

Ronning, Gerd

**Working Paper**

## Stochastische Überlagerung mit Hilfe der Mischungsverteilung

IAW Diskussionspapiere, No. 30

**Provided in Cooperation with:**

Institute for Applied Economic Research (IAW)

*Suggested Citation:* Ronning, Gerd (2007) : Stochastische Überlagerung mit Hilfe der Mischungsverteilung, IAW Diskussionspapiere, No. 30, Institut für Angewandte Wirtschaftsforschung (IAW), Tübingen

This Version is available at:

<https://hdl.handle.net/10419/21868>

**Standard-Nutzungsbedingungen:**

Die Dokumente auf EconStor dürfen zu eigenen wissenschaftlichen Zwecken und zum Privatgebrauch gespeichert und kopiert werden.

Sie dürfen die Dokumente nicht für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, öffentlich zugänglich machen, vertreiben oder anderweitig nutzen.

Sofern die Verfasser die Dokumente unter Open-Content-Lizenzen (insbesondere CC-Lizenzen) zur Verfügung gestellt haben sollten, gelten abweichend von diesen Nutzungsbedingungen die in der dort genannten Lizenz gewährten Nutzungsrechte.

**Terms of use:**

*Documents in EconStor may be saved and copied for your personal and scholarly purposes.*

*You are not to copy documents for public or commercial purposes, to exhibit the documents publicly, to make them publicly available on the internet, or to distribute or otherwise use the documents in public.*

*If the documents have been made available under an Open Content Licence (especially Creative Commons Licences), you may exercise further usage rights as specified in the indicated licence.*



# IAW-Diskussionspapiere

Discussion Paper

| 30 |

## Stochastische Überlagerung mit Hilfe der Mischungsverteilung

Gerd Ronning

April 2007

ISSN: 1617-5654

INSTITUT FÜR  
ANGEWANDTE  
WIRTSCHAFTSFORSCHUNG

Ob dem Himmelreich 1  
72074 Tübingen  
T: (0 70 71) 98 96-0  
F: (0 70 71) 98 96-99  
E-Mail: [iaw@iaw.edu](mailto:iaw@iaw.edu)  
Internet: [www.iaw.edu](http://www.iaw.edu)



# IAW-Diskussionspapiere

Das Institut für Angewandte Wirtschaftsforschung (IAW) Tübingen ist ein unabhängiges außeruniversitäres Forschungsinstitut, das am 17. Juli 1957 auf Initiative von Professor Dr. Hans Peter gegründet wurde. Es hat die Aufgabe, Forschungsergebnisse aus dem Gebiet der Wirtschafts- und Sozialwissenschaften auf Fragen der Wirtschaft anzuwenden. Die Tätigkeit des Instituts konzentriert sich auf empirische Wirtschaftsforschung und Politikberatung.

Dieses **IAW-Diskussionspapier** können Sie auch von unserer IAW-Homepage als pdf-Datei herunterladen:

<http://www.iaw.edu/Publikationen/IAW-Diskussionspapiere>

## ISSN 1617-5654

Weitere Publikationen des IAW:

- IAW-News (erscheinen 4x jährlich)
- IAW-Report (erscheinen 2x jährlich)
- IAW-Wohnungsmonitor Baden-Württemberg (erscheint 1x jährlich kostenlos)
- IAW-Forschungsberichte

Möchten Sie regelmäßig eine unserer Publikationen erhalten, dann wenden Sie sich bitte an uns:

IAW Tübingen, Ob dem Himmelreich 1, 72074 Tübingen,  
Telefon 07071 / 98 96-0  
Fax 07071 / 98 96-99  
E-Mail: [iaw@iaw.edu](mailto:iaw@iaw.edu)

Aktuelle Informationen finden Sie auch im Internet unter: <http://www.iaw.edu>

---

Der Inhalt der Beiträge in den IAW-Diskussionspapieren liegt in alleiniger Verantwortung der Autorinnen und Autoren und stellt nicht notwendigerweise die Meinung des IAW dar.



