aufweist. Damit läßt sich der Wert von y für die m-te Einheit gut durch die Regressionsgerade abschätzen. Die Methode wird allerdings dann weniger guten Erfolg bringen, wenn es viele Werte von x in der Nähe von x_m gibt.[24]

Ein formaler Beweis, dass

$$\lim_{z \to \infty} \hat{y}_m \quad = \quad y_m$$

gilt, wird für den Fall, dass nur die Variable z als Regressor spezifiziert wird (Einfachregression), in Appendix A.3 präsentiert. In der Praxis wird man beispielsweise mit einem $\varepsilon \approx 10^{-4}$ arbeiten.

Ähnlich wie beim Einsatz einer strategischen Dummyvariable (siehe Abschnitt 3.2.1) stellt sich auch hier die Frage, wie der Angreifer erkennen kann, dass er einen einzigen Betrieb identifiziert hat. Allerdings ist im Fall des künstlichen Ausreißers - im Gegensatz zum Einsatz einer strategischen Dummyvariable - die Abfrage von Mittelwert oder Varianz der Regressorvariablen z nicht sinnvoll.

3.3.2 Ein empirisches Beispiel

Entsprechend dem Vorgehen im Fall der strategischen Dummyvariablen wurde auch für den künstlichen Ausreißer das IAB-Betriebspanel für eine empirische Illustration herangezogen. Allerdings entfällt in diesem Fall natürlich die Berücksichtigung von Zusatzinformation bezüglich Bundesland, Rechtsform und Branche. Tabelle 3.3[25] faßt die Ergebnisse zusammen. Die ersten drei Spalten stimmen mit den Angaben in Tabelle 3.2 überein. Und genau wie dort wird auch hier wird nur für die "oberste" Teilmenge der 13 größten Unternehmen für alle Betriebe eine eindeutige Identifikation erreicht, d.h. das Identifikationsrisiko liegt bei 100 %.

Es wurde auch untersucht, ob die Kenntnis bezüglich eines Wertes für eine weitere Variable und die Konstruktion eines künstlichen Ausreißers für diese Variable ein höheres Reiden-

[24] Gomatam et al. (2005 S. 167) unterstellen, dass x_m ein "ungewöhnlicher (unusual)" Wert ist.

[25] Diese Tabelle stammt aus Bleninger et al. (2010). Sie wurde ins Deutsche übersetzt.

Tabelle 3.3: Enthüllungsrisiko bei Verwendung eines künstlichen Ausreißers

Quantil	N	Re-identifikations-Risiko	Δ_{abs}	Fehler bei eindeutiger Identifikation
all	12814	0.034	13773.795	0.001
0.5	6516	0.066	26936.156	0.001
0.75	3217	0.134	51952.053	0.001
0.9	1282	0.335	1.957	0.001
0.99	129	0.969	0.011	0.0001
0.999	13	1	$1.195 * 10^{-6}$	$1.195 * 10^{-6}$

tifikationsrisiko bedeutet. Dafür wurde die Anzahl Auszubildender als Merkmal benutzt. Allerdings stellte sich heraus, dass dadurch die Prognosegütre eher abnimmt. Allerdings müßte sich, falls wirklich beide Außreißer gegen Unendlich streben, auch hier eine perfekte Prognose ergeben. Man vergleiche dazu die Ausführungen in Appendix A.2.2 und insbesondere die Formel (A-9). Warum in den empirischen Ergebnissen dieses Resultat nicht repöroduziert werden kann, soll noch näher untersucht werden.

3.4 Enthüllungsrisiken für Tabellen bei gesättigten Modellen

Bekanntlich kann man in einem (einfachen) Regressionsmodell die n Datenpunkte $(x_i, y_i), i = 1, \ldots, n$ exakt durch eine Kurve anpassen, wenn man ein Polynom (in x) vom Grade $n-1$ wählt. So lassen sich $n = 2$ Punkte exakt durch eine Gerade (Polynomgrad 1) und $n = 3$ Punkte exakt durch eine quadratische Funktion in x (Polynomgrad 2) anpassen. Alle Residuen sind dann Null und die geschätzten sind gleich den beobachteten Werte der abhängigen Variablen. Durch Vertauschen von x und y erhält man außerdem diese Information auch für die Regressorvariable. Entscheidend ist, dass die Anzahl der zu schätzenden Parameter gleich der Anzahl der Datenpunkte ist. Natürlich wird es im Zusammenhang mit Remote Access auffallen, wenn jemand bei einem Datensatz mit 580 Beobachtungen eine Regression anfordert, bei der die Regressorvariable bis zur 579. Potenz spezifiziert wird (sofern das die Programmpakete überhaupt schlucken). Deshalb wird dieses Szenarion auch nicht ausführlich als Enthüllungsszenario in einem separaten Abschnitt behandelt.

Weniger offensichtlich (und damit vom Server nicht unbedingt aufdeckbar) ist die Tatsache, dass bei nominal-skalierten Regressoren, insbesondere also bei binären Regressoren (Dummyvariablen) im Fall eines "gesättigten" Modells (englisch: saturated model) zwar nicht die Einzelwerte, sehr wohl dagegen die Werte bestimmter Tabellen rekonstruiert werden können.[26] Hier spielt die Tatsache eine Rolle, dass die Anzahl der zu schätzenden Pararameter identisch mit der Anzahl Zellen in der Tabelle ist! Dies soll in diesem Unterabschnitt ausführlicher behandelt werden.

Ein gesättigtes Modell liegt vor, wenn bei p nominal-skalierten Regressoren alle Interaktionen bis zum Grade $p-1$ berücksichtigt werden. Man beachte, dass im Fall einer stetigen abhängigen Variablen dieses Modell der Varianzanalyse zuzuordnen ist, während im Fall einer binären oder ganzzahlig positiven abhängigen Variablen diese Modelle in

[26] Siehe Reznek und Riggs (2005) sowie Gomatam et al. (2005 p. 167).

die Klasse der Logit/Probit-Modelle sowie der Poisson-Regression fallen, die unter dem Begriff "mikroökonometrische Modelle'[27] oder auch auch unter dem Begriff der loglinearen (Wahrscheinlichkeits-)Modelle bekannt sind.[28] Alle drei Fälle sollen in den folgenden Unterabschnitten behandelt werden.

3.4.1 Reidentifikation einer Werte-Tabelle mittels Varianzanalyse

Wir wollen zeigen, dass eine Varianzanalyse, also eine Regressionsanalyse mit einer stetigen abhängigen und ausschließlich diskreten Regressorvariablen (predictor variables), die "Tabellenwerte" der abhängigen Variablen[29] offenlegt, wenn die Varianzanalyse mit allen möglichen Interaktionen (auch höherer Ordnung) gefahren wird, d.h. wenn ein "gesättigtes" (saturated) Modell verwendet wird. Der allgemeine Beweis ist aufwendig wegen der notwendigen Symbolik. Für den Fall, dass neben dem Merkmal y (z.B. Umsatz) für zwei binäre Merkmale (x_1 = Groß- und Einzelhandel, x_2 Ost- und Westdeutschland) kann das Ergebnis jedoch leicht illustriert werden.

Der Gesamtumsätze für $n = 10$ Unternehmen soll in diesem Beispiel durch Tabelle 3.4 darstellbar sein. Im folgenden wird unterstellt, dass zumindest einzelne Zellen dieser Tabelle vertraulich und damit gesperrt sind. Es wird untersucht, ob diese Werte aus einer statistischen Prozedur im linearen Modell nachvollzogen werden können, wobei unterstellt wird, dass die beiden binären Variablen als Regressoren eingesetzt werden können, während die Einzelangaben für ein spezielles Unternehmen unbekannt sind.

Tabelle 3.4: Tabelle der Umsätze nach Groß/Einzelhandel und Region West/Ost

	G	E	$\sum$
W	3	7	10
O	36	19	55
$\sum$	39	26	65

Die zehn Beobachtungen von y, die dem Nutzer unbekannt sind, sind durch den n-dimensionalen Vektor $\mathbf{y}$ der abhängigen Variablen gegeben:

$$\mathbf{y}' = \begin{pmatrix} 1 & 2 & 3 & 4 & 17 & 5 & 8 & 7 & 12 & 6 \end{pmatrix}$$

Ferner sei $\mathbf{X}$ die $(n \times K)$-Matrix der Beobachtungen für die Regressoren D_H und D_R, die ebenfalls dem Nutzer unbekannt ist, wie folgt gegeben sein:

[27] Siehe dazu beispielsweise Ronning (1991).
[28] Siehe dazu beispielsweise Kapitel 10 in Fahrmeir et al. (1996).
[29] Siehe Abschnitt 2.2.2.

$$\mathbf{X} = \left(\begin{array}{ccccc} \underline{const.} & \underline{G} & \underline{E} & \underline{W} & \underline{O} \\ 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 1 & 0 \\ 1 & 1 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 & 1 \\ 1 & 1 & 0 & 0 & 1 \end{array} \right)$$

Dabei bezieht sich die erste Spalte auf das Absolutglied im Modell, die zweite und dritte Spalte betreffen die beiden Ausprägungen G = Großhandel und E = Einzelhandel der Handels-Variable D_H und die vierte und fünfte Spalte betreffen die beiden Ausprägungen W = West und O = Ost der Regionalvariable D_R. (Unten werden wir die Matrix reduzieren, weil beispielsweise die Information, dass ein Unternehmen nicht im Westen ist, eindeutig angibt, dass es im Osten ist.)

Beispielsweise ist das erste Unternehmen im Großhandel tätig und im Westen beheimatet, das neunte Unternehmen im Einzelhandel und im Osten zu Haus. Aus den Spaltensummen ergibt sich, dass insgesamt $n_G = 6$ Unternehmen im Großhandel (und $n_E = 4$ im Einzelhandel) sowie $n_W = 4$ im Westen (und $n_O = 6$ im Osten) tätig sind.[30] Auch die übrigen Angaben in der Tabelle 3.4 können aus $\mathbf{y}$ und $\mathbf{X}$ verifiziert werden!!!

Insbesondere erhalten wir die Randsummen aus

$$\mathbf{X}'\mathbf{y} = \left(\begin{array}{c} 1+2+3+4+17+5+8+7+12+6 \\ 1+2+17+5++8+6 \\ 3+4+7+12 \\ 1+2+3+4 \\ 17+5+8+7+12+6 \end{array} \right) = \left(\begin{array}{c} 65 \\ 39 \\ 26 \\ 10 \\ 55 \end{array} \right) = \left(\begin{array}{c} n\,\overline{y} \\ n_G\,\overline{y}_G \\ n_E\,\overline{y}_E \\ n_W\,\overline{y}_W \\ n_O\,\overline{y}_O \end{array} \right)$$

Demnach betrug der Umsatz insgesamt 65, der Umsatz im Großhandel 39 und im Einzelhandel 26 sowie im Westen 10 und im Osten 55 . Dagegen müssen die Angaben in den einzelnen Zellen durch entsprechende Addition verifiziert werden. Da nur das erste und zweite Unternehmen im Großhandel **und** im Westen arbeiten, ergibt sich der betreffende Gesamtumsatz als 1 + 2 = 3 und entspricht damit dem Wert in der Tabelle oben für das Feld "G/W". Entsprechend können die Werte in den anderen Feldern verifiziert werden.

Bekanntlich hat die obige Matrix $\mathbf{X}$ nicht vollen Spaltenrang. Üblicherweise werden deshalb nur

$$D^H = \begin{cases} 1 & \text{falls Unternehmen im Großhandel} \\ 0 & \text{falls Unternehmen im Einzelhandel} \end{cases}$$

[30] Die Tabelle faßt dies zusammen.

Anzahl Unternehmen			
	G	E	$\sum$
W	2	2	4
O	4	2	6
$\sum$	6	4	10

sowie

$$D^R = \begin{cases} 1 & \text{falls Unternehmen im Westen} \\ 0 & \text{falls Unternehmen im Osten} \end{cases}$$

betrachtet und die Matrix $\mathbf{X}$ reduziert sich zu

$$\mathbf{X} = \left(\begin{array}{ccc} \textit{const.} & G & W \\ \hline 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 1 & 0 \end{array} \right) \quad .$$

Damit erhalten wir

$$\mathbf{X'y} = \begin{pmatrix} 1+2+3+4+17+5+8+7+12+6 \\ 1+2+17+5++8+6 \\ 1+2+3+4 \end{pmatrix} = \begin{pmatrix} 65 \\ 39 \\ 10 \end{pmatrix} = \begin{pmatrix} n\ \overline{y} \\ n_G\ \overline{y}_G \\ n_W\ \overline{y}_W \end{pmatrix}$$

mit den Randsummen für 'G' und 'W' sowie dem Gesamtumsatz.

Diese Regressormatrix $\mathbf{X}$ geht in die bekannte Formel für den Kleinst-Quadrate-Schätzer

$$\mathbf{b} \quad = \quad \left(\mathbf{X'X}\right)^{-1}\ \mathbf{X'y} \qquad .$$

ein, die in diesem Fall das Modell der **Varianzanalyse mit Zweifach-Klassifikation** schätzt:

$$y_{ijk} = \mu + \beta_1\, D^H_{ijk} + \beta_2\, D^R_{ijk}\ +\ \varepsilon_{ijk} \qquad .$$

Dabei bezeichnet $j \epsilon \{G, H\}$ die Handelskategorie und $k \epsilon \{W, O\}$ die Regionalkategorie, während i die jeweilige Beobachtung in (j, k) indiziert.

Eine "gesättigte" Variante dieses Modells ergibt sich durch Hinzufügung der Interaktionsvariablen

$$D^{HR} \equiv D^H \cdot D^R = \begin{cases} 1 & \text{falls Unternehmen im Großhandel } \textbf{und} \text{ im Westen} \\ 0 & \text{sonst} \end{cases} \quad ,$$

was zum Modell

$$y_{ijk} = \mu + \beta_1\, D^H_{ijk} + \beta_2\, D^R_{ijk}\ +\ \beta_3\, D^{HR}_{ijk}\ +\ \varepsilon_{ijk}$$

führt. Man beachte, dass die vier Paramater μ, β_1, β_2 und β_3 den jeweiligen Erwartungswert bestimmen. Beispielsweise gilt für den Erwartungswert im Fall, dass das Unternehmen im Großhandel und im Westen ist,

$$E[y|D^H = 1, D^R = 1] = \mu + \beta_1 + \beta_2 + \beta_3$$

während für den Fall, dass Unternehmen im Osten und im Großhandel ist,

$$E[y|D^H = 1, D^R = 0] = \mu + \beta_1$$

gilt.

In diesem Fall ergibt sich für die Regressormatrix

$$\mathbf{X} = \left(\begin{array}{cccc} const. & G & W & G \cap W \\ \hline 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \end{array} \right) \tag{3-4}$$

und damit erhalten wir

$$\mathbf{X}'\mathbf{y} = \begin{pmatrix} 65 \\ 39 \\ 10 \\ 3 \end{pmatrix} = \begin{pmatrix} n\,\overline{y} \\ n_G\,\overline{y}_G \\ n_W\,\overline{y}_W \\ n_{GW}\,\overline{y}_{GW} \end{pmatrix} \quad ,$$

d.h. jetzt wird zusätzlich der Umsatz für das 'G/W'-Feld angegeben, aus dem die übrigen drei Werte in den restlichen Feldern unter Zuhilfenahme der Randsummen bestimmt werden können. Somit ist bereits der Ausdruck $\mathbf{X}'\mathbf{y}$ geeignet, die Detailinformationen aus der Tabelle 3.4 aufzudecken, sofern jemand diesen Ausdruck beim Fernrechnen anfordert.

Übrigens kann man außerdem die in Fußnote 30 gegebenen Häufigkeiten aus der Matrix $\mathbf{X}'\mathbf{X}$ wie folgt bestimmen:

$$\mathbf{X}'\mathbf{X} = \begin{pmatrix} n & n_G & n_W & n_{G\cap W} \\ n_G & n_G & n_{G\cap W} & n_{G\cap W} \\ n_W & n_{G\cap W} & n_W & n_{G\cap W} \\ n_{G\cap W} & n_{G\cap W} & n_{G\cap W} & n_{G\cap W} \end{pmatrix} = \begin{pmatrix} 10 & 6 & 4 & 2 \\ 6 & . & . & . \\ 4 & . & . & . \\ 2 & . & . & . \end{pmatrix} \tag{3-5}$$

Man sieht, dass eigentlich nur die erste Zeile (oder Spalte) dieses Ausdrucks bestimmt werden muß.

Üblicherweise wird man jedoch ganz normal den obigen KQ-Schätzer für das gesättigte Modell bestimmen. In diesem Fall ergibt sich

$$\begin{pmatrix} \hat{\mu} \\ \hat{\beta}_1 \\ \hat{\beta}_2 \\ \hat{\beta}_3 \end{pmatrix} = \begin{pmatrix} 9.50000 \\ -0.50000 \\ -6.00000 \\ -1.50000 \end{pmatrix}$$

Damit ergibt sich für die Schätzung der oben genannten bedingten Erwartungswerte

$$E[y|\widehat{D^H = 1}, D^R = 1] = \hat{\mu} + \hat{\beta}_1 + \hat{\beta}_2 + \hat{\beta}_3 = 1,5$$

sowie

$$E[y|\widehat{D^H = 1}, D^R = 0] = \hat{\mu} + \hat{\beta}_1 = 9,0 \quad .$$

Für den Vektor dieser geschätzten bedingten Erwartungswerte, der üblichweise mit $\hat{y}$ bezeichnet wird, für alle 10 Unternehmen insgesamt ergibt sich

$$\hat{y} \quad = \quad \begin{pmatrix} 1.5 \\ 1.5 \\ 3.5 \\ 3.5 \\ 9.0 \\ 9.0 \\ 9.0 \\ 9.5 \\ 9.5 \\ 9.0 \end{pmatrix}$$

Da es insgesamt 2 Unternehmen in der Klasse "Großhandel und Region West" gibt, ergibt sich der geschätzte **Gesamt**umsatz für diese Gruppe als $2 \cdot 1,5 = 3,0$. Ebenso ergibt sich für die 2 Unternehmen der Gruppe "Einzelhandel und Region Westen" als **Gesamt**umsatz $2 \cdot 3,5 = 7,0$, für die Gruppe "Großhandel und Region Osten" als **Gesamt**umsatz $4 \cdot 9,0 = 36,0$ sowie für die Gruppe "Einzelhandel und Region Osten" als **Gesamt**umsatz $2 \cdot 9,5 = 19,0$. Damit entsprechen die **geschätzten** Gesamtumsätze exakt den beobachteten Umsätzen in Tabelle 3.4. Dies ist kein Zufall ! Zunächst erinnern wir uns daran, dass die Matrix $\mathbf{X}$ wie ein Summenoperator agiert, d.h. mit dem Ausdruck

$$\mathbf{X}'\hat{\mathbf{y}} = \begin{pmatrix} 65.00000 \\ 6.00000 \\ 39.00000 \\ 3.00000 \end{pmatrix}$$

erhalten wir wie oben **dieselben** vier Angaben, die zur vollständigen Rekonstruktion der Tabelle 3.4 notwendig sind. Diese Identität beruht auf der Tatsache, dass

$$\mathbf{X}'\hat{\mathbf{y}} = \mathbf{X}'\,\mathbf{X}\mathbf{b}$$

gilt, woraus

$$\mathbf{X}'\hat{\mathbf{y}} = \mathbf{X}'\,\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\,\mathbf{X}'\mathbf{y} = \mathbf{X}'\mathbf{y}$$

folgt. Demnach bedeutet die zusätzliche Ausgabe von $\mathbf{X}'\hat{\mathbf{y}}$ bzw. von $\mathbf{X}'\,\mathbf{X}\mathbf{b}$ ein Reidentifikationsrisiko, sofern Angaben aus der Tabelle 3.4 geschützt werden sollen.