# Stochastische Überlagerung mit Hilfe der Mischungsverteilung<sup>1</sup>

GERD RONNING

(Stand: 19. April 2007- Version 20)

## Inhaltsverzeichnis

<b>1</b>	<b>Einleitung</b>	<b>5</b>
<b>2</b>	<b>Eindimensionale Mischungsverteilungen</b>	<b>5</b>
2.1	Stetige Verteilung . . . . .	5
2.2	Diskrete Verteilung . . . . .	6
2.3	Beliebig viele stetige (oder diskrete) Zufallsvariable . . . . .	6
<b>3</b>	<b>Mehrdimensionale Mischungsverteilungen</b>	<b>7</b>
<b>4</b>	<b>Stochastische Überlagerung durch Mischungsverteilungen</b>	<b>9</b>
4.1	Eindimensionaler Fall . . . . .	10
4.1.1	Additive Überlagerung . . . . .	10
4.1.2	Multiplikative Überlagerung . . . . .	10
4.2	Mehrdimensionaler Fall . . . . .	11
4.2.1	Additive Überlagerung . . . . .	11
4.2.2	Multiplikative Überlagerung . . . . .	11
<b>5</b>	<b>Das Höhne-Verfahren</b>	<b>12</b>
5.1	Ein einziges Merkmal . . . . .	12
5.1.1	Additiver Fall . . . . .	12
5.1.2	Multiplikativer Fall . . . . .	13
5.1.3	Dichtefunktion im Höhne-Ansatz . . . . .	14

---

<sup>1</sup>Wirtschaftswissenschaftliche Fakultät, Universität Tübingen, D-72074 Tübingen. email: gerd.ronning@uni-tuebingen.de. Ausarbeitung im Rahmen des Projektes "Wirtschaftsstatistische Paneldaten und faktische Anonymisierung", das vom Bundesministerium für Bildung und Forschung finanziell gefördert wird. Frühere Fassungen trugen den Titel "Mischungsverteilung und Höhne-Verfahren - ein Vergleich". Ich danke Martin Rosemann für viele sehr hilfreiche Kommentare und Hinweise.

5.2	Mehrere Merkmale gemeinsam . . . . .	15
5.2.1	Formale Darstellung des Verfahrens . . . . .	15
5.2.2	Ableitung der Dichtefunktion . . . . .	17
<b>6</b>	<b>Anonymisierung mittels stochastischer Überlagerung</b>	<b>18</b>
6.1	Additive Überlagerung . . . . .	18
6.2	Multiplikative Überlagerung in univariaten Verteilungen . . . . .	19
6.3	Multiplikative Überlagerung in multivariaten Verteilungen . . . . .	19
6.3.1	Überlagerung mit Hilfe eines konstanten Faktors . . . . .	20
6.3.2	Überlagerung mit Hilfe unkorrelierter Störvariablen . . . . .	21
6.3.3	Überlagerung mit Hilfe der Mischungsverteilung . . . . .	22
6.3.4	Zusammenfassung . . . . .	24
6.4	Korrelation bei multiplikativer Überlagerung . . . . .	24
<b>7</b>	<b>Einfluß der Überlagerung auf Quotienten</b>	<b>26</b>
7.1	Einleitende Bemerkungen . . . . .	26
7.2	Das Verhältnis von $Y_1$ zu $Y_2$ bei multiplikativer Überlagerung . . . . .	27
7.3	Simulationsergebnisse . . . . .	29
<b>8</b>	<b>Lineare Modelle mit Fehler in den Variablen aus Mischungsverteilungen</b>	<b>31</b>
8.1	Einleitung . . . . .	31
8.2	Spezielle und flexible Höhne-Spezifikation . . . . .	32
8.3	Additive Überlagerung im linearen Regressionsmodell . . . . .	33
8.4	Multiplikative Überlagerung im linearen Regressionsmodell . . . . .	35
<b>9</b>	<b>Ergänzende Überlegungen für Paneldaten</b>	<b>37</b>
9.1	Allgemeines . . . . .	37
9.1.1	Überlagerungsstrategien . . . . .	37
9.1.2	Das einfache lineare Panelmodell . . . . .	37
9.2	Überlagerung von Paneldaten . . . . .	38

9.2.1	Symbolik . . . . .	38
9.2.2	Additive Überlagerung . . . . .	38
9.2.3	Multiplikative Überlagerung . . . . .	39
9.3	Das einfache lineare Panelmodell mit Individualeffekten . . . . .	40
9.3.1	Feste Individualeffekten . . . . .	40
9.3.2	Stochastische Individualeffekte . . . . .	41
9.4	Schätzung des linearen Panelmodells aus anonymisierten Daten . . . . .	41
9.4.1	Der naive Panelschätzer . . . . .	41
9.4.2	Additive Meßfehler allgemein . . . . .	41
9.4.3	Einschub: Korrekter Wahrscheinlichkeitsgrenzwert im Panelfall ? . . . . .	42
9.4.4	Meßfehler mit Faktorstruktur . . . . .	43
9.4.5	Additive Überlagerung a la Höhne . . . . .	43
9.4.6	Multiplikative Überlagerung des Regressors . . . . .	45
9.4.7	Multiplikative Überlagerung beider (aller) Variablen . . . . .	47
9.4.8	Multiplikative Überlagerung der abhängigen Variablen . . . . .	49
9.5	Korrekturschätzer . . . . .	49
<b>10</b>	<b>Literatur</b>	<b>50</b>
<b>A</b>	<b>Die (spezielle) Höhne-Spezifikation im additiven Fall</b>	<b>51</b>
<b>B</b>	<b>Flexible Höhne-Spezifikation</b>	<b>51</b>
B.1	Multiplikativer Fall . . . . .	51
B.2	Additiver Fall . . . . .	53
<b>C</b>	<b>Höhne-Verfahren und Paneldaten für mehrere Regressoren</b>	<b>53</b>
C.1	Allgemeines . . . . .	53
C.2	Ausschließliche Überlagerung der Regressoren . . . . .	55
C.3	Gemeinsame Überlagerung aller Variablen . . . . .	58
C.4	Eine alternative Schreibweise . . . . .	60





# 1 Einleitung

Diese Arbeit beschäftigt sich mit speziellen Aspekten der stochastischen Überlagerung, die als Anonymisierungsmethode zur Sicherstellung der faktischen Anonymität von Mikrodaten eingesetzt wird.<sup>2</sup> Insbesondere soll der Zusammenhang zwischen der Überlagerung mittels einer Mischungsverteilung einerseits und der Überlagerung nach dem sogenannten "Höhne-Verfahren" im Einzelnen dargestellt werden. Wir gehen dabei auch auf die Auswirkungen auf die Beziehung zwischen verschiedenen Merkmalen ein. Dabei wird zunächst nur der Fall von Querschnittsdaten betrachtet. Die Auswirkungen auf die Analyse von Verhältniszahlen sowie auf die Schätzung linearer Modelle wird eingehend untersucht. Im letzten Abschnitt werden ergänzende Überlegungen für Paneldaten angestellt.

## 2 Eindimensionale Mischungsverteilungen

### 2.1 Stetige Verteilung

Es seien  $V_1$  und  $V_2$  zwei stetige Zufallsvariable mit Dichtefunktionen  $f_1(v)$  und  $f_2(v)$  sowie Erwartungswerten  $\mu_i$  und Varianzen  $\sigma_i^2$  für  $i = 1, 2$ . Wir sagen, eine Zufallsvariable  $U$  entstamme einer (diskreten) Mischung der Verteilungen von  $V_1$  und  $V_2$ , falls ihre Dichtefunktion durch

$$g(u) = \alpha f_1(u) + (1 - \alpha) f_2(u) \quad , \quad 0 < \alpha < 1 \quad , \quad (2-1)$$

gegeben ist, wobei bis auf weiteres unerheblich ist, ob die beiden Ausgangsvariablen  $V_1$  und  $V_2$  stochastisch unabhängig sind oder nicht (siehe unten).

Gedanklich kann man sich die Mischungsverteilung auch wie folgt vorstellen: Es gibt zwei (bzw. allgemeiner  $k$  - siehe unten) Zustände, die jeweils mit Wahrscheinlichkeit  $\alpha_i$  auftreten und die sich gegenseitig ausschließende Ereignisse darstellen. Für jeden Zustand gibt es eine Verteilung der Zufallsvariablen  $U$ . Je nachdem welcher Zustand eintritt, wird der Wert der Zufallsvariablen  $U$  aus der betreffenden Verteilung generiert.<sup>3</sup> Diese Modellvorstellung nutzt man aus, um Zufallszahlen aus Mischungsverteilungen zu erzeugen.