3.4.2 Reidentifikation einer Häufigkeitstabelle mittels Logitanalyse

Es soll nun auch der Fall betrachtet werden, dass statt der stetigen Variablen "Umsatz" die binäre Variable "Betriebsrat" (d.h. die Tatsache dass ein Betriebsrat im Unternehmen vorhanden ist oder nicht) in einer Tabelle gegeben ist, bei der einzelne Zellen nicht veröffentlicht werden können bzw. sollen. Dabei knüpfen wir an das Beispiel aus dem Abschnitt 3.4.1 an. Zusätzlich soll jetzt der Vektor der Betriebsrats-Variablen für die 10 Unternehmen wie folgt gegeben sein (der annahmegemäß wieder dem Datennutzer nicht bekannt gegeben wird):

$$\mathbf{y}' = \begin{pmatrix} 0 & 1 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 1 \end{pmatrix}$$

Daraus ergibt sich entsprechend Tabelle 3.4 die Tabelle 3.5 für die Anzahl der Unternehmen mit Betriebsrat nach Handelstyp und Region, wobei zusätzlich die Anzahl Unternehmen angegeben ist, auf die sich diese Angaben beziehen.[31]

Tabelle 3.5: Unternehmen mit Betriebsrat nach Groß/Einzelhandel und Region West/Ost

	G	E	$\sum$
W	1 (2)	1 (2)	2 (4)
O	2 (4)	0 (2)	2 (6)
$\sum$	3 (6)	1 (4)	4 (10)
In Klammern Anzahl Unternehmen			

Sowohl Logit- als auch Probit-Modell schätzen das bedingte Wahrscheinlichkeitsmodell

$$P(Y_{ijk} = 1|x_1, x_2) = F(\mu + \beta_1 D^H_{ijk} + \beta_2 D^R_{ijk} + \beta_3 D^{HR}_{ijk}) \quad ,$$

d.h. es wird die bedingte Wahrscheinlichkeit (bzw. der bedingte Erwartungswert) für einen Betriebsrat im Unternehmen geschätzt, gegeben eine bestimmte Handelsform und eine bestimmte Region. In obiger Formulierung ist bereits wieder die Interaktion zwischen Handelsform und Region berücksichtigt. Dabei ist F entweder die logistische Verteilung (für das Logitmodell) oder die Normalverteilung (für das Probitmodell). Wie im linearen Modell werden die vier unbekannten Parameter μ, β_1, β_2 und β_3 geschätzt, dieses Mal allerdings üblicherweise mit der Maximum-Likelihood-Methode. Für die geschätzten Wahrscheinlichkeiten ergibt sich dann

$$\widehat{P(Y_{ijk} = 1|x_1, x_2)} = F(\hat{\mu} + \hat{\beta}_1 D^H_{ijk} + \hat{\beta}_2 D^R_{ijk} + \hat{\beta}_3 D^{HR}_{ijk}) \quad ,$$

Die Tatsache, dass auch in diesem Fall die geschätzten Wahrscheinlichkeiten mit den empirischen Häufigkeiten übereinstimmen, ist dadurch bedingt, dass auch in diesem Fall der Ausdruck

$$\mathbf{X}'\mathbf{y}$$

mit $\mathbf{X}$ aus (3-4) eine Rolle spielt. Dies soll im Folgenden - für das (binäre) Logitmodell - gezeigt werden.

Unter den üblichen Annahmen des Logitmodells ergibt sich bei Maximum-Likelihood-Schätzung der folgende Gradient (score function)[32]:

$$\frac{\partial L}{\partial \boldsymbol{\beta}} = -\sum_{t=1}^{n} \left[(1 - y_t) - \frac{1}{1 + \exp(\mathbf{x}'_t \boldsymbol{\beta})}\right] \mathbf{x}_t \tag{3-6}$$

mit

$$\begin{aligned} \mathbf{x}_t &= (1\,,\, x_{t2}\,,\, x_{t3}\,,\, \ldots\,,\, x_{tK})' \\ \boldsymbol{\beta} &= (\beta_1\,,\, \beta_2\,,\, \beta_3\,,\, \ldots\,,\, \beta_K)' \end{aligned}$$

[31] Üblicherweise würde deshalb wohl die Tabelle mit relativen Anteilen oder auch Prozenten dargestellt werden. Beispielsweise könnte die Tabelle so aussehen:

Anteil Unternehmen mit Betriebsrat			
	G	E	G+E
W	0,50	0,50	0,50
O	0,50	0	0,33
W+O	0.50	0,25	0,40

[32] Siehe beispielsweise Ronning (1991) Seite 37.

Unter Verwendung von

$$p\left(x_t;\boldsymbol{\beta}\right)=\frac{1}{1+\exp\left(-x_t'\,\beta\right)} \qquad . \tag{3-7}$$

sowie

$$\mathbf{X}=\begin{pmatrix}\mathbf{x}_1'\\ \mathbf{x}_2'\\ \dots\\ \mathbf{x}_n'\end{pmatrix}$$

wird daraus[33]

$$\frac{\partial\,L}{\partial\,\boldsymbol{\beta}}=\sum_{t=1}^{n}\left[y_t-p(\mathbf{x}_t;\boldsymbol{\beta})\right]\,\mathbf{x}_t$$

oder auch

$$\frac{\partial\,L}{\partial\,\boldsymbol{\beta}}=\mathbf{X}'\mathbf{y}-\mathbf{X}'\begin{pmatrix}p(\mathbf{x}_1;\boldsymbol{\beta})\\ p(\mathbf{x}_2;\boldsymbol{\beta})\\ \vdots\\ p(\mathbf{x}_n;\boldsymbol{\beta})\end{pmatrix}$$

Das Maximum wird erreicht, wenn die ersten partiellen Ableitungen gleich Null gesetzt werden. Demnach gilt für den ML-Schätzer $\hat{\boldsymbol{\beta}}$ (sowie die geschätzten Wahrscheinlichkeiten $\hat{p}$)

$$\mathbf{X}'\mathbf{y} \quad = \quad \mathbf{X}'\begin{pmatrix}\hat{p}(\mathbf{x}_1;\hat{\boldsymbol{\beta}})\\ \hat{p}(\mathbf{x}_2;\hat{\boldsymbol{\beta}})\\ \vdots\\ \hat{p}(\mathbf{x}_n;\hat{\boldsymbol{\beta}})\end{pmatrix} \tag{3-8}$$

oder auch - unter Beachtung von $\mathbf{X}$ aus (3-4) -

$$\begin{pmatrix}\sum_{i=1}^{n} y_i\\ \sum_{i\epsilon G} y_i\\ \sum_{i\epsilon W} y_i\\ \sum_{i\epsilon G\cap W} y_i\end{pmatrix} \quad = \quad \begin{pmatrix}\sum_{i=1}^{n} \hat{p}(\mathbf{x}_i;\hat{\boldsymbol{\beta}})\\ \sum_{i\epsilon G} \hat{p}(\mathbf{x}_i;\hat{\boldsymbol{\beta}})\\ \sum_{i\epsilon W} \hat{p}(\mathbf{x}_i;\hat{\boldsymbol{\beta}})\\ \sum_{i\epsilon G\cap W} \hat{p}(\mathbf{x}_i;\hat{\boldsymbol{\beta}})\end{pmatrix} \qquad .$$

Weil

$$\begin{pmatrix}\sum_{i=1}^{n} y_i\\ \sum_{i\epsilon G} y_i\\ \sum_{i\epsilon W} y_i\\ \sum_{i\epsilon G\cap W} y_i\end{pmatrix} \quad = \quad \begin{pmatrix}\text{Anzahl Betriebsräte insgesamt}\\ \text{Anzahl Betriebsräte im Großhandel}\\ \text{Anzahl Betriebsräte im Westen}\\ \text{Anzahl Betriebsräte im Großhandel und im Westen}\end{pmatrix} \quad = \quad \begin{pmatrix}4\\ 3\\ 2\\ 1\end{pmatrix}$$

gilt, erhält man die jeweilige Anzahl Betriebsräte (= Anzahl "Erfolge") durch entsprechende Aufsummation der geschätzten Wahrscheinlichkeiten bzw. durch den Ausdruck

$$\mathbf{X}'\begin{pmatrix}\hat{p}(\mathbf{x}_1;\hat{\boldsymbol{\beta}})\\ \hat{p}(\mathbf{x}_2;\hat{\boldsymbol{\beta}})\\ \vdots\\ \hat{p}(\mathbf{x}_n;\hat{\boldsymbol{\beta}})\end{pmatrix} \qquad .$$

Zusätzlich kann man die dazugehörige Anzahl "Versuche" aus der Regressormatrix gemäß (3-5) bestimmen. Es lassen sich also durch ein "gesättigtes" Logit-Modell **alle** Angaben aus der Tabelle 3.5 rekonstruieren!

[33] Man beachte, dass

$$\frac{1}{1+\exp\left(\mathbf{x}_t'\,\boldsymbol{\beta}\right)}=1-\frac{1}{1+\exp\left(-\,\mathbf{x}_t'\,\boldsymbol{\beta}\right)}=1-p(\mathbf{x}_t)$$

gilt.

3.4.3 Ein Poisson-Regressionsmodell mit binären Regessoren

Es soll nun auch der Fall betrachtet werden, dass statt der stetigen Variablen "Umsatz" die diskrete Variable "Beschäftigung" in einer Tabelle gegeben ist, bei der einzelne Zellen nicht veröffentlicht werden können. Dabei knüpfen wir an das Beispiel aus den vorhergehenden Abschnitten 3.4.1 und 3.4.2 an. Zusätzlich soll jetzt der Vektor der Beschäftigtenzahlen für die 10 Unternehmen wie folgt gegeben sein (der wiederum dem Datennutzer nicht bekannt gegeben wird):

$$\mathbf{y}' = \begin{pmatrix} 31 & 22 & 73 & 24 & 17 & 35 & 18 & 97 & 124 & 67 \end{pmatrix}$$

Daraus ergibt sich entsprechend Tabelle 3.4 die Häufigkeitstabelle 3.6 für die Beschäftigung nach Handelstyp und Region.

Tabelle 3.6: Tabelle der Beschäftigten nach Groß/Einzelhandel und Region West/Ost

	G	E	$\sum$
W	53	137	190
O	73	185	258
$\sum$	126	322	448

Wenn die ganzzahlige Variable Y (= Beschäftigung) analysiert werden soll, verwendet man üblicherweise das Poissonmodell oder auch dessen Erweiterung, das sogenannte NEGBIN-Modell, das die flexiblere negative Binomialverteilung unterstellt.[34] Im Poissonmodell ergibt sich der bedingte Erwartungswert als

$$\lambda(x_1, x_2) \equiv E(Y_{ijk}|x_{1ij}, x_{2ik}) = \exp(\mu + \beta_1 D^H_{ijk} + \beta_2 D^R_{ijk} + \beta_3 D^{HR}_{ijk}) \quad ,$$

wobei auch hier wieder die beiden Dummyvariablen für Handelsform und Region sowie deren Interaktion spezifiziert sind.

Auch dieses Modell wird üblicherweise mit der Maximum-Likelihood-Methode geschätzt. Für den Vektor der ersten partiellen Ableitungen der Log-Likelihoodfunktion ergibt sich[35]

$$\frac{\partial L}{\partial \boldsymbol{\beta}} = \mathbf{X}'\mathbf{y} - \mathbf{X}' \begin{pmatrix} \lambda(\mathbf{x}_1; \boldsymbol{\beta}) \\ \lambda(\mathbf{x}_2; \boldsymbol{\beta}) \\ \vdots \\ \lambda(\mathbf{x}_n; \boldsymbol{\beta}) \end{pmatrix}$$

und für die mit der ML-Methode geschätzten Koeffizienten gilt:

$$\mathbf{X}'\mathbf{y} \quad = \quad \mathbf{X}' \begin{pmatrix} \hat{\lambda}(\mathbf{x}_1; \hat{\boldsymbol{\beta}}) \\ \hat{\lambda}(\mathbf{x}_2; \hat{\boldsymbol{\beta}}) \\ \vdots \\ \hat{\lambda}(\mathbf{x}_n; \hat{\boldsymbol{\beta}}) \end{pmatrix} \quad .$$

[34] Siehe beispielsweise Ronning (1991) Abschnitt 4.2.

[35] Siehe dazu beliebiges Buch über Mikroökonometrie. In Ronning(1991) wird eine explizite Schreibweise gewählt.

Demnach erhält man die jeweilige Beschäftigung in der Tabelle 3.6 durch entsprechende Aufsummation der geschätzten bedingten Erwartungswerte bzw. durch den Ausdruck

$$\mathbf{X}' \begin{pmatrix} \hat{\lambda}(\mathbf{x}_1;\hat{\boldsymbol{\beta}}) \\ \hat{\lambda}(\mathbf{x}_2;\hat{\boldsymbol{\beta}}) \\ \vdots \\ \hat{\lambda}(\mathbf{x}_n;\hat{\boldsymbol{\beta}}) \end{pmatrix} \quad .$$

Es lassen sich also durch ein "gesättigtes" Poisson-Modell **alle** Angaben aus der Tabelle 3.6 rekonstruieren! Im Übrigen sei hier ohne Beweis oder formale Darstellung angemerkt, dass dies Ergebnis auch mit dem NEGBIN-Modell erreicht werden kann.

3.5 Inferenz-Enthüllung

Bei der Inferenz-Enthüllung[36] geht es vor allem um die Aufdeckung von bestimmten "engen" Zusammenhängen, die geheimgehalten werden sollen. Dies dürfte vor allem dann der Fall sein, wenn die abhängige Variable in dieser Beziehung eine sensible sprich streng vertrauliche Variable ist.[37] Man denke beispielsweise an den Forschungsaufwand von Unternehmen in bestimmten hochtechnologischen Bereichen.

Im folgenden unterstellen wir, dass in $\mathcal{D}$ die $(n \times K)$-Datenmatrix

$$\mathbf{X} \quad = \quad \left(\begin{array}{cccc} \mathbf{x}_1 & \mathbf{x}_2 & \ldots & \mathbf{x}_{K-1} \quad \mathbf{x}_K \end{array} \right)$$

mit n-dimensionalen Vektoren $\mathbf{x}_k$ gegeben ist. Für beliebige Regression

$$\mathbf{x_a} \quad = \quad \mathbf{X}_B\,\boldsymbol{\beta} \,+\, \mathbf{u}$$

ergibt sich der Kleinstquadrate-Schätzer

$$\mathbf{b}_{a|B} \quad = \quad \left(\mathbf{X}_B'\,\mathbf{X}_B\right)^{-1}\,\mathbf{X}_B'\,\mathbf{x}_a$$

sowie Kovarianzmatrix

$$\hat{\Sigma}_{a|B} \quad = \quad s^2\,\left(\mathbf{X}_B'\,\mathbf{X}_B\right)^{-1}$$

mit

$$s^2 \quad = \quad \hat{\mathbf{u}}'\hat{\mathbf{u}}/(n-K) = \left(\mathbf{x}_a'\mathbf{x}_a - \mathbf{b}_{a|B}'\mathbf{X}_B'\,\mathbf{x}_a\right)/(n-K)$$

und dem Bestimmtheitsmaß

$$R^2_{a|B} \quad = \quad 1 - \frac{\left(\mathbf{x}_a'\mathbf{x}_a - \mathbf{b}_{a|B}'\mathbf{X}_B'\,\mathbf{x}_a\right)}{\mathbf{x}_a'\mathbf{x}_a \,-\, n(\overline{x}_a)^2} \quad .$$

Diese Angaben sind, wie die Formeln zeigen, dann verfügbar, wenn die Matrix

$$\mathbf{X}'\mathbf{X} \quad = \quad \left(\begin{array}{ccc} \mathbf{x}_a & \mathbf{X}_B & \mathbf{X}_{\text{Rest}} \end{array} \right)' \left(\begin{array}{ccc} \mathbf{x}_a & \mathbf{X}_B & \mathbf{X}_{\text{Rest}} \end{array} \right)$$

der Kreuzprodukte für alle Variablen in $\mathcal{D}$ und damit insbesondere die Matrix

$$\mathbf{S}_{a|B} \quad = \quad \left(\begin{array}{cc} \mathbf{x}_a & \mathbf{X}_B \end{array} \right)' \left(\begin{array}{cc} \mathbf{x}_a & \mathbf{X}_B \end{array} \right)$$

[36] Dieser Unterabschnitt orientiert sich eng an Gomatam et al.(2005).

[37] Siehe auch auch Heitzig (2004, 2005).

gegeben ist.[38]

Falls nun die Beziehung zwischen einer bestimmten Variablen x_a und den Regressoren in X_B geheimgehalten werden soll, könnte man daran denken, nur Anfragen bezüglich einer solchen Regression nicht zu erfüllen. Allerdings ist dies nicht erfolgreich, denn es gilt

$$\mathbf{S}_{a|B}(i,j) \quad = \quad \mathbf{S}_{a|B}(j,j)\,\mathbf{b}_{i|j} \quad = \quad \left(\mathbf{X}_j'\,\mathbf{X}_j\right)\,\left(\mathbf{X}_j'\,\mathbf{X}_j\right)^{-1}\,\mathbf{X}_j'\,\mathbf{x}_i \quad = \quad \mathbf{X}_j'\,\mathbf{x}_i$$

oder in Worten: Beliebiges Element aus $\mathbf{S}_{a|B}$ kann durch Regression von $\mathbf{x}_i$ auf $\mathbf{X}_j$ bestimmt werden, wenn beide Variablen aus dem Datensatz $(\mathbf{x}_a, \mathbf{X}_B)$ stammen. Deshalb muß der Server prüfen, ob die angefragte Regression aus diesem Datensatz ist. Zumindest für eine Regression aus diesem Datensatz sollte die Erfüllung der Anfrage verweigert werden.[39]

4 Enthüllungsrisiken bei multivariaten Verfahren

Das Thema Reidentifikation bzw. Enthüllungsrisiko wird üblicherweise nur am Beispiel der (linearen) Regressionsanalyse abgehandelt. Aber zumindest "die" multivariaten Verfahren, die im Folgenden abgehandelt werden, sind inzwischen Standard-Instrumente und sollten deshalb auch auf mögliche Risiken untersucht werden. Dabei ist natürlich die zuerst angesprochene Clusteranalyse trivialerweise ein Enthüllungsrisiko, sofern das betreffende Programm die Darstellung von Clustern mit weniger als vier (drei, zwei) Elementen erlaubt. Hinzuweisen ist in diesem Zusammenhang auch auf Reidentifikations-Strategien, die auf Matching-Algorithmen basieren, bei denen ebenfalls die Ähnlichkeit von Objekten überprüft wird.[40]

4.1 Clusteranalyse

4.1.1 Allgemeines

Clusteranalyse[41] bewertet die Ähnlichkeit von Objekten anhand ihrer beobachtbaren Merkmalsausprägungen. Dabei sind sowohl stetige als auch diskrete Merkmale zugelassen, wenn auch die Clusteranalyse vor allem stetige Merkmale betrachtet. Vor allem die gemeinsame Behandlung von stetigen **und** diskreten Merkmalen ist problematisch. Im allgemeinen dient die Clusteranalyse der deskriptiven statistischen Analyse, ist also nicht durch ein strukturelles (ökonomisches) Modell motiviert.

Die einzelnen Clusterverfahren unterscheiden sich in mannigfaltiger Weise.[42] Wesentliche Komponenten der Clusteranalyse sind

[38] Annahmegemäß sind auch alle Mittelwerte gegeben. Die bei Gomatam et al. (2005) unterstellte Zentrierung der Variablen ist überflüssig!

[39] Gomatam et al. (2005 , S. 169) beschreiben auch noch eine Abschätzung für unbekannte Elemente der Matrix $\mathbf{S}_{a|B}$, wobei ausgenutzt wird, dass diese Matrix positiv definit ist. Darauf gehen wir nicht weiter ein.

[40] Siehe beispielsweise die Arbeiten von Rainer Lenz.

[41] Einen guten Überblick geben beispielsweise Kaufmann und Pape (1996) sowie Backhaus et al. (2008).

[42] Das Folgende orientiert sich weitgehend an dem Buch von Backhaus et al. (2008).

(a) die Bestimmung der Ähnlichkeit von verschiedenen Untersuchungseinheiten bzw. Distanz zwischen diesen Untersuchungseinheiten. Man unterscheidet deshalb auch zwischen **Ähnlichkeitsmaßen** und **Distanzmaßen**.

(b) die Methode, mit der die Gruppen gebildet (bzw. umgebildet) werden.

Wenn eine Distanz (oder Ähnlichkeits-)Matrix zur Verfügung steht, dann geht es darum zu entscheiden, welche Objekte gemeinsam in einem Cluster erscheinen sollen. Dies hängt bereits sehr stark von der Wahl des Distanz- oder Ähnlichkeitsmaßes abhängt. Aber auch bei gegebener Entscheidung ist

- die Zahl der Cluster wie auch
- die Art der **Messung des Abstandes zwischen Clustern bzw. Objekten**

von Bedeutung. Abbildung 4/1 gibt einen Überblick über verschiedene Verfahren.

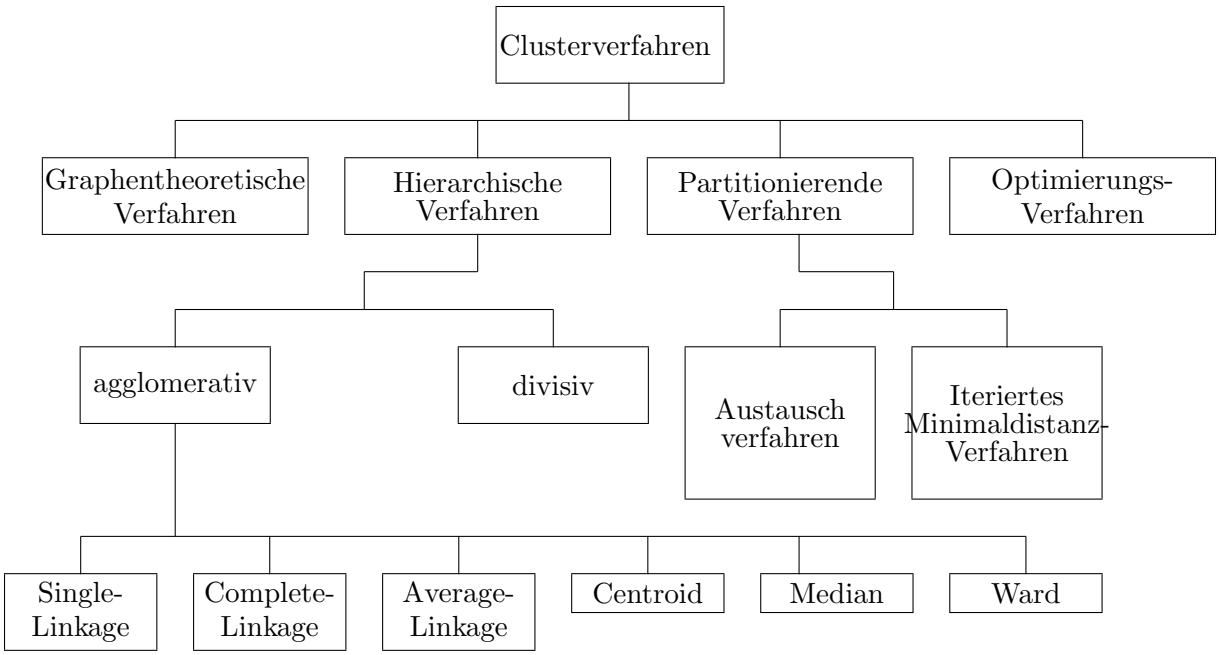


Abbildung 4/1: Ein Überblick über die Algorithmen (in Anlehnung an das Lehrbuch von Backhaus)

Am bedeutsamsten sind die **hierarchischen Verfahren**, die sukzessiv ein Cluster auf- oder abbauen. Im ersten Fall spricht man von **agglomerativen**, sonst von **divisiven** Verfahren. Von besonderer Bedeutung sind in der Unterklasse der agglomerativen hierarchischen Verfahren mit den Varianten

- das Single-Linkage-Verfahren (nearest neighbour)
- das Complete-Linkage-Verfahren (furthest neighbour)
- das Verfahren von Ward

Mit dem jeweiligen Distanzmaß wird in einem ersten Schritt der aus den am nächsten zueinander liegenden Objekten ein Cluster (Klumpen) gebildet und sodann erneut die Distanz zwischen diesem Cluster und allen anderen Objekten bestimmt. In weiteren Schritten werden aus den zweit-(dritt-, viert, usw.) nächsten Objekten bzw. Clustern neue Cluster gebildet, bis alle Objekte in einem einzigen Cluster vereint sind (= Abbruchkriterium). Interessant sind natürlich die zuvor gebildeten Cluster.

Abbildung 4/2: Der Befehl " cluster dendrogram, labels(besch) "

Abbildung 4/3: Der Befehl "cluster dendrogram, showcount cutvalue(1000)"

4.1.2 Enthüllungsrisiken in der Clusteranalyse

Das Enthüllungsrisiko bei der Clusteranalyse ist vielfältig:

- Es können die einzelnen Objekte eines Datensatzes dargestellt werden. Allerdings lassen sich die jeweiligen Merkmalsausprägungen nur dann abrufen, wenn nicht die Distanzmatrix als Ausgangsinformation verwendet wird.
- Die **graphische Darstellung** der Positionierung der einzelnen Objekte im zweidimensionalen Raum - möglichst unter Verwendung von zwei ' 'Schlüsselmerkmalen" - gibt für jedes Objekt die Werte zumindest approximativ an.
- Das Verfahren kann - ggfs. durch Variation des Distanzmaßes - Extremwerte lokalisieren, was auf eine Reidentifikation großer Unternehmen hinausläuft. Siehe auch das Beispiel in Abschnitt 4.1.3.

Beispielsweise bewirken die Stata-Befehle

```
cluster wardslinkage umsatz besch , name(bsp1) cluster dendrogram,
labels(umsatz) cluster dendrogram, labels(besch) cluster dendrogram,
showcount cutvalue(1000)
```

folgendes: Zunächst wird eine hierarchische Clusteranalyse auf Basis der Orginaldaten für BESCH und UMSATZ (also nicht auf Basis der Ähnlichkeitsmatrix) durchgeführt, wobei die Ward-Alternative für die Distanzmessung verwendet wird und bsp1 der Name der durchgeführten Clusteranalyse ist. Sodann wird das resultierende Dendrogramm dargestellt, wobei einmal für jedes Objekt die Merkmalsausprägung UMSATZ und zum anderen die Ausprägung BESCH angegeben wird. Den zweiten Fall zeigt Abbildung 4/2. Ferner bewirkt der letzte Befehl, dass das Dendrogramm nur oberhalb einer cutvalue-Grenze (für die Distanz) dargestellt wird und dass für jedes Cluster die Anzahl der zugrundeliegenden Beobachtungen angegeben wird. Siehe dazu Abbildung 4/3. Diese eher willkürlichen Beispiele zeigen, dass die Clusteranalyse ein großes Enthüllungsrisiko bieten **kann**, wenn sie entsprechend angewendet wird.

4.1.3 Ein empirisches Beispiel

Welche Enthüllungsrisiken sich auch bei absolut unbefangener Anwendung der Clusteranalyse ergeben können, hat das FDZ des Bundes anhand der Monatsberichte anhand einer fünfzehnprozentigen Teilstichprobe für das Jahr 1999 untersucht. Diese Abgrenzung war erforderlich, weil andernfalls der Umfang des Datensatzes eine Anwendung in STATA nicht zugelassen hätte.[43]

Entsprechend den Beispielen im vorigen Unterabschnitt wurden Cluster mit den beiden Klassifikationsvariablen Umsatz und Beschäftigung gebildet, wobei die Anzahl Cluster mit 5 vorgegeben wurde. Die Tabelle 4.1 berichtet über das Resultat für unterschiedliche Optionen bezüglich der Klumpenbildung[44].

Wesentliches Ergebnis ist, dass in allen Varianten der Klumpenbildung **mindestens zwei** Unternehmen in jeweils einem eigenen Cluster, die wir im Folgenden als "singuläre Cluster" bezeichnen wollen, identifiziert werden, wobei natürlich die vorgegebene Klumpenzahl von 5, vor allem aber die Wahl der Klassifikationsmerkmale Umsatz und Beschäftigung zu beachten ist. Die Wahl von fünf Klumpen ist sicher alles andere als übertrieben, und auch die Wahl der beiden Klassifikationsmerkmale dürfte sich für Unternehmensdaten anbieten. Im übrigen ist interessant, dass nicht nur das Single-Linkage-Verfahren, dem die Eigenschaft der Erkennung von Ausreißern zugeschrieben wird, mehr als zwei singuläre Cluster erkennt.

Tabelle 4.1: Besetzungshäufigkeit der Klumpen

	Verfahren						
Klumpen	WL	CL	SL	AL	CL	WAL	ML
1	7,341	7,505	1	7,505	7,523	7,527	7,527
2	182	24	1	24	6	4	4
3	9	3	1	3	3	1	1
4	1	1	2	1	1	1	1
5	1	1	7,529	1	1	1	1

Datenbasis: 15 %-Stichprobe, Monatsberichte 1999
Erläuterung der Abkürzungen:
WL = Wards Linkage, CL = Centroid Linkage, SL = Single Linkage,
AL = Average Linkage,CL = Complete Linkage,
WAL = Weighted Average Linkage, ML = Median Linkage

Die Bestimmung von singulären Clustern ansich ist noch kein Enthüllungsrisiko. Man könnte hier von einer "spontanen Identifikation" sprechen. Wird jedoch nun die Regressionsanalyse mit einem strategisch gesetzten Dummy (siehe Abschnitt 3.2.1 sowie Appendix A.3) eingesetzt, dann läßt sich sowohl der Umsatz- wie auch der Beschäftigungswert für dieses Unternehmen leicht bestimmen. Dabei ist natürlich hier der Dummy in der Art zu bestimmen, dass das jeweilige Cluster ausgewählt wird. Beispielsweise in STATA ließe sich dies (zur Bestimmung des jeweiligen Umsatzwertes) durch folgende Befehlsequenz erreichen:

[43] Die im Folgenden untersuchte Teilmenge umfaßt 7.435 Unternehmen/Betriebe. Für den Datensatz 1999 insgesamt ergibt sich ein Umfang von rund 50.000 Unternehmen/Betriebe.

[44] Siehe dazu Unterabschnitt 4.1.1.

```
cluster wardslinkage umsatz besch        **** Aufruf des Clusterverfahrens ****
cluster generate beschtype1=groups(5)    **** Generierung der Cluster *********
tab beschtype1, generate(dummy1)         **** Häufigkeiten der Cluster ********
for values i=1(1)5
{regress umsatz dummy1‘i’ if dummy1‘i’==[_n==1]
sort dummy1‘i’ by dummy1‘i’:
tab umsatz if dummy1‘i’==[_n==1] }
```

Aus Gründen der Vertraulichkeit der Amtlichen Mikrodaten können die Ergebnisse dazu hier nicht präsentiert werden. Jedoch ist wegen der zuvor erwähnten theoretischen Ergebnisse bezüglich "strategischer Dummies" klar, dass in dieser Art die Angaben für einzelne Unternehmen in Erfahrung gebracht werden können.

4.2 Hauptkomponenten- und Faktorenanalyse

4.2.1 Allgemeines zur Faktorenanalyse

Die Faktorenanalyse (wie auch die Hauptkomponenten-Analyse) versucht, aus der beobachtbaren Information

$$\boldsymbol{Y} = \begin{pmatrix} y_{11} & y_{12} & \cdots & y_{1m} \\ y_{21} & y_{22} & \cdots & y_{2m} \\ y_{31} & y_{32} & \cdots & y_{3m} \\ \vdots & \vdots & & \vdots \\ y_{n1} & y_{n2} & \cdots & y_{nm} \end{pmatrix}$$

für m verschiedene Merkmale mit jeweils n Beobachtungen auf zugrundeliegende "Faktoren" zurückzuschließen. Formaler: Im **Modell der Faktorenanalyse** soll der $m-$Vektor $\boldsymbol{y}$ der m Merkmale[45] durch einen Vektor von $p(< m)$ "Faktoren" $\boldsymbol{f}^\star$ und einen Störvektor $\boldsymbol{u}^\star$ erklärt werden:

$$\boldsymbol{y} = \boldsymbol{\Lambda}\, \boldsymbol{f}^\star + \boldsymbol{u}^\star$$

Die $(m \times p)$ Parameter-Matrix

$$\boldsymbol{\Lambda} \quad = \quad \begin{pmatrix} \lambda_{11} & \lambda_{12} & \ldots & \lambda_{1p} \\ \lambda_{21} & \lambda_{22} & \ldots & \lambda_{2p} \\ \vdots & \vdots & \ldots & \vdots \\ \lambda_{m1} & \lambda_{m2} & \ldots & \lambda_{mp} \end{pmatrix}$$

bezeichnet man als Matrix der "Faktorladungen". Der p-dimensionale Vektor $\mathbf{f}$ gibt die p "gemeinsamen" Faktoren an. Außerdem stehen im Vektor $\boldsymbol{u}^\star$ die "spezifischen" Faktoren. Dabei deutet der Stern $(*)$ darauf hin, dass die betreffende Variable unbeobachtbar ist. Das Modell sieht formal wie ein Regressionsmodell aus, hat aber den entscheidenden Unterschied, dass auf der rechten Seite nur unbeobachtbare Größen stehen. Deshalb gibt es auch ein Identifikations- bzw. Schätzproblem: Das Produkt $\boldsymbol{\Lambda}\, \boldsymbol{f}^\star$ ist nicht eindeutig parametrisierbar. Man sagt auch, die Ladungsmatrix sei nur bis auf orthonormale Rotationen bestimmbar. Deshalb kann man die Ladungsmatrix unterschiedlich schätzen. Besonders bekannt ist die Option "Varimax-Rotation".

[45] Im Gegensatz zu den vorhergehenden Abschnitten steht $\boldsymbol{y}$ hier für m Merkmale und nicht für n Beobachtungen.

Man beachte, dass die Hauptkomponentenanalyse eng mit der Faktorenanalyse verwandt ist. Dabei gilt insbesondere:

- Die Hauptkomponenten können als Faktoren interpretiert werden.
- In diesem "Modell" gibt es keine spezifischen Störvariablen, die betreffenden spezifischen Varianzen sind Null und die Kommunalitäten (im Fall einer Korrelationsmatrix) damit gleich Eins.

Insofern gelten die hier betrachteten Probleme bezüglich Enthüllungsrisiken auch für die Hauptkomponentenanalyse.

Wesentlich ist die Interpretation der geschätzten Ladungsmatrix. Dazu verwenden wir ein künstliches Beispiel mit $m = 5$ Merkmalen und $p = 3$ Faktoren. Die geschätzte Ladungsmatrix habe die folgende Form:[46]

$$\hat{\mathbf{\Lambda}} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \frac{1}{\sqrt{2}} & 0 \\ 0 & \frac{1}{\sqrt{2}} & 0 \\ 0 & 0 & \frac{1}{\sqrt{2}} \\ 0 & 0 & \frac{1}{\sqrt{2}} \end{pmatrix}$$

Diese idealisierte Situation würde man wie folgt beschreiben: Der erste Faktor "lädt" ausschließlich auf dem ersten Merkmal, der zweite Faktor auf dem zweiten und dritten Merkmal und der dritte Faktor auf dem vierten und fünften Merkmal. Solch eine Situation entspricht der folgenden Kovarianzstruktur:

$$\mathbf{\Sigma} = \left(\begin{array}{c|cc|cc} \sigma_{11} & & & & \\ \hline & \sigma_{22} & \sigma_{23} & & \\ & \sigma_{32} & \sigma_{33} & & \\ \hline & & & \sigma_{44} & \sigma_{43} \\ & & & \sigma_{54} & \sigma_{55} \end{array}\right)$$

Das heißt, das erste Merkmal ist mit allen vier anderen Merkmalen nicht korreliert, das zweite und das dritte Merkmal sind ausschließlich miteinander korreliert, ebenso das vierte und fünfte Merkmal. In der Empirie werden wir diese "ideale" Sitation natürlich nie so antreffen.[47] Weiter unten in diesem Abschnitt wird uns vor allem die erste Spalte von $\mathbf{\Lambda}$ beschäftigen, die ausschließlich an erster Position einen von Null verschiedenen Wert besitzt; dies korrespondiert zu dem ersten Merkmal, dass mit allen anderen unkorreliert ist.

4.2.2 Bestimmung der Faktorwerte

Um die geschätzten Faktorladungen besser interpretieren zu können, stellt man sich vor, dass für die n Beobachtungswerte y_{ij} der m Merkmale bestimmte **realisierte Werte**

[46] Die Werte sind so gewählt, dass die einzelnen Spalten-Vektoren die Länge 1 haben.

[47] Genau so selten, nämlich nie, werden wir die obige Ladungsmatrix erhalten. Vielmehr wird man sehr hohe und weniger hohe Werte bekommen.

$\hat{f}_{ik}, i = 1, \ldots, n; k = 1, \ldots, p$, der Faktoren verantwortlich waren. Daraus möchte man eine Methode gewinnen, um vorzugsweise für $p = 2$ Faktoren eine graphische Darstellung der einzelnen Beobachtungspunkte im Zwei-Faktorenraum zu erreichen. Die dafür verwendete Formel soll im folgenden kurz erläutert werden. Sie benutzt Ergebnisse aus der linearen Algebra zur Lösung von linearen Gleichungssystemen. Siehe beispielsweise Graybill (1969).

Unter Vernachlässigung der spezifischen Faktoren (deren Streuung hier als klein angenommen wird) unterstellen wir, dass (zumindest approximativ)[48]

$$y_{ij} \quad = \quad \sum_{k=1}^{p} \tilde{\lambda}_{jk} \hat{f}_{ik} \tag{4-1}$$

oder auch kompakter

$$\boldsymbol{Y} \quad = \quad \tilde{\boldsymbol{F}} \hat{\boldsymbol{\Lambda}}' \quad . \tag{4-2}$$

gilt. Dabei ist

$$\boldsymbol{Y} = \begin{pmatrix} y_{11} & y_{12} & \cdots & y_{1m} \\ y_{21} & y_{22} & \cdots & y_{2m} \\ y_{31} & y_{32} & \cdots & y_{3m} \\ \vdots & \vdots & & \vdots \\ y_{n1} & y_{n2} & \cdots & y_{nm} \end{pmatrix}$$

die $(n \times m)$-**Datenmatrix** und

$$\tilde{\boldsymbol{F}} \quad = \quad \begin{pmatrix} \hat{f}_{11} & \hat{f}_{12} & \cdots & \hat{f}_{1p} \\ \hat{f}_{21} & \hat{f}_{22} & \cdots & \hat{f}_{2p} \\ \hat{f}_{31} & \hat{f}_{32} & \cdots & \hat{f}_{3p} \\ \vdots & \vdots & & \vdots \\ \hat{f}_{n1} & \hat{f}_{n2} & \cdots & \hat{f}_{np} \end{pmatrix}$$

die $(n \times p)$-**Matrix der zu schätzenden Faktorwerte**. Außerdem ist $\hat{\boldsymbol{\Lambda}}$ die bereits oben eingeführte $(m \times p)-$ Matrix der Faktorladungen; das Dach $\hat{}$ weist darauf hin, dass diese Matrix geschätzt wurde. Dies besagt auch, dass sich je nach Schätz- und Rotationsmethode eine andere Matrix ergibt, was natürlich auch die Lösung für $\tilde{\boldsymbol{F}}$ beeinflußt. Dagegen sind die Faktorwerte in der Matrix $\tilde{\boldsymbol{F}}$ bisher nicht bekannt und sollen nun unter Verwendung der bereits geschätzten Ladungskoeffizienten $\hat{\lambda}_{jk}$ bzw. der geschätzten Ladungsmatrix $\hat{\boldsymbol{\Lambda}}$ bestimmt werden.