Symbolisch schreiben wir die "Mischung" auch wie folgt:

$$U \sim \alpha F_1(\boldsymbol{\theta}_1) + (1 - \alpha) F_2(\boldsymbol{\theta}_2) \quad ,$$

wobei  $F_i(\boldsymbol{\theta}_i)$  die Verteilungsfunktion von  $V_i$  bezeichnet und  $\boldsymbol{\theta}_i$  deren Parametervektor. Im Fall der Normalverteilung schreiben wir

$$U \sim \alpha N(\mu_1, \sigma_1^2) + (1 - \alpha) N(\mu_2, \sigma_2^2) \quad . \quad (2-2)$$

Für den Erwartungswert von  $U$  ergibt sich unter Verwendung von (2-1) direkt

$$E[U] = \alpha \mu_1 + (1 - \alpha) \mu_2 \quad . \quad (2-3)$$

---

<sup>2</sup>Siehe Ronning et al (2005).

<sup>3</sup>Siehe z.B. McLachlan und Peel (2000 section 1.4 "Interpretation of Mixture Models").

Ferner ergibt sich für den Erwartungswert der quadrierten Zufallsvariablen

$$E[U^2] = \alpha E[V_1^2] + (1 - \alpha) E[V_2^2] .$$

Daraus folgt für die Varianz von  $U^4$

$$\begin{aligned} V[U] &= [E[U^2] - (E[U])^2] \\ &= \alpha E[V_1^2] + (1 - \alpha) E[V_2^2] - [\alpha \mu_1 + (1 - \alpha) \mu_2]^2 \\ &= \alpha (\sigma_1^2 + \mu_1^2) + (1 - \alpha) (\sigma_2^2 + \mu_2^2) - [\alpha \mu_1 + (1 - \alpha) \mu_2]^2 \quad (2-4) \\ &= \alpha \sigma_1^2 + (1 - \alpha) \sigma_2^2 + \alpha(1 - \alpha) \mu_1^2 + (1 - \alpha) \alpha \mu_2^2 - 2\alpha(1 - \alpha) \mu_1 \mu_2 \\ &= \alpha \sigma_1^2 + (1 - \alpha) \sigma_2^2 + \alpha(1 - \alpha) (\mu_1 - \mu_2)^2 . \end{aligned}$$

Die letzte Zeile zeigt, daß nicht die absolute Größe der beiden Parameter  $\mu_1$  und  $\mu_2$ , sondern nur deren Differenz relevant ist. Im übrigen unterscheidet sich dies Ergebnis von dem, das man bekommt, wenn man die **Varianz der Summe aus  $V_1$  und  $V_2$**  bestimmt.<sup>5</sup> Dort würde auch von Bedeutung sein, ob die beiden Zufallsvariablen stochastisch unabhängig sind oder nicht !

## 2.2 Diskrete Verteilung

Für diesen Fall gelten dieselben Formeln wie im Fall von zwei stetigen Zufallsvariablen. Allerdings ist die Interpretation jetzt etwas anders. Dazu unterstellen wir, daß für  $V_1$  und  $V_2$  folgende Verteilungen gegeben sein sollen:

$$V_1 = \begin{cases} 1 & \text{mit Wahrscheinlichkeit } 0.5 \\ 3 & \text{mit Wahrscheinlichkeit } 0.5 \end{cases}, \quad V_2 = \begin{cases} 2 & \text{mit Wahrscheinlichkeit } 0.5 \\ 4 & \text{mit Wahrscheinlichkeit } 0.5 \end{cases}$$

Man beachte, daß die aus der Mischung von zwei **binären** Zufallsvariablen resultierende Zufallsvariable  $Y$  nun die vier verschiedenen Werte 1 , 2 , 3 und 4 annehmen kann. Alle vier Werte werden jeweils mit Wahrscheinlichkeit 1/4 realisiert, wenn  $\alpha = 0.5$  gilt.

## 2.3 Beliebige viele stetige (oder diskrete) Zufallsvariable

Für eine beliebig große Zahl  $k$  von Zufallsvariablen lautet die Formel für die Dichtefunktion

$$g(u) = \sum_{i=1}^k \alpha_i f_i(u) \quad , \quad 0 < \alpha_i < 1 \text{ und } \sum_i \alpha_i = 1 \quad , \quad (2-5)$$

und für Erwartungswert und Varianz erhalten wir

$$E[U] = \sum_{i=1}^k \alpha_i \mu_i \quad , \quad (2-6)$$

<sup>4</sup>Im folgenden wird  $V[U]$  und  $\sigma_u^2$  gleichgesetzt. Davon zu unterscheiden ist der später eingeführte Mischungsverteilungs-Parameter  $\sigma_m^2$ .

<sup>5</sup> $Var[V_1 + V_2] = Var[V_1] + Var[V_2] + 2 cov[V_1, V_2]$ .

sowie<sup>6</sup>

$$V[U] = \sum_{i=1}^k \alpha_i (\sigma_i^2 + \mu_i^2) - \left( \sum_i \alpha_i \mu_i \right)^2 . \quad (2-7)$$

### 3 Mehrdimensionale Mischungsverteilungen

Es sei  $\mathbf{V}_i$  ein  $r$ -dimensionaler Zufallsvektor mit (gemeinsamer) Dichtefunktion

$$f_i(v_{1i}, v_{2i}, \dots, v_{ri})$$

sowie Erwartungsvektor

$$\boldsymbol{\mu}_i = \begin{pmatrix} \mu_{1i} \\ \mu_{2i} \\ \vdots \\ \mu_{r-1,i} \\ \mu_{r,i} \end{pmatrix}$$

und Kovarianzmatrix

$$\boldsymbol{\Sigma}_i = \begin{pmatrix} \sigma_{11i} & \sigma_{12i} & & \sigma_{1,r-1,i} & \sigma_{1ri} \\ \sigma_{21i} & \sigma_{22i} & & \sigma_{2,r-1,i} & \sigma_{2ri} \\ & & \ddots & & \\ \sigma_{r-1,1i} & \sigma_{r-1,2i} & & \sigma_{r-1,r-1,i} & \sigma_{r-1,ri} \\ \sigma_{r1i} & \sigma_{r2i} & & \sigma_{r,r-1,i} & \sigma_{rri} \end{pmatrix}$$

Wir sagen, der  $r$ -dimensionale Zufallsvektor  $\mathbf{U}$  sei durch eine diskrete Mischung von  $k$  verschiedenen  $r$ -dimensionalen Verteilungen erzeugt worden, wenn die (gemeinsame) Dichtefunktion für diesen Zufallsvektor durch

$$g(u_1, u_2, \dots, u_r) = \sum_{i=1}^k \alpha_i f_i(u_1, u_2, \dots, u_r) \quad (3-1)$$

gegeben ist.

Symbolisch schreiben wir - speziell im Fall der Normalverteilung - für die Mischung:

$$\mathbf{U} \sim \sum_{i=1}^k \alpha_i N(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) .$$

Für den  $r$ -dimensionalen Vektor der Erwartungswerte des Zufallsvektors  $\mathbf{U}$  ist für jedes  $j$  das Integral

$$E[U_j] = \int \dots \int u_j g(u_1, \dots, u_j, \dots, u_r) du_r \dots du_j \dots du_1$$

---

<sup>6</sup>Eine Vereinfachung wie in Zeile 5 von Formel (2-4) für den Spezialfall von zwei Mischungskomponenten ist nicht möglich.

zu bestimmen.<sup>7</sup> Der Einfachheit halber betrachten wir die erste Komponente, d.h.  $j = 1$  und schreiben für das Integral

$$\begin{aligned} E[U_1] &= \int u_1 \left\{ \underbrace{\int \int \dots \int g(u_1, \dots, u_j \dots u_r) du_r du_{r-1} \dots du_2}_{g_1(u_1)} \right\} du_1 \\ &= \int u_1 g_1(u_1) du_1 \quad . \end{aligned}$$

Andererseits läßt sich unter Verwendung der Definition (3-1) für die gemeinsame Dichtefunktion schreiben:

$$\begin{aligned} E[U_1] &= \sum_{i=1}^k \alpha_i \int u_1 \left\{ \underbrace{\int \int \dots \int f_i(u_1, \dots, u_r) du_r du_{r-1} \dots du_2}_{f_{i1}(u_1)} \right\} du_1 \\ &= \sum_{i=1}^k \alpha_i \int u_1 f_{i1}(u_1) du_1 \\ &= \sum_{i=1}^k \alpha_i \mu_{1i} \quad . \end{aligned}$$