Kleinst-Quadrate-Lösung Die Frage ist, welche Werte die geschätzten Werte der Faktoren, also die Variablen $\hat{f}_{ik}$, besitzen. Aus der obigen expliziten Gleichung (4-1) oder auch (4-2) erkennt man, dass die Lösung nicht eindeutig ist. Trotzdem verwendet man die folgende, in gewisser Weise willkürliche Methode, die ihre Berechtigung jedoch aus Ergebnissen der linearen Algebra bezieht.[49] Man postmultipliziert die Gleichung (4-2) mit $\hat{\boldsymbol{\Lambda}}$ und er-

[48] Es soll alternativ unterstellt werden, dass die p gewählten Faktoren durch das Eigenwert- bzw. Kaiser-Kriterium gut abgesichert sind.

[49] Es wird die Lösung für das Gleichungssystem $\boldsymbol{AX} = \boldsymbol{B}$ betrachtet. In unserem Fall ist $\boldsymbol{X} = \hat{\boldsymbol{F}}'$ sowie $\boldsymbol{A} = \boldsymbol{\Lambda}'$ und $\boldsymbol{B} = \boldsymbol{Y}'$. Die Lösung lautet

$$\mathbf{X} = \mathbf{A}^{+}\mathbf{B} = (\mathbf{A}'\mathbf{A})^{-1}\mathbf{A}'\mathbf{B} \quad ,$$

wobei $\mathbf{A}^{+}$ die Moore-Penrose-Inverse von $\mathbf{A}$ ist. Dabei wurde hier unterstellt, dass diese Matrix vollen Spaltenrang hat. Siehe dazu beliebiges Buch über lineare Algebra, insbesondere Ausführungen zur verallgemeinerten Inversen, etwa bei Graybill (1969).

hält dann, weil $\hat{\mathbf{\Lambda}}'\hat{\mathbf{\Lambda}}$ invertierbar ist, die folgende Lösung von (4-2) für die Matrix der "Faktorenwerte" (englisch: factor scores):

$$\hat{\boldsymbol{F}}_{KQ} = \boldsymbol{Y}\,\hat{\mathbf{\Lambda}}\,(\hat{\mathbf{\Lambda}}'\hat{\mathbf{\Lambda}})^{-1} \tag{4-3}$$

Die Lösung ist formal äquivalent mit der Kleinstquadrate-Lösung im multivariaten Regressionsmodell, und das ist kein Zufall, sondern deutet daraufhin, dass die in gewisser Weise beste Approximation an die Lösung bestimmt wird. Deshalb wird sie in der Literatur auch als "Kleinst-Quadrate-Lösung" (Least Squares Solution) bezeichnet.[50]

Bartletts Methode (ML-Schätzung) [51]

Führt man den Gedanken der Kleinst-Quadrate-Schätzung des Vektors $\mathbf{f}$ im Modell[52]

$$\boldsymbol{y} \quad = \quad \mathbf{\Lambda}\,\boldsymbol{f} \quad + \quad \boldsymbol{u} \tag{4-4}$$

weiter, dann liegt es nahe, eine gewichtete Kleinstquadrate Schätzung zu verwenden, d.h. man löst das Minimierungsproblem

$$\min_{d}\,(\mathbf{y} - \hat{\mathbf{\Lambda}})'\,\mathbf{f})\,\hat{\mathbf{\Psi}}^{-1}\,(\mathbf{y} - \hat{\mathbf{\Lambda}})$$

und erhält

$$\hat{\boldsymbol{F}}_{BA} = \boldsymbol{Y}\,\hat{\mathbf{\Psi}}^{-1}\,\hat{\mathbf{\Lambda}}\;(\hat{\mathbf{\Lambda}}'\,\hat{\mathbf{\Psi}}^{-1}\hat{\mathbf{\Lambda}})^{-1} \tag{4-5}$$

wobei $\hat{\mathbf{\Psi}}$ die (geschätzte) Kovarianzmatrix der m spezifischen Faktoren ist, die annahmegemäß Diagonalgestalt hat, d.h.

$$\hat{\mathbf{\Psi}} = \begin{pmatrix} \hat{\psi}_1 & & & & \\ & \hat{\psi}_2 & & & \\ & & \ddots & & \\ & & & \hat{\psi}_{m-1} & \\ & & & & \hat{\psi}_m \end{pmatrix}$$

Diese Schätzmethode wird Barlett (1937) zugeschrieben. Fahrmeir et al (1996) zeigen, dass dies äquivalent mit einer (Pseudo-)ML-Schätzung ist, "Pseudo" deshalb, weil es sich um ein geschätztes Modell handelt, in dem der Vektor $\mathbf{f}$ als Parametervektor interpretiert wird, obwohl er eigentlich ein Zufallsvektor ist!

Die Regressionsmethode von Thurstone/Thomson Wenn man fordert, dass der Schätzer für die Faktorwerte derart gewählt werden soll, dass die Abweichung zu den "wahren" Faktorwerten möglichst klein ist, dann gelangt man zu der Schätzformel[53]

$$\hat{\boldsymbol{F}}_{RE} = \boldsymbol{Y}\,\left(\hat{\mathbf{\Lambda}}\hat{\mathbf{\Lambda}}' + \hat{\mathbf{\Psi}}\right)^{-1}\,\hat{\mathbf{\Lambda}}\,. \tag{4-6}$$

[50] Siehe McDonald und Burr (1967) S. 386. Die Methode geht zurück auf Horst (1965). Backhaus et al. (2008) geben diese Formel ohne jede Quelle an.

[51] Siehe zum Folgenden auch den Abschnitt "Schätzung der Faktorenwerte" in Fahrmeir et al. (1996), Kapitel 11 Abschnitt 5.

[52] Dabei ist $\mathbf{\Lambda}$ eine $(m \times p)$ - Matrix der Ladungskoeffizienten und $\mathbf{f}$ ein p-dimensionaler Zufalls-Vektor der Faktoren, ferner $\mathbf{u}$ ein m-dimensionaler Vektor der "spezifischen Faktoren".

[53] Siehe beispielsweise Roderick et al. (1967) S.387 oder Fahrmeir et al. (1996) S. 691 sowie Bartholomew et al. (2009).

Alternativ läßt sich diese Formel auch wie folgt schreiben:[54]

$$\hat{\boldsymbol{F}}_{RE} = \boldsymbol{Y}\,\hat{\boldsymbol{\Psi}}^{-1}\,\hat{\boldsymbol{\Lambda}}\,\left(\hat{\boldsymbol{\Lambda}}'\hat{\boldsymbol{\Psi}}\hat{\boldsymbol{\Lambda}} + \mathbf{I}_p\right)^{-1} \quad , \tag{4-7}$$

und in dieser Form wird sie auch meistens zitiert. Diese Methode scheint zuerst von Thomson vorgeschlagen worden zu sein, wird aber auch Thurstone (1935) zugeschrieben.[55]

Eine Bewertung der drei Schätzformeln Macdonald and Burr (1967), die die drei Ansätze (gemeinsam mit einem vierten, der hier vernachlässigt wurde, ausführlich bewerten, weisen darauf hin, dass nur die beiden ersten Varianten erwartungstreue Schätzungen der Faktorwerte garantieren, während die dritte Variante zwar die kleinste Varianz besitzt, jedoch verzerrt ist.

Interessant ist, dass in STATA vor allem diese dritte Methode (Regressions-Methode von Thomson bzw. Thurstone) als Option angeboten wird. Alternativ kann man die Methode von Bartlett auswählen. STATA gibt folgende Beschreibung zur Option "predict", die die Faktorwerte bestimmt:

```
regression produces factors scored by the regression method.

bartlett produces factors scored by the method suggested by Bartlett.
    This method produces unbiased factors, but they may be less accurate
    than those produced by the default regression method suggested by
    Thomson.  Regression-scored factors have the smallest mean squared
    error from the true factors but may be biased.
```

Man beachte, dass in STATA die Begriffe exakt vertauscht sind: "regression" meint die Methode von Bartlett gemäß (4-5) und "bartlett" meint die Methode von Thomson/Thurstone gemäß (4-6) bzw. (4-7) .

[54] Diese alternative Form folgt aus folgendem Resultat:

$$\begin{aligned}
\left(\hat{\boldsymbol{\Lambda}}\hat{\boldsymbol{\Lambda}}' + \hat{\boldsymbol{\Psi}}\right)^{-1}\,\hat{\boldsymbol{\Lambda}} &= \left\{\hat{\boldsymbol{\Psi}}^{-1} - \hat{\boldsymbol{\Psi}}^{-1}\hat{\boldsymbol{\Lambda}}\left(\hat{\boldsymbol{\Lambda}}'\hat{\boldsymbol{\Psi}}^{-1}\hat{\boldsymbol{\Lambda}} + \mathbf{I}_p\right)^{-1}\hat{\boldsymbol{\Lambda}}'\hat{\boldsymbol{\Psi}}^{-1}\right\}\hat{\boldsymbol{\Lambda}} \\
&= \hat{\boldsymbol{\Psi}}^{-1}\,\hat{\boldsymbol{\Lambda}}\left\{\mathbf{I} - \left(\hat{\boldsymbol{\Lambda}}'\hat{\boldsymbol{\Psi}}^{-1}\hat{\boldsymbol{\Lambda}} + \mathbf{I}_p\right)^{-1}\hat{\boldsymbol{\Lambda}}'\hat{\boldsymbol{\Psi}}^{-1}\,\hat{\boldsymbol{\Lambda}}\right\} \\
&= \hat{\boldsymbol{\Psi}}^{-1}\,\hat{\boldsymbol{\Lambda}}\left\{\left(\hat{\boldsymbol{\Lambda}}'\hat{\boldsymbol{\Psi}}^{-1}\hat{\boldsymbol{\Lambda}} + \mathbf{I}_p\right)^{-1}\left(\hat{\boldsymbol{\Lambda}}'\hat{\boldsymbol{\Psi}}^{-1}\hat{\boldsymbol{\Lambda}} + \mathbf{I}_p\right) - \left(\hat{\boldsymbol{\Lambda}}'\hat{\boldsymbol{\Psi}}^{-1}\hat{\boldsymbol{\Lambda}} + \mathbf{I}_p\right)^{-1}\hat{\boldsymbol{\Lambda}}'\hat{\boldsymbol{\Psi}}^{-1}\,\hat{\boldsymbol{\Lambda}}\right\} \\
&= \hat{\boldsymbol{\Psi}}^{-1}\,\hat{\boldsymbol{\Lambda}}\left\{\left(\hat{\boldsymbol{\Lambda}}'\hat{\boldsymbol{\Psi}}^{-1}\hat{\boldsymbol{\Lambda}} + \mathbf{I}_p\right)^{-1}\left(\hat{\boldsymbol{\Lambda}}'\hat{\boldsymbol{\Psi}}^{-1}\hat{\boldsymbol{\Lambda}} + \mathbf{I}_p - \hat{\boldsymbol{\Lambda}}'\hat{\boldsymbol{\Psi}}^{-1}\,\hat{\boldsymbol{\Lambda}}\right)\right\} \\
&= \hat{\boldsymbol{\Psi}}^{-1}\,\hat{\boldsymbol{\Lambda}}\left\{\left(\hat{\boldsymbol{\Lambda}}'\hat{\boldsymbol{\Psi}}^{-1}\hat{\boldsymbol{\Lambda}} + \mathbf{I}_p\right)^{-1}(\mathbf{I}_p)\right\}
\end{aligned}$$

Dabei wird bei der Umformung der Inversen in der ersten Zeile eine Formel verwendet, wie sie beispielsweise bei Press (1972) S. 23 ("Binomial Inverse Theorem") zu finden ist.

[55] Eine ausführliche Diskussion der verschiedenen Methoden zur Schätzung von Faktorwerten wird in Roderick et al. (1967) gegeben. Vor allem die Methode von Thomson wird ferner sehr anschaulich in Bartholomew et al. (2009) beschrieben.

4.2.3 Enthüllungsrisiken bei speziellen Korrelationsstrukturen

Verwendung der Kleinst-Quadrate-Formel Wir wollen uns nun ausführlicher mit der für die Faktorwerte gefundenen Lösung (4-3) beschäftigen. In unserem idealisierten Fall hat die Matrix $\mathbf{\Lambda}$ eine blockdiagonale Struktur und wir erhalten deshalb

$$(\hat{\mathbf{\Lambda}}'\hat{\mathbf{\Lambda}}) \quad = \quad \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

und damit ergibt sich in unserem speziellen Fall für die Lösung (4-3) der Ausdruck

$$\begin{aligned} \hat{\boldsymbol{F}}_{KQ} &= \mathbf{Y}\,\hat{\mathbf{\Lambda}} \\ &= \begin{pmatrix} y_{11} & y_{12} & \cdots & y_{15} \\ y_{21} & y_{22} & \cdots & y_{25} \\ y_{31} & y_{32} & \cdots & y_{35} \\ \vdots & \vdots & & \vdots \\ y_{n1} & y_{n2} & \cdots & y_{n5} \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & \frac{1}{\sqrt{2}} & 0 \\ 0 & \frac{1}{\sqrt{2}} & 0 \\ 0 & 0 & \frac{1}{\sqrt{2}} \\ 0 & 0 & \frac{1}{\sqrt{2}} \end{pmatrix} \\ &= \left(\begin{pmatrix} y_{11} \\ y_{21} \\ y_{31} \\ \vdots \\ y_{n1} \end{pmatrix} \middle| \frac{1}{\sqrt{2}} \cdot \left\{ \begin{pmatrix} y_{12} \\ y_{22} \\ y_{32} \\ \vdots \\ y_{n2} \end{pmatrix} + \begin{pmatrix} y_{13} \\ y_{23} \\ y_{33} \\ \vdots \\ y_{n3} \end{pmatrix} \right\} \middle| \frac{1}{\sqrt{2}} \cdot \left\{ \begin{pmatrix} y_{14} \\ y_{24} \\ y_{34} \\ \vdots \\ y_{n4} \end{pmatrix} + \begin{pmatrix} y_{15} \\ y_{25} \\ y_{35} \\ \vdots \\ y_{n5} \end{pmatrix} \right\} \right) \end{aligned} \tag{4-8}$$

Die Matrix $\hat{\boldsymbol{F}}$ hat 3 Spalten: Jede Spalte gibt die Faktorwerte des jeweiligen Faktors an. In unserem speziellen Fall ist allerdings der Vektor der n Faktorwerte für den **ersten** Faktor **exakt identisch mit den n Beobachtungswerten für das erste Merkmal**. Dies bedeutet, dass bei entsprechender Korrelationsstruktur durchaus ein Enthüllungsrisiko für einzelne Merkmale bestehen kann.

Verwendung der Bartlett-Formel Allerdings ist dieses Resultat für die spezielle Schätzformel (4-3) abgeleitet worden. Verwendet man dagegen die Schätzformel von Bartlett, dann erhalten wir aus (4-5)

$$\begin{aligned} \hat{\boldsymbol{F}}_{BA} &= \boldsymbol{Y}\,\hat{\mathbf{\Psi}}^{-1}\,\hat{\mathbf{\Lambda}}\,(\hat{\mathbf{\Lambda}}'\,\hat{\mathbf{\Psi}}^{-1}\hat{\mathbf{\Lambda}})^{-1} \\ &= \begin{pmatrix} y_{11} & y_{12} & \cdots & y_{15} \\ y_{21} & y_{22} & \cdots & y_{25} \\ y_{31} & y_{32} & \cdots & y_{35} \\ \vdots & \vdots & & \vdots \\ y_{n1} & y_{n2} & \cdots & y_{n5} \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & \frac{1}{\sqrt{2}} & 0 \\ 0 & \frac{1}{\sqrt{2}} & 0 \\ 0 & 0 & \frac{1}{\sqrt{2}} \\ 0 & 0 & \frac{1}{\sqrt{2}} \end{pmatrix} \end{aligned}$$

Wir betrachten zunächst für $m = 5$

$$\hat{\mathbf{\Psi}}^{-1}\,\hat{\mathbf{\Lambda}} = \begin{pmatrix} \frac{1}{\hat{\psi}_1} & & & & \\ & \frac{1}{\hat{\psi}_2} & & & \\ & & \frac{1}{\hat{\psi}_3} & & \\ & & & \frac{1}{\hat{\psi}_4} & \\ & & & & \frac{1}{\hat{\psi}_5} \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & \frac{1}{\sqrt{2}} & 0 \\ 0 & \frac{1}{\sqrt{2}} & 0 \\ 0 & 0 & \frac{1}{\sqrt{2}} \\ 0 & 0 & \frac{1}{\sqrt{2}} \end{pmatrix}$$

$$= \begin{pmatrix} \frac{1}{\hat{\psi}_1} & 0 & 0 \\ 0 & \frac{1}{\sqrt{2}\,\hat{\psi}_2} & 0 \\ 0 & \frac{1}{\sqrt{2}\,\hat{\psi}_3} & 0 \\ 0 & 0 & \frac{1}{\sqrt{2}\,\hat{\psi}_4} \\ 0 & 0 & \frac{1}{\sqrt{2}\,\hat{\psi}_5} \end{pmatrix}$$

sowie

$$\hat{\mathbf{\Lambda}}'\,\hat{\mathbf{\Psi}}^{-1}\,\hat{\mathbf{\Lambda}} = \begin{pmatrix} \frac{1}{\hat{\psi}_1} & 0 & 0 \\ 0 & \frac{1}{2\,\hat{\psi}_2} + \frac{1}{2\,\hat{\psi}_3} & 0 \\ 0 & 0 & \frac{1}{2\,\hat{\psi}_4} + \frac{1}{2\,\hat{\psi}_5} \end{pmatrix}$$

und demnach

$$\hat{\mathbf{\Psi}}^{-1}\hat{\mathbf{\Lambda}}\left(\hat{\mathbf{\Lambda}}'\hat{\mathbf{\Psi}}^{-1}\hat{\mathbf{\Lambda}}\right)^{-1} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \frac{1}{\sqrt{2}\,\hat{\psi}_2}\left(\frac{1}{2\,\hat{\psi}_2} + \frac{1}{2\,\hat{\psi}_3}\right)^{-1} & 0 \\ 0 & \frac{1}{\sqrt{2}\,\hat{\psi}_3}\left(\frac{1}{2\,\hat{\psi}_2} + \frac{1}{2\,\hat{\psi}_3}\right)^{-1} & 0 \\ 0 & 0 & \frac{1}{\sqrt{2}\,\hat{\psi}_4}\left(\frac{1}{2\,\hat{\psi}_4} + \frac{1}{2\,\hat{\psi}_5}\right)^{-1} \\ 0 & 0 & \frac{1}{\sqrt{2}\,\hat{\psi}_5}\left(\frac{1}{2\,\hat{\psi}_4} + \frac{1}{2\,\hat{\psi}_5}\right)^{-1} \end{pmatrix} \quad (4\text{-}9)$$

Ein Vergleich von (4-8) mit (4-9) offenbart, dass auch in diesem Fall, also bei Verwendung der Bartlett-Formel, wegen der ersten Spalte, die eine Eins und sonst Nullen aufweist, die Faktorwerte für den ersten Faktor wieder den Datenvektor der ersten Spalte aus der Datenmatrix $\mathbf{Y}$ identifizieren.

Verwendung der Thomson/Thurstone-Formel Wir untersuchen jetzt auch die Schätzformel von Thomson/Thurstone, die durch

$$\hat{\boldsymbol{F}}_{RE} = \boldsymbol{Y}\left(\hat{\mathbf{\Lambda}}\hat{\mathbf{\Lambda}}' + \hat{\mathbf{\Psi}}\right)^{-1}\hat{\mathbf{\Lambda}}$$

gegeben ist. Für unser Beispiel erhalten wir

$$\hat{\mathbf{\Lambda}}\hat{\mathbf{\Lambda}}' = \left(\begin{array}{c|cc|cc} 1 & 0 & 0 & 0 & 0 \\ \hline 0 & \frac{1}{2} & \frac{1}{2} & 0 & 0 \\ 0 & \frac{1}{2} & \frac{1}{2} & 0 & 0 \\ \hline 0 & 0 & 0 & \frac{1}{2} & \frac{1}{2} \\ 0 & 0 & 0 & \frac{1}{2} & \frac{1}{2} \end{array}\right)$$

sowie

$$\left(\hat{\mathbf{\Lambda}}\hat{\mathbf{\Lambda}}' + \hat{\mathbf{\Psi}}\right)^{-1} \hat{\mathbf{\Lambda}}$$

$$= \begin{pmatrix} (1+\hat{\psi}_1)^{-1} & & \\ & \begin{pmatrix} \frac{1}{2}+\hat{\psi}_2 & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2}+\hat{\psi}_3 \end{pmatrix}^{-1} & \\ & & \begin{pmatrix} \frac{1}{2}+\hat{\psi}_4 & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2}+\hat{\psi}_5 \end{pmatrix}^{-1} \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & \frac{1}{\sqrt{2}} & 0 \\ 0 & \frac{1}{\sqrt{2}} & 0 \\ 0 & 0 & \frac{1}{\sqrt{2}} \\ 0 & 0 & \frac{1}{\sqrt{2}} \end{pmatrix}$$

$$= \begin{pmatrix} (1+\hat{\psi}_1)^{-1} & & \\ & \begin{pmatrix} \frac{1}{2}+\hat{\psi}_2 & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2}+\hat{\psi}_3 \end{pmatrix}^{-1} \begin{pmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{pmatrix} & \\ & & \begin{pmatrix} \frac{1}{2}+\hat{\psi}_4 & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2}+\hat{\psi}_5 \end{pmatrix}^{-1} \begin{pmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{pmatrix} \end{pmatrix}$$

In diesem Fall würde sich also für die Faktorwerte für den ersten Faktor der Datenvektor für das erste Merkmal ergeben, allerdings für alle n Beobachtungswerte mit dem Faktor

$$\frac{1}{1+\hat{\psi}_1}$$

multipliziert. Da die Schätzwerte für die spezifischen Varianzen in der Faktorenanalyse ausgewiesen werden, wäre also auch in diesem Fall - nach Multiplikation mit dem Kehrwert - eine Re-Identifikation der Beobachtungswerte möglich.

Wir wollen auch noch die spezielle Situation für die alternative Formel (4-7) darstellen. Dazu können wir auf Ergebnisse zurückgreifen, die wir oben für den Bartlett-Fall bestimmt

haben. Wir erhalten

$$
\begin{aligned}
&\hat{\mathbf{\Psi}}^{-1}\hat{\mathbf{\Lambda}}\left(\hat{\mathbf{\Lambda}}'\hat{\mathbf{\Psi}}\hat{\mathbf{\Lambda}}+\mathbf{I}_p\right)^{-1}\\
&= \begin{pmatrix} \frac{1}{\hat{\psi}_1} & 0 & 0 \\ 0 & \frac{1}{\sqrt{2}\hat{\psi}_2} & 0 \\ 0 & \frac{1}{\sqrt{2}\hat{\psi}_3} & 0 \\ 0 & 0 & \frac{1}{\sqrt{2}\hat{\psi}_4} \\ 0 & 0 & \frac{1}{\sqrt{2}\hat{\psi}_5} \end{pmatrix} \begin{pmatrix} \frac{1}{\hat{\psi}_1}+1 & 0 & 0 \\ 0 & \frac{1}{2\hat{\psi}_2}+\frac{1}{2\hat{\psi}_3}+1 & 0 \\ 0 & 0 & \frac{1}{2\hat{\psi}_4}+\frac{1}{2\hat{\psi}_5}+1 \end{pmatrix}^{-1}\\
&= \begin{pmatrix} \frac{1}{\hat{\psi}_1} & 0 & 0 \\ 0 & \frac{1}{\sqrt{2}\hat{\psi}_2} & 0 \\ 0 & \frac{1}{\sqrt{2}\hat{\psi}_3} & 0 \\ 0 & 0 & \frac{1}{\sqrt{2}\hat{\psi}_4} \\ 0 & 0 & \frac{1}{\sqrt{2}\hat{\psi}_5} \end{pmatrix} \begin{pmatrix} \frac{\hat{\psi}_1}{\hat{\psi}_1+1} & 0 & 0 \\ 0 & \left(\frac{1}{2\hat{\psi}_2}+\frac{1}{2\hat{\psi}_3}+1\right)^{-1} & 0 \\ 0 & 0 & \left(\frac{1}{2\hat{\psi}_4}+\frac{1}{2\hat{\psi}_5}+1\right)^{-1} \end{pmatrix}\\
&= \begin{pmatrix} \frac{1}{1+\hat{\psi}_1} & 0 & 0 \\ 0 & \frac{1}{\sqrt{2}\hat{\psi}_2}\left(\frac{1}{2\hat{\psi}_2}+\frac{1}{2\hat{\psi}_3}+1\right)^{-1} & 0 \\ 0 & \frac{1}{\sqrt{2}\hat{\psi}_3}\left(\frac{1}{2\hat{\psi}_2}+\frac{1}{2\hat{\psi}_3}+1\right)^{-1} & 0 \\ 0 & 0 & \frac{1}{\sqrt{2}\hat{\psi}_4}\left(\frac{1}{2\hat{\psi}_4}+\frac{1}{2\hat{\psi}_5}+1\right)^{-1} \\ 0 & 0 & \frac{1}{\sqrt{2}\hat{\psi}_5}\left(\frac{1}{2\hat{\psi}_4}+\frac{1}{2\hat{\psi}_5}+1\right)^{-1} \end{pmatrix}
\end{aligned}
$$

Auch aus dieser Formel folgt, dass der erste Faktor proportional zum ersten Datenvektor ist. Siehe die Diskussion oben.

Ergänzende Bemerkungen dass das hier beschriebene Enthüllungsrisiko auch bei "sehr geringer Korrelation" (statt Null-Korrelation) zumindest in verringertem Umfang besteht, wird im folgenden Unterabschnitt untersucht.

4.2.4 Daten mit empirischer Fast-Null-Korrelation

Wenn man als Angreifer die oben abgeleiteten Ergebnisse in der Weise ausnutzt, dass man die "Zielvariable", also beispielsweise den Umsatz, in einem Datensatz mit anderen Merkmalen kombiniert, die relativ gering mit dieser Zielvariablen korreliert sind, dann kann man in einer Faktorenanalyse durch Berechnung der Faktorwerte zumindest approximativ die Werte der Zielvariablen identifizieren. Erste empirische Ergebnisse dazu, die auf dem IAB-Beschäftigten-Panel basieren, zeigen, dass bei "geschickter" Wahl der zusätzlichen Merkmale die Reidentifikation des Datenvektors der Zielvariablen mit erstaunlich hoher Präzision gelingt.

Dazu wurde der (logarithmierte) Umsatz mit folgenden Merkmalen in einem Datensatz kombiniert:

(1) Anteil (%) der Investitionen in Information und Kommunikation,

(2) Anzahl der Beamtenanwärter, Anzahl an sofort freier Stellen für einfache Tätigkeiten

(3) insgesamt und

(4) davon bei der BA gemeldet,

(5) Anzahl an sofort freier, bei der BA gemeldeter Stellen für Ausgebildete,

(6) Anzahl geförderter Personen,

(7) davon über 50 jährige.

Für alle 7 Merkmale gilt, dass sie eine sehr geringe Korrelation (< 0.10) mit dem (logarithmierten) Umsatz aufweisen. Anders gesagt: Diese Merkmale wurden aufgrund einer Korrelationsmatrix für alle verfügbaren Merkmale ausgewählt.

Die Ladungsmatrix (in diesem Fall mit der Hauptkomponenten-Analyse erzeugt) ergab eine entsprechend "zerlegte" Struktur: Der erste Faktor (PC1) lädt mit Gewicht 0.99 auf dem Merkmal Umsatz, während die übrigen Merkmale für diesen Faktor praktisch keine Rolle spielen. Siehe die folgende Tabelle.

	$PC1$	$PC2$	$PC3$	$PC4$	$PC5$	$PC6$	$PC7$	$PC8$
Umsatz	**0.999**	$-5.4\,10^{-9}$	0	0	$-1.2\,10^{-9}$	0	0	0
InvIuK	$6.2\,10^{-11}$	-0.005	0.999	0.0009	-0.029	-0.002	0.007	$3.7\,10^{-5}$
BeaAnw	$-1.6\,10^{-13}$	$8.2\,10^{-6}$	$-2.9\,10^{-5}$	-0.0008	-0.0002	-0.001	-0.002	0.999
FreiEinfTä	$2.4\,10^{-10}$	0.005	$-8.6\,10^{-5}$	-0.419	0.180	0.089	0.885	0.002
FreiEinfTäBA	$8.9\,10^{-11}$	0.004	0.005	-0.885	0.055	-0.216	-0.408	-0.002
FreiAusTäBA	$1.2\,10^{-9}$	0.009	0.029	0.140	0.977	-0.093	-0.123	$-5.7\,10^{-5}$
GeföPers	$5.2\,10^{-9}$	0.981	0.003	0.032	-0.027	-0.185	0.034	-0.0002
GeföPers50	$1.1\,10^{-9}$	0.192	0.006	-0.142	0.086	0.950	-0.182	0.0005

Die Abbildung 4/4 zeigt das Ergebnis für die Faktorwerte des ersten Faktors graphisch: Die größten Abweichungen zwischen den wahren (logarithmierten) Umsatz-Werten und den aus den Faktorwerten bestimmten Ergebnissen sind von der Größenordnung 10^{-8} !! Für die 1000 größten Unternehmen ist die Abweichung noch deutlich geringer !! Ergebnisse, die inzwischen für den Umsatz selbst analysiert wurden, kommen zu ähnlich deutlichen Ergebnissen bezüglich des Enthüllungsrisikos.

4.3 Multidimensionale Skalierung

4.3.1 Einführende Bemerkungen

Die Multidimensionale Skalierung (MDS)[56] hat in der Psychologie, vor allem in der Psychometrie, eine lange Tradition. Auch im Marketing wird das Instrument seit langem genutzt. Wir werden sehen, dass eine enge Beziehung zur **Faktorenanalyse** oder auch **Hauptkomponentenanalyse** besteht. Die gilt vor allem für die **"klassische" MDS**, die metrische Merkmale oder besser Euklidische Distanzen zwischen verschiedenen Objekten voraussetzt. Der wesentliche Unterschied zwischen Faktoren/Hauptkomponentenanalyse und MDS läßt sich wie folgt beschreiben:

- In der Faktorenanalyse/Hauptkomponentenanalyse stehen die Beobachtungswerte für die Merkmale, die die Objekte charakterisieren, zur Verfügung.Die Objekte werden sodann in einem niedrig-dimensionalen Faktor-Raum positioniert.

[56]Der Text dieses Abschnitts orientiert sich an eigenen Vorlesungsunterlagen zur Veranstaltung "Multivariate Verfahren".

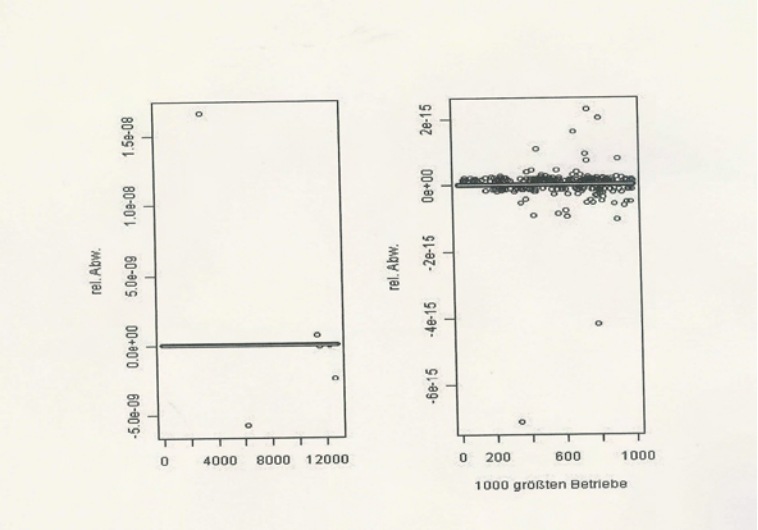


Abbildung 4/4: Die Abweichungen der Faktorwerte für den ersten Faktor von den (logarithmierten) Umsatzwerten (links insgesamt, rechts für die größten 1000 Unternehmen)

- In der MDS stehen typischerweise nur **eindimensionale Informationen über die Ähnlichkeit von Objekten** zur Verfügung. Aus dieser Information soll die Positionierung in einem niedrig-dimensionalen Raum, eine sogenannte "**Konfiguration**" abgeleitet werden. Im Gegensatz zur Faktorenanalyse können dabei die Faktoren nicht als Linearkombinationen von Merkmalsvektoren dargestellt werden (weil diese Information nicht vorhanden ist).

Der Begriff der **(UN-)Ähnlichkeitsmaße** wird auch im Zusammenhang mit der Clusteranalyse behandelt. Dort geht es vorwiegend darum, ob sich die Ähnlichkeitsmaße auf metrische oder (binäre) Nominalvariablen bezogen. Bei der MDS ist es dagegen wesentlicher, ob man - wie bei der klassischen MDS – das Maß als Distanz im Euklidischen Sinne versteht oder bezüglich der (Un)Ähnlichkeit **nur die Ranginformation** verwendet, also beispielsweise die Information, dass Gut A und Gut B einander ähnlicher sind als Gut A und Gut C und dieses Paar wiederum einander ähnlicher ist als die Güter B und C. Siehe dazu Tabelle 4.2. Wenn nur derartige Information vorliegt, dann spricht man von einer **nichtmetrischen** MDS.

Tabelle 4.2: Ordinale Ähnlichkeitsmaße für drei Güter

	Gut A	Gut B	Gut C
Gut A	0		
Gut B	1	0	
Gut C	2	3	0
Hinweis: 0 = maximale Ähnlichkeit			

Ein bekanntes Beispiel ist die "Positionierung" von Städten in einem zweidimensionalen Raum mittels MDS.[57] Falls man die exakten Entfernungen für alle Städtepaare kennt,

[57] Siehe dazu beispielsweise das Lehrbuch von Backhaus et al. (2008) neueste oder frühere Auflagen.

handelt es sich um eine "klassische" MDS. Stehen dagegen (ggfs. aus Gründen der Anonymität!!) nur die Ränge für die einzelnen Distanzen zur Verfügung, dann kommt die nichtmetrische MDS zum Einsatz.

dass die Bestimmung der "Lage" der einzelnen Städte aus der Distanzmatrix Probleme aufwirft, zeigt die folgende Betrachtung der drei Distanzen zwischen den Städten Berlin, Frankfurt und Basel. Für diese drei Städte ist bekannt, dass Berlin und Basel 874 km voneinander entfernt sind. Wenn wir zwei Punkte für Berlin und Basel im Abstand von 874 km zeichnen und dann die Lage von Frankfurt dadurch bestimmen, dass wir um Berlin einen Kreis mit dem Radius von 555 km und um Basel einen Kreis mit dem Radius 337 km ziehen, so ergeben sich **zwei** Schnittpunkte. Siehe Abbildung 4/5. Daraus sieht man, dass die Darstellung **nicht eindeutig** ist.

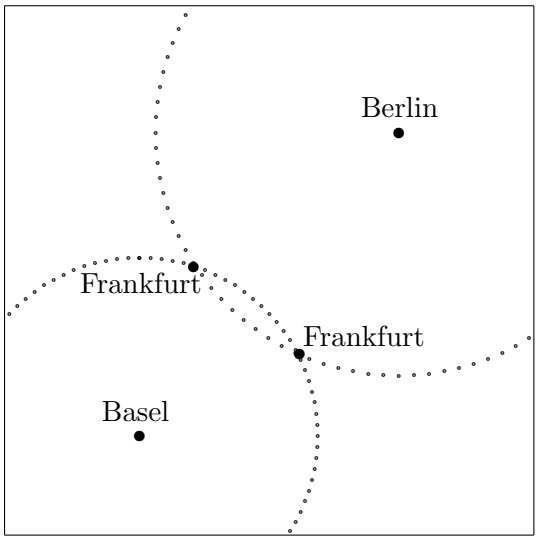


Abbildung 4/5: Die Lösung für drei Städte

4.3.2 MDS und Remote Access

Die vorausgegangenen Ausführungen machen deutlich, dass das Enthüllungsrisiko bei der Verwendung der MDS allenfalls im "klassischen" Fall eine Rolle spielt: Nur hier stehen die Mikrodaten überhaupt zur Verfügung, während im "nichtmetrischen" Fall nur Distanz- bzw. Ähnlichkeitsinformationen vorliegen. Andererseits ist die klassische MDS äquivalent mit der Hauptkomponentenanalyse. Insofern genügt es bezüglich eines möglichen Enthüllungsrisikos bei Remote Access, auf den Abschnitt 4.2 zu verweisen.

5 Abschließende Bemerkungen

Wir haben diese Ausarbeitung mit dem Titel "Remote Access. Eine Welt ohne Mikrodaten ??" versehen. Dabei stehen die beiden Fragezeichen dafür, dass Datennutzer bei entsprechender Schurken-Mentalität[58] durchaus an bestimmte Einzeldaten gelangen können. Die

[58]Uns gefällt dieses Wort weit besser als "Angreifer-Mentalität", weil es darauf hinweist, dass wissenschaftlich orientierte Datennutzer im allgemeinen kein Interesse an der Bestimmung von Einzelwerten haben.

hier präsentierten Beispiele sind sicher nur ein kleiner Ausschnitt aus dem, was möglich ist, vor allem wenn mathematisch-statistisch Begabte sich mit diesem "Problem" beschäftigen.

Die präsentierten Beispiele zeigen aber auch, dass die Überprüfung der Outputs bei Remote Access einen hohen formal-statistischen Sachverstand erfordert; eine entsprechend qualifizierte Gruppe ist beim Server, im Zweifel also in den Statistischen Ämtern, vorzuhalten, was die Kosten dieses Ansatzes deutlich erhöhen wird. Und als Folge (sprich Kostenreduzierung) wird es sicher eine Tendenz geben, nicht die angeforderten Analysen (queries) zu liefern, sondern solche, die auf reduzierter Dateninformation basieren, womit sich dieser Ansatz dem Ansatz des Rechnens mit anonymisierten Daten annähert.

6 Literatur

Backhaus, K., Erichson, B., Plinke, W., Weiber, R. (2008). *Multivariate Analysemethoden. Eine anwendungsorientierte Einführung.* 12. vollständig überarbeitete Auflage, 2008, Springer: Berlin.

Bartolomew, D.J, I.J.Deary und M. Lawn (2009). The origin of factor scores: Spearman, Thomson and Bartlett. *British Journal of Mathematical and Statistical Psychology 62*, 569 - 582.

Bartlett, M.S. (1937). The statistical conception of mental factors. *British Journal of Psychology 28*, 97-104.

Bleninger, P,. Drechsler, J., and Ronning, G. (2010). Remote data access and the risk of disclosure from linear regression: An empirical study. Paper submitted for presentation at the conference 'PRIVACY IN STATISTICAL DATABASES 2010 (PSD 2010)', May 2010.

T.F. Cox und M.A.A. Cox (1994): Multidimensional Scaling. London.

Dobra, A, A.F. Karr und A.P. Sanil (2002). Preserving confidentiality of high-dimensional tabulated data: Statistical and computations issues. *Statistics and Computing 13*, 363-370.

Dobra, A, A.F. Karr, A.P. Sanil und S.E. Fienberg (2002). Sofware Systems for tabular data releases. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems 10*, 529-544.

Duncan, G.T., und Mukherjee, S. (2000). Optimal Disclosure Limitation strategy in statistical databeses: Deterring tracker attacks through additive noise. *Journal of the American Statistical Association 95*, 720-729.

Everitt, Brian S. und Graham Dunn (1991): Applied Multivariate Analysis. Edward Arnold: London.

Fahrmeir, L., A. Hamerle und G. Tutz (Herausgeber) 1996). *Multivariate statistische Verfahren.* De Gruyter: Berlin, zweite Auflage.

Gelman, A., Y. Goegebeur, F. Tuerlinckx und I. Van Mechelen (2000). Diagnostic Checks for Discrete Data Regression Models Using Posterior Predictive Simulations. *Applied Statistics 49*, 247-268.

Giessing,S., und S. Dittrich (2005). Tabellengeheimhaltung im statistischen Verbund – ein Verfahrensvergleich am Beispiel der Umsatzsteuerstatistik. *Wirtschaft und Statistik 8/2006*, 805-814.

Gomatam, S., A.F. Karr, J.P. Reiter und A.P. Sanil (2005). Data Disemination and Disclosure Limitation in a World Without Microdata: A Risk-Utility Framework for Remote Access Analysis Servers. *Statistical Science 20*, 163-177.

Graybill, F.A. (1969). *Introduction to Matrices with Applications in Statistics.* Wadsworth Publ. Co: Belmont.

Hamerle, A. und H. Pape (1996). *Grundlagen der mehrdimensionalen Skalierung.* In: Fahrmeir et al. (1996), 765-793.

Heitzig, J. (2004) “Protection of Confidential Data when Publishing Correlation Matrices”, in: Proceedings in Computational Statistics (16th COMPSTAT Symposium), 1163–1170

Heitzig, J. (2005): The ”Jackknife” Method: Confidentiality Protection for complex statistical Analyses, UNECE/Eurostat Work Session on Statistical Data Confidentiality, Ge neva., http://www.unece.org/stats/documents/ece/ces/ge.46/2005/wp.39.e.pdf.

Hoaglin, D.C., und R.E. Welsh (1978). The Hat Matrix in Regression and ANOVA. *The American Statistician 32*, 17-22.

Horst, P. (1965). *Factor analysis of data matrices*. Holt, Rinehart & Winston, New York.

Hundepool, A., J. Domingo-Ferrer, L. Franconi, S. Giessing, R. Lenz, J. Longhurst, E. Schulte Nordholt, G. Seri und P.-P. de Wolf (2009). *Handbook on Statistical Disclosure Control*. Version 1.1. January 2009. ESSNet SDC.
http://neon.vb.cbs.nl/casc/SDC_Handbook.pdf

Karr, A.F., A. Dobra und A.P. Sanil (2003). Table servers protect confidentiality in tabular data releases. *Communications of the ACM 46*, 57-58.

Kaufmann, H. und H. Pape (1996). *Clusteranalyse*. In: Fahrmeir et al. (1996), 437-536.

Keller-McNulty, S. und E.A. Unger (1998). A database system prototype for remote access to information based on confidential data. *Journal of Official Statistics 14*, 347-360.

Landwehr, J.M., D. Pregibon and A.C. Shoemaker (1984). Graphical Methods for Assessing Logistic Regression Models (with discussion). *Journal of the American Statistical Association 79*, 61-71.

McDonald, R.P., and Burr, E.J. (1967). A comparison of four methods for constructing factor scores. *Psychometrika 32*, 381-401.

Press, S.J. (1972). *Applied Multivariate Analysis*. Holt Rinehart and Winston: New York.

Reiter, J.P. (2003). Model diagnostics for remote access regression servers. *Statistics and Computing 13*, 371-380.

Reiter, J.P., und C.Kohnen (2005). Categorical data regression diagnostics for remote servers. *Journal of Statistical Computation and Simulation 75*, 889-903.

Reznek, A.P., and T. L. Riggs (2005). Disclosure Risks in Releasing Output Based on Regression Residuals. *American Statistical Association 2005*. Proceedings of the Section on Government Statistics and Section on Social Statistics, 1397-1404.

Ronning, G. (1991). *Mikroökonometrie*. Springer, Berlin.

Schönfeld, P. (1969). *Methoden der Ökonometrie. Band I*. Vahlen:Berlin.

Schouten, B., und Cigrang,M. (2003). Remote Access Systems for Statistical Analysis of Microdata. *Statistics and Computing 13*, 381-389.

Thurstone, L.L. (1935). *The vectors of mind*. Chicago, University of Chicago Press.

A Effekt von künstlichen Ausreißern und strategischen Dummies auf den Vektor $\hat{\mathbf{y}}$

Die von Gomatam et al. (2005) angeregte Diskussion über die Risiken der Enthüllung bei Transformation, die zu einem künstlichen Ausreißer führt, oder Setzung einer Dummyvariable, die die Kenntnis für ein bestimmtes Unternehmen ausnutzt (siehe Abschnitte 3.3 und 3.2.1), sollen hier formal konkretisiert werden.

A.1 Hebelwirkung (Leverage) und Hatmatrix

Wir betrachten das Modell der Einfachregression, das in Vektorschreibweise durch

$$\mathbf{y} \quad = \quad \mathbf{X}\,\boldsymbol{\beta} + \mathbf{u} \tag{A-1}$$

mit

$$\mathbf{X} \quad = \quad \left(\boldsymbol{\iota} \quad \mathbf{x} \right)$$

gegeben ist. Dabei sind $\mathbf{y}, \mathbf{x}$ und $\mathbf{u}$ n-dimensionale Vektoren, $\mathbf{X}$ demnach eine $(n \times 2)$-Matrix und $\boldsymbol{\iota}$ ist der Eins-Vektor. Für die geschätzten Werte der abhängigen Variablen y gilt

$$\hat{\mathbf{y}} \quad = \quad \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{y} \tag{A-2}$$

Oftmals[59] schreibt man auch für die Projektionsmatrix

$$\mathbf{H} \quad = \quad \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}' \qquad ,$$

und nennt sie im Englischen "hat matrix", was angesichts (A-2) eine plausible Bezeichnung ist. Wegen der Beziehung

$$\mathbf{HX} \quad = \quad \mathbf{X}$$

ist klar, dass für jedes $i, i = 1, \ldots, n$ die Beziehung

$$\sum_j h_{ij} \quad = \quad 1$$

gilt, sofern die Matrix $\mathbf{X}$ den Einsvektor als Spalte enthält (inhomogene Regression). Da das Element h_{ii} stets nichtnegativ ist und ferner wegen der Idempotenz von $\mathbf{H}$

$$h_{ii} \quad = \quad h_{ii}^2 + \sum_{j \neq i} h_{ij}^2 \quad \Rightarrow \quad h_{ii} \geq h_{ii}^2$$

gilt, ergibt sich ferner

$$0 \quad \leq \quad h_{ii} \quad \leq 1 \qquad .$$

Da außerdem der Rang einer idempotenten Matrix gleich deren Spur ist und für diese Matrix

$$tr(\mathbf{H}) \quad = \quad \sum_{i=1}^{n} h_{ii} \qquad = \quad p$$

[59] Siehe für das Folgende beispielsweise Hoaglin und Welsh (1978).

gilt (p die Anzahl Spalten in der Matrix $\mathbf{X}$), ist der durchschnittliche Wert für alle n Diagonalelemente p/n. Als Daumenregel bezeichnet man ein Diagonalelement mit

$$h_{ii} \quad > \quad \frac{2\,p}{n}$$

als einflußreich bzw. mit hoher Hebelwirkung behaftet. In den folgenden Abschnitten werden wir Situationen betrachten, in denen für bestimmtes i der Diagonalwert den Wert 1 erreicht und alle h_{ij}, $j \neq i$, gleich 0 sind. Dies ist der Fall einer extremen Hebelwirkung.

A.2 Künstliche Ausreißer

Im folgenden nehmen wir an, dass die Daten derart geordnet sind, dass die Informationen für das **erste** Unternehmen gegeben sind, d.h. der Wert der Regressorvariablen, den wir mit z_1 bezeichnen, geht gegen Unendlich. Siehe dazu die Formel (3-3). Statt des Index $''m''$ dort verwenden wir hier den Index $''1''$.

Im speziellen Fall der obigen Regressormatrix für die Einfachregression sind die einzelnen Elemente durch

$$h_{jk} \quad = \quad \frac{1}{n^2 s_z^2}\left(\sum_{i=1}^{n} z_i^2 - x_j \sum_{i=1}^{n} z_i - z_k \sum_{i=1}^{n} z_i + n z_j z_k\right) \quad , j,k = 1,2,\ldots,n,$$

gegeben.[60] Für den j-ten Wert von $\hat{\mathbf{y}}$ ergibt sich dann

$$\begin{aligned} \hat{y}_j &= \sum_{k=1}^{n} h_{jk} y_k \\ &= \frac{1}{n^2 s_z^2} \sum_{k=1}^{n}\left(\sum_{i=1}^{n} z_i^2 - z_j \sum_{i=1}^{n} z_i - z_k \sum_{i=1}^{n} z_i + n z_j z_k\right) y_k \,. \end{aligned} \tag{A-3}$$

und speziell für den ersten Wert, für den der Regressorwert gegen Unendlich strebt:

$$\begin{aligned} \hat{y}_1 &= \frac{1}{n\sum z_i^2 - (\sum z_i)^2} \sum_{k=1}^{n}\left[\sum_{i=1}^{n} z_i^2 - z_1 \sum_{i=1}^{n} z_i - z_k \sum_{i=1}^{n} z_i + n z_1 z_k\right] y_k \\ &= \frac{1}{n\sum z_i^2 - (\sum z_i)^2}\left[\left(\sum_{i=1}^{n} z_i^2 - z_1 \sum_{i=1}^{n} z_i\right)\sum_{k=1}^{n} y_k \;-\; \left(\sum_{i=1}^{n} z_i - n z_1\right)\sum_{k=1}^{n} z_k y_k\right] \\ &= \frac{\left(\sum_{i=1}^{n} z_i^2 - z_1 \sum_{i=1}^{n} z_i\right)\sum_{k=1}^{n} y_k}{n\sum z_i^2 - (\sum z_i)^2} \;-\; \frac{\left(\sum_{i=1}^{n} z_i - n z_1\right)\sum_{k=1}^{n} z_k y_k}{n\sum z_i^2 - (\sum z_i)^2} \\ &= \frac{\left(z_1^2 + \sum_{i>1} z_i^2 - z_1(z_1 + \sum_{i>1} z_i)\right)\sum_{k=1}^{n} y_k}{n(z_1^2 + \sum_{i>1} z_i^2) - (z_1 + \sum_{i>1} z_i)^2)} \;-\; \frac{\left(z_1 + \sum_{i>1} z_i - n z_1\right)(z_1 y_1 + \sum_{k>1} z_k y_k)}{n(z_1^2 + \sum_{i>1} z_i^2) - (z_1 + \sum_{i>1} z_i)^2)} \\ &= A \quad - \quad B \end{aligned}$$

[60] Es gilt

$$\boldsymbol{X}'\boldsymbol{X} = \begin{bmatrix} n & \sum z_i \\ \sum z_i & \sum z_i^2 \end{bmatrix} \quad , \quad \boldsymbol{X}'\boldsymbol{y} = \begin{bmatrix} \sum y_i \\ \sum z_i y_i \end{bmatrix}$$

$$(\boldsymbol{X}'\boldsymbol{X})^{-1} = \frac{1}{n\sum z_i^2 - (\sum z_i)^2}\begin{bmatrix} \sum z_i^2 & -\sum z_i \\ -\sum z_i & n \end{bmatrix} = \frac{1}{n^2 s_z^2}\begin{bmatrix} \sum z_i^2 & -\sum z_i \\ -\sum z_i & n \end{bmatrix} \quad .$$

wobei wir hier für den Nenner den ursprünglichen Ausdruck (siehe Fußnote 60) verwenden. Um den Grenzwert für $z_1 \to \infty$ bezüglich $\hat{y}_1$ zu bestimmen, erweitern wir beide Brüche um den Faktor $1/z_1^2$ und erhalten

$$A = \frac{\left[\left(1 + \frac{\sum_{i>1} z_i^2}{z_1^2}\right) - \left(1 + \frac{\sum_{i>1} z_i}{z_1}\right)\right] \sum_{k=1}^{n} y_k}{n\left(1 + \frac{\sum_{i>1} z_i^2}{z_1^2}\right) - \left(1 + \frac{\sum_{i>1} z_i}{z_1}\right)^2}$$

sowie

$$B = \frac{\left(1 + \frac{\sum_{i>1} z_i}{z_1} - n\right)\left(y_1 + \frac{\sum_{k>1} z_k y_k}{z_1}\right)}{n\left(1 + \frac{\sum_{i>1} z_i^2}{z_1^2}\right) - \left(1 + \frac{\sum_{i>1} z_i}{z_1}\right)^2}$$

und daraus

$$\lim_{z_1 \to \infty} \hat{y}_1 \;=\; \lim_{z_1 \to \infty} (A - B) \;=\; \frac{0}{n-1} - \frac{(1-n)\, y_1}{n-1} \;=\; y_1 \qquad . \tag{A-4}$$

Demnach läßt sich durch ein genügend großes z_1 der Wert von y_1 durch das geschätzte $\hat{y}_1$ beliebig genau annähern.

A.2.1 Berücksichtigung zusätzlicher Regressoren

Es soll nun der Fall betrachtet werden, dass neben dem künstlichen Außreißer für die erste Beobachtung auch weitere Regressorvariable berücksichtigt werden. Dabei nutzen wir aus, dass das lineare Modell auch in **zerlegter Form** geschrieben werden kann:[61]

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u} = \begin{pmatrix} \boldsymbol{X}_1 & \boldsymbol{X}_2 \end{pmatrix} \begin{pmatrix} \boldsymbol{\beta}_2 \\ \boldsymbol{\beta}_2 \end{pmatrix} + \boldsymbol{u} = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{X}_2\boldsymbol{\beta}_2 + \boldsymbol{u}\,.$$

Unter Verwendung von

$$\boldsymbol{M}_1 = \boldsymbol{I} - \boldsymbol{X}_1(\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1' \quad \text{und} \quad \boldsymbol{M}_2 = \boldsymbol{I} - \boldsymbol{X}_2(\boldsymbol{X}_2'\boldsymbol{X}_2)^{-1}\boldsymbol{X}_2'$$

lassen sich die **Teilschätzungen** für $\boldsymbol{\beta}_1$ bzw. $\boldsymbol{\beta}_2$ wie folgt schreiben:

$$\begin{aligned} \hat{\boldsymbol{\beta}}_1 &= (\boldsymbol{X}_1'\boldsymbol{M}_2\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1'\boldsymbol{M}_2\boldsymbol{y} \\ \hat{\boldsymbol{\beta}}_2 &= (\boldsymbol{X}_2'\boldsymbol{M}_1\boldsymbol{X}_2)^{-1}\boldsymbol{X}_2'\boldsymbol{M}_1\boldsymbol{y}\,. \end{aligned}$$

Eine alternative Form ist:

$$\begin{aligned} \hat{\boldsymbol{\beta}}_1 &= (\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1'\left(\boldsymbol{y} - \boldsymbol{X}_2\hat{\boldsymbol{\beta}}_2\right) \\ \hat{\boldsymbol{\beta}}_2 &= (\boldsymbol{X}_2'\boldsymbol{X}_2)^{-1}\boldsymbol{X}_2'\left(\boldsymbol{y} - \boldsymbol{X}_1\hat{\boldsymbol{\beta}}_1\right)\,. \end{aligned}$$

Wir teilen das Modell in der Art auf, dass in $\mathbf{X}_1$ alle Regressoren (samt dem Einsvektor) enthalten ist und $\mathbf{X}_2$ nur den oben bereits betrachteten Regressor z (mit künstlichem Ausreißer z_1) enthält, d.h. diese Matrix hat die Form

$$\mathbf{X}_2 \quad = \quad \mathbf{z} \quad = \quad \begin{pmatrix} z_1 \\ \mathbf{z}_2 \end{pmatrix} \qquad ,$$

[61] Siehe beispielsweise Schönfeld (1969).

wobei z_1 gegen Unendlich strebt und $\mathbf{z}_2$ der $(n-1)$-dimensionale Vektor ist, der die restlichen Beobachtungswerte von $\mathbf{z}$ enthält.

Im folgender benötigen wir außerdem die folgende Zerlegung der Regressormatrix $\mathbf{X}_1$:

$$\mathbf{X}_1 \quad = \quad \begin{pmatrix} \mathbf{x}'_{11} \\ \mathbf{X}_{12} \end{pmatrix} \qquad ,$$

d.h. die erste Zeile dieser Matrix, die mit $\mathbf{x}'_{11}$ bezeichnet ist, wird separat betrachtet. Entsprechend wird der Vektor $\mathbf{y}$ zerlegt:

$$\mathbf{y} \quad = \quad \begin{pmatrix} y_1 \\ \mathbf{y}_2 \end{pmatrix} \qquad .$$

Nun schreiben wir den Vektor der Prognosewerte unter Verwendung der Ergebnisse für das "zerlegte" Modell wie folgt:

$$\begin{aligned} \hat{\mathbf{y}} &= \mathbf{X}_1 \hat{\boldsymbol{\beta}}_1 + \mathbf{X}_2 \hat{\boldsymbol{\beta}}2 \\ &= \mathbf{X}_1 \hat{\boldsymbol{\beta}}_1 + \mathbf{X}_2 (\boldsymbol{X}'_2 \boldsymbol{X}_2)^{-1} \boldsymbol{X}'_2 \left(\boldsymbol{y} - \boldsymbol{X}_1 \hat{\boldsymbol{\beta}}_1 \right) \end{aligned} \tag{A-5}$$

Wegen

$$\mathbf{X}_2 (\boldsymbol{X}'_2 \boldsymbol{X}_2)^{-1} \boldsymbol{X}'_2 \quad = \quad \frac{1}{z_1^2 + \sum_{i>1} z_i^2} \begin{pmatrix} z_1^2 & z_1 \mathbf{z}'_2 \\ z_1 \mathbf{z}_2 & \mathbf{z}_2 \mathbf{z}'_2 \end{pmatrix}$$

ergibt sich

$$\hat{\mathbf{y}} \quad = \quad \mathbf{X}_1 \hat{\boldsymbol{\beta}}_1 + \frac{1}{z_1^2 + \sum_{i>1} z_i^2} \begin{pmatrix} z_1^2 & z_1 \mathbf{z}'_2 \\ z_1 \mathbf{z}_2 & \mathbf{z}_2 \mathbf{z}'_2 \end{pmatrix} \left(\boldsymbol{y} - \boldsymbol{X}_1 \hat{\boldsymbol{\beta}}_1 \right)$$

Daraus erhalten wir (unter Beachtung der obigen Zerlegung für die Teilmatrix $\mathbf{X}1$ sowie den Vektor $\mathbf{y}$)

$$\hat{\mathbf{y}} \quad = \quad \begin{pmatrix} \mathbf{x}'_{11} \hat{\boldsymbol{\beta}}_1 \\ \mathbf{X}_{12} \hat{\boldsymbol{\beta}}_1 \end{pmatrix} + \frac{1}{z_1^2 + \sum_{i>1} z_i^2} \begin{pmatrix} z_1^2 & z_1 \mathbf{z}'_2 \end{pmatrix} \begin{pmatrix} y_1 - \mathbf{x}'_{11} \hat{\boldsymbol{\beta}}_1 \\ (\mathbf{y}_2 - \mathbf{X}_{12} \hat{\boldsymbol{\beta}}_1 \end{pmatrix} \tag{A-6}$$

und insbesondere für das erste Element

$$\begin{aligned} \hat{y}_1 &= \mathbf{x}'_{11} \hat{\boldsymbol{\beta}}_1 + \frac{1}{z_1^2 + \sum_{i>1} z_i^2} \left(z_1^2 (y_1 - \mathbf{x}'_{11} \hat{\boldsymbol{\beta}}_1) + z_1 \mathbf{z}'_2 (\mathbf{y}_2 - \mathbf{X}_{12} \hat{\boldsymbol{\beta}}_1) \right) \\ &= \mathbf{x}'_{11} \hat{\boldsymbol{\beta}}_1 + \frac{1}{1 + \frac{\sum_{i>1} z_i^2}{z_1^2}} \left((y_1 - \mathbf{x}'_{11} \hat{\boldsymbol{\beta}}_1) + \frac{1}{z_1} \mathbf{z}'_2 (\mathbf{y}_2 - \mathbf{X}_{12} \hat{\boldsymbol{\beta}}_1) \right) \end{aligned} \tag{A-7}$$

Für unendlich großes z_1 ergibt sich demnach entsprechend (A-4) auch im Fall einer multiplen Regression

$$\lim_{z_1 \to \infty} \hat{y}_1 = y_1 \qquad . \tag{A-8}$$

A.2.2 Einsatz von mehreren künstlichen Ausreißern

Falls zwei verschiedene Merkmale für die Konstruktion von künstlichen Ausreißern verwendet werden, hat die $(n \times 2)-$Matrix $\mathbf{X}_2$ die Form

$$\mathbf{X}_2 \quad = \quad \mathbf{Z} \quad = \quad \begin{pmatrix} z_{11} & z_{12} \\ \mathbf{z}_{21} & \mathbf{z}_{22} \end{pmatrix} \qquad ,$$

wobei z_{11} und z_{12} gegen Unendlich streben und $\mathbf{z}_{21}$ und $\mathbf{z}_{22}$ $(n-1)$-dimensionale Vektoren sind, die die restlichen Beobachtungswerte von $\mathbf{z}_1$ und $\mathbf{z}_2$ enthalten.

Wir erhalten weiter

$$
\begin{aligned}
&\mathbf{X}_2\,(\boldsymbol{X}_2'\boldsymbol{X}_2)^{-1}\boldsymbol{X}_2' \\
&= \quad \mathbf{Z}\,(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}' \\
&= \quad \begin{pmatrix} z_{11} & z_{12} \\ \mathbf{z}_{21} & \mathbf{z}_{22} \end{pmatrix} \begin{pmatrix} z_{11}^2+\mathbf{z}_{21}'\mathbf{z}_{21} & z_{11}z_{12}+\mathbf{z}_{21}'\mathbf{z}_{22} \\ z_{11}z_{12}+\mathbf{z}_{21}'\mathbf{z}_{22} & z_{12}^2+\mathbf{z}_{22}'\mathbf{z}_{22} \end{pmatrix}^{-1} \begin{pmatrix} z_{11} & z_{12} \\ \mathbf{z}_{21} & \mathbf{z}_{22} \end{pmatrix}' \\
&= \tfrac{1}{c} \begin{pmatrix} z_{11} & z_{12} \\ \mathbf{z}_{21} & \mathbf{z}_{22} \end{pmatrix} \begin{pmatrix} z_{12}^2+\mathbf{z}_{22}'\mathbf{z}_{22} & -(z_{11}z_{12}+\mathbf{z}_{21}'\mathbf{z}_{22}) \\ -(z_{11}z_{12}+\mathbf{z}_{21}'\mathbf{z}_{22}) & z_{11}^2+\mathbf{z}_{21}'\mathbf{z}_{21} \end{pmatrix} \begin{pmatrix} z_{11} & z_{12} \\ \mathbf{z}_{21} & \mathbf{z}_{22} \end{pmatrix}' \\
&= \tfrac{1}{c} \begin{pmatrix} z_{11}(z_{12}^2+\mathbf{z}_{22}'\mathbf{z}_{22})-z_{12}(z_{11}z_{12}+\mathbf{z}_{21}'\mathbf{z}_{22}) & -z_{11}(z_{11}z_{12}+\mathbf{z}_{21}'\mathbf{z}_{22})+z_{12}(z_{11}^2+\mathbf{z}_{21}'\mathbf{z}_{21}) \\ (z_{12}^2+\mathbf{z}_{22}'\mathbf{z}_{22})\mathbf{z}_{21}-(z_{11}z_{12}+\mathbf{z}_{21}'\mathbf{z}_{22})\mathbf{z}_{22} & -(z_{11}z_{12}+\mathbf{z}_{21}'\mathbf{z}_{22})\mathbf{z}_{21}+(z_{11}^2+\mathbf{z}_{21}'\mathbf{z}_{21})\mathbf{z}_{22} \end{pmatrix} \begin{pmatrix} z_{11} & \mathbf{z}_{21}' \\ z_{12} & \mathbf{z}_{22}' \end{pmatrix} \\
&= \tfrac{1}{c} \left(\begin{array}{cc|c} z_{11}^2(z_{12}^2+\mathbf{z}_{22}'\mathbf{z}_{22})-z_{11}z_{12}(z_{11}z_{12}+\mathbf{z}_{21}'\mathbf{z}_{22}) & & \left[z_{11}(z_{12}^2+\mathbf{z}_{22}'\mathbf{z}_{22})-z_{12}(z_{11}z_{12}+\mathbf{z}_{21}'\mathbf{z}_{22})\right]\mathbf{z}_{21}' \\ -z_{11}z_{12}(z_{11}z_{12}+\mathbf{z}_{21}'\mathbf{z}_{22})+z_{12}^2(z_{11}^2+\mathbf{z}_{21}'\mathbf{z}_{21}) & & \left[-z_{11}(z_{11}z_{12}+\mathbf{z}_{21}'\mathbf{z}_{22})+z_{12}(z_{11}^2+\mathbf{z}_{21}'\mathbf{z}_{21})\right]\mathbf{z}_{22}' \\ \hline z_{11}(z_{12}^2+\mathbf{z}_{22}'\mathbf{z}_{22})\mathbf{z}_{21}-z_{11}(z_{11}z_{12}+\mathbf{z}_{21}'\mathbf{z}_{22})\mathbf{z}_{22} & & (z_{12}^2+\mathbf{z}_{22}'\mathbf{z}_{22})\mathbf{z}_{21}\mathbf{z}_{21}'-(z_{11}z_{12}+\mathbf{z}_{21}'\mathbf{z}_{22})\mathbf{z}_{22}\mathbf{z}_{21}' \\ -z_{12}(z_{11}z_{12}+\mathbf{z}_{21}'\mathbf{z}_{22})\mathbf{z}_{21}+z_{12}(z_{11}^2+\mathbf{z}_{21}'\mathbf{z}_{21})\mathbf{z}_{22} & & -(z_{11}z_{12}+\mathbf{z}_{21}'\mathbf{z}_{22})\mathbf{z}_{22}\mathbf{z}_{22}'+(z_{11}^2+\mathbf{z}_{21}'\mathbf{z}_{21})\mathbf{z}_{22}\mathbf{z}_{22}' \end{array}\right)
\end{aligned}
$$

mit

$$
\begin{aligned}
c \equiv det(\mathbf{Z}'\mathbf{Z}) &= \left(z_{11}^2+\mathbf{z}_{21}'\mathbf{z}_{21}\right)\left(z_{12}^2+\mathbf{z}_{22}'\mathbf{z}_{22}\right) - \left(z_{11}z_{12}+\mathbf{z}_{21}'\mathbf{z}_{22}\right)^2 \\
&= (z_{11}z_{12})^2\left[\left(1+\tfrac{\mathbf{z}_{21}'\mathbf{z}_{21}}{z_{11}^2}\right)\left(1+\tfrac{\mathbf{z}_{22}'\mathbf{z}_{22}}{z_{12}^2}\right) - \left(1+\tfrac{\mathbf{z}_{21}'\mathbf{z}_{22}}{z_{11}z_{12}}\right)^2\right] \\
&= (z_{11}z_{12})^2 \quad d
\end{aligned}
$$

Den für den Prognosewert $\hat{y}_1$ relevanten "oberen" Teil der Matrix $\mathbf{X}_2\,(\boldsymbol{X}_2'\boldsymbol{X}_2)^{-1}\boldsymbol{X}_2'$ formen wir wie folgt um:

$$
\begin{aligned}
&\tfrac{1}{c}\left(\begin{array}{cc|c} z_{11}^2(z_{12}^2+\mathbf{z}_{22}'\mathbf{z}_{22})-z_{11}z_{12}(z_{11}z_{12}+\mathbf{z}_{21}'\mathbf{z}_{22}) & & \left[z_{11}(z_{12}^2+\mathbf{z}_{22}'\mathbf{z}_{22})-z_{12}(z_{11}z_{12}+\mathbf{z}_{21}'\mathbf{z}_{22})\right]\mathbf{z}_{21}' \\ -z_{11}z_{12}(z_{11}z_{12}+\mathbf{z}_{21}'\mathbf{z}_{22})+z_{12}^2(z_{11}^2+\mathbf{z}_{21}'\mathbf{z}_{21}) & & \left[-z_{11}(z_{11}z_{12}+\mathbf{z}_{21}'\mathbf{z}_{22})+z_{12}(z_{11}^2+\mathbf{z}_{21}'\mathbf{z}_{21})\right]\mathbf{z}_{22}' \end{array}\right) \\
&= \tfrac{1}{d}\left(\begin{array}{cc|c} (1+\tfrac{\mathbf{z}_{22}'\mathbf{z}_{22}}{z_{12}^2})-2(1+\tfrac{\mathbf{z}_{21}'\mathbf{z}_{22}}{z_{11}z_{12}}) & & \tfrac{1}{z_{11}}\left[(1+\tfrac{\mathbf{z}_{22}'\mathbf{z}_{22}}{z_{12}^2})-(1+\tfrac{\mathbf{z}_{21}'\mathbf{z}_{22}}{z_{11}z_{12}}\right]\mathbf{z}_{21}' \\ +(1+\tfrac{\mathbf{z}_{21}'\mathbf{z}_{21}}{z_{11}^2}) & & +\tfrac{1}{z_{12}}\left[(1+\tfrac{\mathbf{z}_{21}'\mathbf{z}_{21}}{z_{12}^2})-(1+\tfrac{\mathbf{z}_{21}'\mathbf{z}_{22}}{z_{11}z_{12}})\right]\mathbf{z}_{22}' \end{array}\right)
\end{aligned}
$$

Falls sich für $z_{11}\to\infty$ und $z_{12}\to\infty$ für diese Matrix ein Ausdruck von der Form

$$\left(\begin{array}{c|c} 1 & \mathbf{0} \end{array}\right)$$

ergeben würde, würde sich

$$\hat{y}_1 \quad = \quad \mathbf{x}_{11}'\hat{\boldsymbol{\beta}}_1 + \left(\begin{array}{c|c} 1 & \mathbf{0} \end{array}\right)\begin{pmatrix} y_1 - \mathbf{x}_{11}'\hat{\boldsymbol{\beta}}_1 \\ (\mathbf{y}_2 - \mathbf{X}_{12}\hat{\boldsymbol{\beta}}_1 \end{pmatrix} \quad = \quad y_1$$

ergeben. Dafür untersuchen wir zunächst das linke Element in der Matrix und erhalten

den Ausdruck

$$\frac{(1+\frac{\mathbf{z}'_{22}\mathbf{z}_{22}}{z_{12}^2})-2(1+\frac{\mathbf{z}'_{21}\mathbf{z}_{22}}{z_{11}z_{12}})+(1+\frac{\mathbf{z}'_{21}\mathbf{z}_{21}}{z_{11}^2})}{\left(1+\frac{\mathbf{z}'_{21}\mathbf{z}_{21}}{z_{11}^2}\right)\left(1+\frac{\mathbf{z}'_{22}\mathbf{z}_{22}}{z_{12}^2}\right)-\left(1+\frac{\mathbf{z}'_{21}\mathbf{z}_{22}}{z_{11}z_{12}}\right)^2}$$

$$= \frac{\frac{\mathbf{z}'_{21}\mathbf{z}_{21}}{z_{11}^2}+\frac{\mathbf{z}'_{22}\mathbf{z}_{22}}{z_{12}^2}-2\frac{\mathbf{z}'_{21}\mathbf{z}_{22}}{z_{11}z_{12}}}{\frac{\mathbf{z}'_{21}\mathbf{z}_{21}}{z_{11}^2}+\frac{\mathbf{z}'_{22}\mathbf{z}_{22}}{z_{12}^2}-2\frac{\mathbf{z}'_{21}\mathbf{z}_{22}}{z_{11}z_{12}}+\frac{1}{z_{11}^2z_{12}^2}(\mathbf{z}'_{21}\mathbf{z}_{21}\,\mathbf{z}'_{22}\mathbf{z}_{22}-(\mathbf{z}'_{21}\mathbf{z}_{22})^2)} \quad ,$$

$$= \frac{1}{1+\frac{1}{z_{11}^2z_{12}^2}(\mathbf{z}'_{21}\mathbf{z}_{21}\,\mathbf{z}'_{22}\mathbf{z}_{22}-(\mathbf{z}'_{21}\mathbf{z}_{22})^2)} \quad ,$$

der offensichtlich gegen 1 strebt. Ferner ist klar, dass der Vektorausdruck "rechts" in der Matrix $\mathbf{X}_2\,(\boldsymbol{X}'_2\boldsymbol{X}_2)^{-1}\boldsymbol{X}'_2$, der durch

$$\frac{1}{z_{11}}\left[(1+\frac{\mathbf{z}'_{22}\mathbf{z}_{22}}{z_{12}^2})-(1+\frac{\mathbf{z}'_{21}\mathbf{z}_{22}}{z_{11}z_{12}}\right]\mathbf{z}'_{21}+\frac{1}{z_{12}}\left[(1+\frac{\mathbf{z}'_{21}\mathbf{z}_{21}}{z_{12}^2})-(1+\frac{\mathbf{z}'_{21}\mathbf{z}_{22}}{z_{11}z_{12}})\right]\mathbf{z}'_{22}$$

gegeben ist, für $z_{11}\to\infty$ und $z_{12}\to\infty$ gegen einen Nullvektor strebt. Demach gilt auch für einen Vektor von künstlichen Ausreißern, hier speziell für einen 2-dimensionalen Vektor

$$\lim_{\substack{z_{11}\to\infty\\ z_{12}\to\infty}} \hat{y}_1 = y_1 \quad . \tag{A-9}$$

A.2.3 Einfluß von Kovariablen bei nicht eindeutiger Identifikation

Es soll nun noch der Fall untersucht werden, dass durch den Wert x_m und den daraus resultierenden Wert von z (siehe Formel (3-3)) mehr als ein Unternehmen "gepickt" wird. Im Gegensatz zur Situation bei strategischen Dummyvariablen (siehe unten Abschnitt A.3) dürfte dies eher selten vorkommen.

Formal bedeutet dies, dass die im vorigen Unterabschnitt eingeführte Matrix folgende Gestalt hat:

$$\mathbf{X}_2 = \begin{pmatrix} z_1\,\boldsymbol{\iota}_q \\ \mathbf{z}_2 \end{pmatrix} \quad .$$

Dabei gehen wir davon aus, dass q Unternehmen gepickt werden (und diese die ersten q Beobachtungen darstellen). $\boldsymbol{\iota}_q$ ist ein Einsvektor von der Dimension q und $\mathbf{0}$ steht für einen Nullvektor mit $n-q$ Elementen. Weiter erhalten wir

$$\mathbf{X}_2\,(\boldsymbol{X}'_2\boldsymbol{X}_2)^{-1}\boldsymbol{X}'_2 = \frac{1}{q\,z_1^2+\sum_{i>q}z_i^2}\begin{pmatrix} z_1^2\boldsymbol{\iota}_q\boldsymbol{\iota}'_q & z_1\,\boldsymbol{\iota}_q\,\mathbf{z}'_2 \\ z_1\,\mathbf{z}_2\,\boldsymbol{\iota}'_q & \mathbf{z}_2\mathbf{z}'_2 \end{pmatrix}$$

und

$$\mathbf{X}_2\,(\boldsymbol{X}'_2\boldsymbol{X}_2)^{-1}\boldsymbol{X}'_2\left(\boldsymbol{y}-\boldsymbol{X}_1\hat{\boldsymbol{\beta}}_1\right) = \frac{1}{q\,z_1^2+\sum_{i>q}z_i^2}\begin{pmatrix} z_1^2\boldsymbol{\iota}_q\boldsymbol{\iota}'_q & z_1\,\boldsymbol{\iota}_q\,\mathbf{z}'_2 \\ z_1\,\mathbf{z}_2\,\boldsymbol{\iota}'_q & \mathbf{z}_2\mathbf{z}'_2 \end{pmatrix}\left(\boldsymbol{y}-\boldsymbol{X}_1\hat{\boldsymbol{\beta}}_1\right)$$

und damit für den Vektor der Vorhersagewerte

$$\hat{\mathbf{y}} = \boldsymbol{X}_1\hat{\boldsymbol{\beta}}_1+\frac{1}{q\,z_1^2+\sum_{i>q}z_i^2}\begin{pmatrix} z_1^2\boldsymbol{\iota}_q\boldsymbol{\iota}'_q & z_1\,\boldsymbol{\iota}_q\,\mathbf{z}'_2 \\ z_1\,\mathbf{z}_2\,\boldsymbol{\iota}'_q & \mathbf{z}_2\mathbf{z}'_2 \end{pmatrix}\left(\boldsymbol{y}-\boldsymbol{X}_1\hat{\boldsymbol{\beta}}_1\right)$$

bzw. in zerlegter Form (und offensichtlicher Symbolik)

$$\begin{pmatrix} \hat{\mathbf{y}}_q \\ \hat{\mathbf{y}}_{n-q} \end{pmatrix}$$

$$= \begin{pmatrix} \mathbf{X}_{1q} \\ \mathbf{X}_{1,n-q} \end{pmatrix} \hat{\boldsymbol{\beta}}_1 + \frac{1}{q\, z_1^2 + \sum_{i>q} z_i^2} \begin{pmatrix} z_1^2 \boldsymbol{\iota}_q \boldsymbol{\iota}_q' & z_1\, \boldsymbol{\iota}_q\, \mathbf{z}_2' \\ z_1\, \mathbf{z}_2\, \boldsymbol{\iota}_q' & \mathbf{z}_2 \mathbf{z}_2' \end{pmatrix} \left\{ \begin{pmatrix} \mathbf{y}_q \\ \mathbf{y}_{n-q} \end{pmatrix} - \begin{pmatrix} \mathbf{X}_{1q} \\ \mathbf{X}_{1,n-q} \end{pmatrix} \hat{\boldsymbol{\beta}}_1 \right\}$$

Für die ersten q Elemente des Prognosevektors $\hat{\mathbf{y}}$ erhalten wir demnach

$$\begin{aligned} \hat{\mathbf{y}}_q &= \mathbf{X}_{1q}\, \hat{\boldsymbol{\beta}}_1 + \frac{1}{q\, z_1^2 + \sum_{i>q} z_i^2} \begin{pmatrix} z_1^2 \boldsymbol{\iota}_q \boldsymbol{\iota}_q' & z_1\, \boldsymbol{\iota}_q\, \mathbf{z}_2' \end{pmatrix} \left\{ \begin{pmatrix} \mathbf{y}_q \\ \mathbf{y}_{n-q} \end{pmatrix} - \begin{pmatrix} \mathbf{X}_{1q} \\ \mathbf{X}_{1,n-q} \end{pmatrix} \hat{\boldsymbol{\beta}}_1 \right\} \\ &= \mathbf{X}_{1q}\, \hat{\boldsymbol{\beta}}_1 \\ &\quad + \frac{1}{q\, z_1^2 + \sum_{i>q} z_i^2}\, z_1^2\, \boldsymbol{\iota}_q \boldsymbol{\iota}_q' \left(\mathbf{y}_q - \mathbf{X}_{1q}\, \hat{\boldsymbol{\beta}}_1\right) \\ &\quad + \frac{1}{q\, z_1^2 + \sum_{i>q} z_i^2}\, z_1\, \boldsymbol{\iota}_q\, \mathbf{z}_2' \left(\mathbf{y}_{n-q} - \mathbf{X}_{1,n-q}\, \hat{\boldsymbol{\beta}}_1\right) \\ &= \mathbf{X}_{1q}\, \hat{\boldsymbol{\beta}}_1 \\ &\quad + \frac{1}{q + \frac{\sum_{i>q} z_i^2}{z_1^2}}\, \boldsymbol{\iota}_q \boldsymbol{\iota}_q' \left(\mathbf{y}_q - \mathbf{X}_{1q}\, \hat{\boldsymbol{\beta}}_1\right) \\ &\quad + \frac{1}{q + \frac{\sum_{i>q} z_i^2}{z_1^2}}\, \frac{1}{z_1}\, \boldsymbol{\iota}_q\, \mathbf{z}_2' \left(\mathbf{y}_{n-q} - \mathbf{X}_{1,n-q}\, \hat{\boldsymbol{\beta}}_1\right) \end{aligned} \tag{A-10}$$

Demnach erhalten wir für unendlich großes z_1

$$\lim_{z_1 \to \infty} \hat{\mathbf{y}}_q = \frac{1}{q}\, \boldsymbol{\iota}_q \boldsymbol{\iota}_q'\, \mathbf{y}_q + \mathbf{X}_{1q}\, \hat{\boldsymbol{\beta}}_1 - \frac{1}{q}\, \boldsymbol{\iota}_q \boldsymbol{\iota}_q'\, \mathbf{X}_{1q}\, \hat{\boldsymbol{\beta}}_1 \tag{A-11}$$

oder auch

$$\hat{y}_i = \overline{y}_q + \left(1, \quad x_{i2} - \overline{x}_q^{(2)}, \quad \ldots, \quad x_{iK} - \overline{x}_q^{(K)} \right) \hat{\boldsymbol{\beta}}_1 \,, i = 1, 2, \ldots, q \quad . \tag{A-12}$$

Dabei verwenden wir

$$\overline{y}_q = \frac{1}{q} \sum_{i=1}^{q} y_i \quad \text{und} \quad \overline{x}_q^{(k)} = \frac{1}{q} \sum_{i=1}^{q} x_{ik} \,,\, k = 2, \ldots, K \,.$$

Die Formeln (A-11) und (A-12) zeigt folgendes:

- Falls $q = 1$ ist und damit eine eindeutige Identifikation vorliegt, reduziert sich dies Ergebnis auf die Formel (A-8), denn dann ist $\overline{y}_q = y_1$ und für alle Regressoren gilt $\overline{x}_q^{(k)} = x_{1k}$.
- Sofern nur der künstliche Ausreißer in einer Einfachregression verwendet wird, gilt in der Tat

  $$\lim_{z_1 \to \infty} \hat{y}_i = \overline{y}_q \,, \quad i = 1, \ldots, q \,,$$

  für alle q Unternehmen.

- Im allgemeinen wird jedoch im Fall der nicht-eindeutigen Identifikation keine Aussage über Richtung und Ausmaß der Abweichung zwischen $\hat{y}_i$ und y_i , $i = 1, 2, \ldots, q$, möglich sein.

A.3 Strategische Dummyvariable

A.3.1 Einfachregression

Im Fall einer Dummyvariablen gemäß Abschnitt 3.2.1 kann man eine Einfachregression durchführen, in der eine der beiden folgenden Dummyvariablen als einziger Regressor auftaucht:

$$D_i = \begin{cases} 1 & \text{falls } x = x_1 \\ 0 & \text{sonst} \end{cases}$$

oder

$$D_i = \begin{cases} 1 & \text{falls } x - \delta < x_1 < x + \delta \\ 0 & \text{sonst} \end{cases}$$

Natürlich ist diese Spezifikation - ohne weitere Regressoren - besonders offensichtlich und wird vermutlich bei der Outputkontrolle sofort entdeckt.[62] Wir erhalten in diesem Fall die Regressormatrix

$$\mathbf{X} = \begin{pmatrix} \boldsymbol{\iota} & \mathbf{e}_1 \end{pmatrix} ,$$

wobei $\mathbf{e}_1$ ein n-dimensionaler Vektor ist, der an der ersten Stelle eine 1 und sonst nur Nullen aufweist. Diese Konstellation tritt natürlich nur dann ein, wenn kein anderer Wert der Regressorvariablen exakt gleich x_1 ist bzw. "zu nahe" bei x_1 liegt!! In diesem Fall ergibt sich

$$\boldsymbol{X}'\boldsymbol{X} = \begin{pmatrix} n & 1 \\ 1 & 1 \end{pmatrix}$$

sowie

$$\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} = \frac{1}{(n-1)} \begin{pmatrix} 1 & -1 \\ -1 & n \end{pmatrix}$$

und damit

$$\mathbf{H} = \frac{1}{n-1} \begin{pmatrix} n-1 & \mathbf{0}' \\ \mathbf{0} & \boldsymbol{\iota}_{n-1}\boldsymbol{\iota}'_{n-1} \end{pmatrix} ,$$

wobei $\mathbf{0}$ der $(n-1)$-dimensionale Nullvektor und $\boldsymbol{\iota}_{n-1}$ der $(n-1)$-dimensionale Einsvektor sind. Man beachte , dass demnach $h_{11} = 1$ und $h_{1j} = 0$, $j > 1$, gilt. Damit ergibt sich für den geschätzten Wert von y_1 sofort

$$\hat{y}_1 = \sum_{k=1}^{n} h_{1k} y_k = \frac{1}{n-1} \left((n-1)\, y_1 + \sum_{k>1} 0 \cdot y_k \right) = y_1 .$$

A.3.2 Berücksichtigung zusätzlicher Regressoren

Es soll nun genau wie bei der Behandlung eines künstlichen Ausreißers der Fall betrachtet werden, dass neben dem Dummy für die erste Beobachtung auch weitere Regressorvariable berücksichtigt werden. Dabei nutzen wir wieder aus, dass das lineare Modell auch in **zerlegter Form** geschrieben werden kann. Siehe den obigen Abschnitt Abschnitt A.2.1

[62] Für den allgemeineren Fall verweisen wir auf Unterabschnitt A.3.2.

Wir teilen in diesem Fall das Modell in der Art auf, dass in $\mathbf{X}_1$ alle Regressoren (samt dem Einsvektor) enthalten ist und $\mathbf{X}_2$ nur den oben definierten Dummy D_i enthält, d.h. diese Matrix hat die Form

$$\mathbf{X}_2 \quad = \quad \mathbf{e}_1 \qquad .$$

Nun schreiben wir den Vektor der Prognosewerte unter Verwendung der Ergebnisse für das "zerlegte" Modell wie folgt:

$$\begin{aligned} \hat{\mathbf{y}} &= \mathbf{X}_1 \hat{\boldsymbol{\beta}}_1 + \mathbf{X}_2 \hat{\boldsymbol{\beta}} 2 \\ &= \mathbf{X}_1 \hat{\boldsymbol{\beta}}_1 + \mathbf{X}_2 (\boldsymbol{X}_2' \boldsymbol{X}_2)^{-1} \boldsymbol{X}_2' \left(\boldsymbol{y} - \boldsymbol{X}_1 \hat{\boldsymbol{\beta}}_1 \right) \end{aligned} \tag{A-13}$$

Wegen

$$\mathbf{X}_2 (\boldsymbol{X}_2' \boldsymbol{X}_2)^{-1} \boldsymbol{X}_2' \quad = \quad \begin{pmatrix} 1 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}$$

und

$$\begin{aligned} \mathbf{X}_2 (\boldsymbol{X}_2' \boldsymbol{X}_2)^{-1} \boldsymbol{X}_2' \left(\boldsymbol{y} - \boldsymbol{X}_1 \hat{\boldsymbol{\beta}}_1 \right) &= \begin{pmatrix} 1 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \left(\boldsymbol{y} - \boldsymbol{X}_1 \hat{\boldsymbol{\beta}}_1 \right) \\ &= \begin{pmatrix} y_1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} - \begin{pmatrix} 1 & x_{12} & \dots & x_{1K} \\ 0 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 \end{pmatrix} \hat{\boldsymbol{\beta}}_1 \qquad , \end{aligned}$$

erhalten wir für den Vektor $\hat{\mathbf{y}}$ in (A-13):

$$\hat{\mathbf{y}} \quad = \quad \mathbf{X}_1 \hat{\boldsymbol{\beta}}_1 + \begin{pmatrix} y_1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} - \begin{pmatrix} 1 & x_{12} & \dots & x_{1K} \\ 0 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 \end{pmatrix} \hat{\boldsymbol{\beta}}_1 \tag{A-14}$$

und insbesondere für das erste Element

$$\hat{y}_1 \quad = \quad \begin{pmatrix} 1 & x_{12} & \dots & x_{1K} \end{pmatrix} \hat{\boldsymbol{\beta}}_1 + y_1 - \begin{pmatrix} 1 & x_{12} & \dots & x_{1K} \end{pmatrix} \hat{\boldsymbol{\beta}}_1 \quad = \quad y_1 \,. \tag{A-15}$$

Demnach ergibt sich im Prognosevektor $\hat{\mathbf{y}}$ unabhängig von der Wahl anderer Einflußgrößen stets für $\hat{y}_1$ der Wert y_1, wenn der in Abschnitt A.3.1 definierte Dummy verwendet wird. Dies gilt allerdings nur, wenn ein einziges Unternehmen identifiziert wird, wie der folgende Unterabschnitt deutlich macht.

A.3.3 Einfluß von Kovariablen bei nicht eindeutiger Identifikation

Es soll nun (genau wie bei der Behandlung eines künstlichen Ausreißers) noch der Fall untersucht werden, dass durch den Dummy mehr als ein Unternehmen "gepickt" wird. Dies dürfte vor allem für die zweite Variante des Dummys (siehe Abschnitt A.3.1 gelten. Formal bedeutet dies, dass die im vorigen Unterabschnitt eingeführte Matrix folgende Gestalt hat:

$$\mathbf{X}_2 \quad = \quad \begin{pmatrix} \boldsymbol{\iota}_q \\ \mathbf{0} \end{pmatrix} \qquad .$$

Dabei gehen wir davon aus, dass q Unternehmen gepickt werden (und diese die ersten q Beobachtungen darstellen). $\boldsymbol{\iota}_q$ ist ein Einsvektor von der Dimension q und $\mathbf{0}$ steht für einen Nullvektor mit $n-q$ Elementen. Weiter erhalten wir

$$\mathbf{X}_2\left(\boldsymbol{X}_2'\boldsymbol{X}_2\right)^{-1}\boldsymbol{X}_2' \quad = \quad \frac{1}{q}\begin{pmatrix} \boldsymbol{\iota}_q\boldsymbol{\iota}_q' & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}$$

und

$$\begin{aligned}\mathbf{X}_2\left(\boldsymbol{X}_2'\boldsymbol{X}_2\right)^{-1}\boldsymbol{X}_2'\left(\boldsymbol{y}-\boldsymbol{X}_1\hat{\boldsymbol{\beta}}_1\right) &= \tfrac{1}{q}\begin{pmatrix} \boldsymbol{\iota}_q\boldsymbol{\iota}_q' & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}\left(\boldsymbol{y}-\boldsymbol{X}_1\hat{\boldsymbol{\beta}}_1\right) \\ &= \begin{pmatrix} \overline{y}_q \\ \overline{y}_q \\ \vdots \\ \overline{y}_q \\ 0 \\ \vdots \\ 0 \end{pmatrix} - \begin{pmatrix} 1 & \overline{x}_q^{(2)} & \dots & \overline{x}_q^{(K)} \\ 1 & \overline{x}_q^{(2)} & \dots & \overline{x}_q^{(K)} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & \overline{x}_q^{(2)} & \dots & \overline{x}_q^{(K)} \\ 0 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 \end{pmatrix}\hat{\boldsymbol{\beta}}_1 \quad ,\end{aligned}$$

Dabei verwenden wir

$$\overline{y}_q = \frac{1}{q}\sum_{i=1}^{q} y_i \quad \text{und} \quad \overline{x}_q^{(k)} = \frac{1}{q}\sum_{i=1}^{m} x_{ik}\,,\; k=2,\dots,K\;.$$

Für die ersten q Elemente des Prognosevektors $\hat{\mathbf{y}}$ erhalten wir demnach

$$\hat{y}_i \quad = \quad \overline{y}_q \;+\; \left(\; 1, \quad x_{i2}-\overline{x}_q^{(2)}, \quad \dots, \quad x_{iK}-\overline{x}_q^{(K)} \;\right)\hat{\boldsymbol{\beta}}_1\,, i=1,2,\dots,q \quad . \qquad \text{(A-16)}$$

Die Formel (A-16) zeigt folgendes:

- Falls $q=1$ ist und damit eine eindeutige Identifikation vorliegt, reduziert sich dies Ergebnis auf die Formel (A-15), denn dann ist $\overline{y}_q = y_1$ und für alle Regressoren gilt $\overline{x}_q^{(k)} = x_{1k}$.
- Die Abweichung zwischen $\hat{y}_i$ und y_i beschränkt sich auf die Werte der abhängigen Variablen, sofern für alle Regressoren $x_{ik} = const, i=1,2,\dots,m$ gilt, soll heißen, sofern der zweite Summand in (A-16) wegfällt. Dieser Fall ist jedoch wenig wahrscheinlich.
- Sofern allerdings nur der identifizierende Dummy in einer Einfachregression verwendet wird, gilt in der Tat
  $$\hat{y}_i \quad = \quad \overline{y}_q$$
  für alle q Unternehmen. Siehe dazu die empirische Analyse in Abschnitt 3.2.2.
- Im allgemeinen wird jedoch im Fall der nicht-eindeutigen Identifikation keine Aussage über Richtung und Ausmaß der Abweichung zwischen $\hat{y}_i$ und y_i , $i=1,2,\dots,m$, möglich sein.

Man beachte, dass diese Ergebnisse formal mit denen in Abschnitt A.2.3 übereinstimmen. Allerdings wird in der Praxis im Fall einer Dummyvariable ein Unternehmen entweder "gepickt" oder nicht. Dagegen können sich im Fall des künstlichen Ausreißers unterschiedliche (unterschiedlich große) Werte für die verschiedenen Ausreißer ergeben.

Bisher sind in der Reihe folgende FDZ-Arbeitspapiere erschienen:

Arbeitspapier Nr. 32: Compiling a Harmonized Database from Germany´s 1978 to 2003 Sample Surveys of Income and Expenditure., T. Bönke/ C. Schröder/ C. Werdt, Mai 2010

Arbeitspapier Nr. 31: The Research Potential of New Types of Enterprise Data based on Surveys from Official Statistics in Germany., J. Wagner, Oktober 2009

Arbeitspapier Nr. 30: Geschlechterspezifische Einkommensunterschiede bei Selbstständigen im Vergleich zu abhängig Beschäftigten - Ein empirischer Vergleich auf der Grundlage steuerstatistischer Mikrodaten, P. Eilsberger/ M. Zwick, Januar 2008

Arbeitspapier Nr. 29: Reichtum in Niedersachsen und anderen Bundesländern -Ergebnisse der Steuergeschäftsstatistik 2003 für Selbstständige (Freie Berufe und Unternehmer) und abhängig Beschäftigte, P. Böhm/J. Merz, November 2008

Arbeitspapier Nr. 28: Exports and Productivity in the German Business Services Sector. First Evidence from the Turnover Tax Statistics Panel, A. Vogel, Juli 2009

Arbeitspapier Nr. 27: Künstler in den Daten der amtlichen Statistik, C. Haak, August 2008

Arbeitspapier Nr. 26: Union Density and Varieties of Coverage: The Anatomy of Union Wage Effects in Germany, B. Fitzenberger/ K. Kohn/ A. C. Lembcke, August 2008

Arbeitspapier Nr. 25: German engineering firms during the 1990's. How efficient are export champions?, A. Schiersch, Juli 2008

Arbeitspapier Nr. 24: Zum Einkommensreichtum Älterer in Deutschland – Neue Reich-tumskennzahlen und Ergebnisse aus der Lohn- und Einkommensteuerstatistik (FAST 2001), P. Böhm/J. Merz, Februar 2008

Arbeitspapier Nr. 23: Neue Datenangebote in den Forschungsdatenzentren. Betriebs- und Unternehmensdaten im Längsschnitt, M. Brandt/ D. Oberschachtsiek/ R. Pohl, November 2007

Arbeitspapier Nr. 22: Stichprobendaten von Versicherten der gesetzlichen Krankenversicherung - Grundlage und Struktur des Datenmaterials, P. Lugert, Dezember 2007

Arbeitspapier Nr. 21: KombiFid - Kombinierte Firmendaten für Deutschland, S. Bender/ J. Wagner/M. Zwick, November 2007

Arbeitspapier Nr. 20: Neue Möglichkeiten zur Nutzung vertraulicher amtlicher Personen- und Firmendaten, U. Kaiser/ J. Wagner, Juni 2007

Arbeitspapier Nr. 18: Die Gehalts- und Lohnstrukturerhebung: Methodik, Datenzugang und Forschungspotential, H.-P. Hafner/ R. Lenz, Mai 2007

Arbeitspapier Nr. 17: Anonymisation of Linked Employer Employee Datasets. Theoretical Thoughts and an Application to the German Structure of Earnings Survey, H.-P. Hafner/ R. Lenz, Dezember 2006

Arbeitspapier Nr. 16: Die europäische Union - Integration von unten oder Eliteprojekt? Eine Sekundäranalyse von Mikrodaten der amtlichen Statistik, R. Nauenburg, November 2006

Arbeitspapier Nr. 15: Keeping in Touch - A Benefit of Public Holidays Using German Time Use diary Data, J. Merz/ L. Osberg, November 2006

Arbeitspapier Nr. 14: Zur Konzeption eines Taxpayer-Panels für Deutschland, D. Vorgrimler/ C. Gräb/ S. Kriete-Dodds, November 2006

Arbeitspapier Nr. 13: Anonymisierte Daten der amtlichen Steuerstatistik, D. Vorgrimler, September 2006

Arbeitspapier Nr. 12: Mikrosimulation in der Betriebswirtschaftlichen Steuerlehre, R. Maiterth, August 2006

Arbeitspapier Nr. 11: Der Anteil der freien Berufe und der Gewerbetreibenden an der Gemeindefinanzierung, M. Zwick, September 2006

Arbeitspapier Nr. 10: Konstruktion und Bewertung eines ökonomischen Einkommens aus der Faktisch Anonymisierten Lohn- und Einkommensteuerstatistik, T. Bönke/F. Neher/C. Schröder, August 2006

Arbeitspapier Nr. 9: Anonymising business micro data - results of a German project, R. Lenz/ M. Rosemann/D. Vorgrimler/R. Sturm, Juni 2006

Arbeitspapier Nr. 8: Scientific analyses using the Continuing Vocational Training Survey 2000, R. Lenz/ H.-P. Hafner/ D. Schmidt, Juni 2006

Arbeitspapier Nr. 7: A standard for the release of microdata, R. Lenz/ D. Vorgrimler/ M. Scheffler, Juni 2006

Arbeitspapier Nr. 6: Measuring the disclosure protection of micro aggregated business microdata, R. Lenz, Juni 2006

Arbeitspapier Nr. 5: De facto anonymised microdata file on income tax statistics 1998, J. Merz/ D. Vorgrimler/M. Zwick, Oktober 2005

Arbeitspapier Nr. 4: Matching German turnover tax statistics, R. Lenz/ D. Vorgrimler, Juni 2005

Arbeitspapier Nr. 3: The research data centres of the Federal Statistical Office and the statistical offices of the Länder, S. Zühlke/ M. Zwick/ S. Scharnhorst/ T. Wende, März 2005

Arbeitspapier Nr. 2: Eine kommunale Einkommen- und Körperschaftsteuer als Alternative zur deutschen Gewerbesteuer: Eine empirische Analyse für ausgewählte Gemeinden, R. Maiterth/ M. Zwick, April 2005

Arbeitspapier Nr. 1: Ein Vergleich der Ergebnisse von Mikrosimulationen mit denen von Gruppensimulationen auf Basis der Einkommensteuerstatistik, H. Müller, März 2005

Stand: 02.08.2010

Statistische Ämter des Bundes und der Länder,
FDZ-Arbeitspapier Nr. 33, Remote Access. Eine Welt ohne Mikrodaten?, 2011