Dabei bezeichnet  $f_{i1}$  die Randdichte der Zufallsvariablen  $V_{i1}$ , d.h. der jeweils ersten Komponente bei den Zufallsvektoren  $\mathbf{V}_i$ . Daraus folgt

$$E[\mathbf{U}] = \sum_{i=1}^k \alpha_i \boldsymbol{\mu}_i \quad (3-2)$$

Für die Kovarianzmatrix des Zufallsvektors  $\mathbf{U}$  läßt sich unter Verwendung der Definition

$$\text{cov}[\mathbf{U}] = E[\mathbf{U}\mathbf{U}'] - E[\mathbf{U}]E[\mathbf{U}]'$$

außerdem das folgende Ergebnis herleiten:<sup>8 9</sup>

$$\text{cov}[\mathbf{U}] = \sum_{i=1}^k \alpha_i (\boldsymbol{\Sigma}_i + \boldsymbol{\mu}_i \boldsymbol{\mu}_i') - \left( \sum_{i=1}^k \alpha_i \boldsymbol{\mu}_i \right) \left( \sum_{i=1}^k \alpha_i \boldsymbol{\mu}_i \right)' \quad (3-3)$$

<sup>7</sup>Siehe z.B. die Definition bei Graybill, Introduction to Matrices with Applications in Statistics. 1969. Kap. 10.4.

<sup>8</sup>Man beachte, daß

$$E[\mathbf{U}\mathbf{U}'] = \begin{pmatrix} E[U_1^2] & E[U_1 U_2] & \dots & E[U_1 U_k] \\ E[U_1 U_2] & E[U_2^2] & \dots & E[U_2 U_k] \\ \vdots & \vdots & \ddots & \vdots \\ E[U_k U_1] & \dots & \dots & E[U_k^2] \end{pmatrix}$$

und beispielsweise

$$E[U_1 U_2] = \sigma_{12} + \mu_1 \mu_2$$

gilt.

<sup>9</sup>Aus (3-3) folgt, daß das  $(t, s)$ -Element der Kovarianz-Matrix (3-3) wie folgt aussieht:

$$\text{cov}(u_t, u_s) = \sum_i \alpha_i (\sigma_i(t, s) + \mu_i(t) \mu_i(s)) - \left( \sum_i \alpha_i \mu_i(t) \right) \left( \sum_i \alpha_i \mu_i(s) \right).$$

Dabei ist  $\sigma_i(t, s)$  das  $(t, s)$ -Element aus Matrix  $\boldsymbol{\Sigma}_i$  und  $\mu_i(t)$  das  $t$ -Element aus dem Vektor  $\boldsymbol{\mu}_i$ . In der Überlagerung zwecks Anonymisierung dürften die  $\sigma_i(t, s)$  alle gleich Null sein, ausserdem ist  $\alpha_i = 0.5, i = 1, 2$  und  $\mu_i(t) = \pm \mu$  in der additiven Überlagerung bzw.  $\mu_i(t) = 1 \pm \delta$  in der multiplikativen Überlagerung.

Falls man unterstellt, daß alle  $k$  Komponenten der Mischungsverteilung von  $\mathbf{U}$  die gleiche Korrelation aufweisen, ist - beispielsweise für den Fall  $r = 4$  - die Kovarianzmatrix durch

$$\Sigma_j = \sigma_j^2 \begin{pmatrix} 1 & \varrho_j & \varrho_j & \varrho_j \\ \varrho_j & 1 & \varrho_j & \varrho_j \\ \varrho_j & \varrho_j & 1 & \varrho_j \\ \varrho_j & \varrho_j & \varrho_j & 1 \end{pmatrix} = \sigma_j^2 [(1 - \varrho_j) \mathbf{I} + \varrho_j \boldsymbol{\iota} \boldsymbol{\iota}'] \quad (3-4)$$

gegeben, wobei  $\varrho_j$  der Korrelationsparameter ist und  $\boldsymbol{\iota}$  den Einsvektor bezeichnet. Man beachte, daß dieser Parameter nicht beliebig negativ werden kann. Es muß

$$-\frac{1}{r-1} < \varrho_j \leq 1$$

gelten, damit die obige Matrix positiv definit ist.

Falls wir unterstellen, daß alle involvierten Mischungsverteilungen von der Struktur (3-4) sind, dann ergibt sich für die Kovarianzmatrix von  $\mathbf{U}$

$$\text{cov}[\mathbf{U}] = \sum_{i=1}^k \alpha_i (\sigma_i^2 [(1 - \varrho_i) \mathbf{I} + \varrho_i \boldsymbol{\iota} \boldsymbol{\iota}'] + \boldsymbol{\mu}_i \boldsymbol{\mu}_i') - \left( \sum_{i=1}^k \alpha_i \boldsymbol{\mu}_i \right) \left( \sum_{i=1}^k \alpha_i \boldsymbol{\mu}_i \right)' \quad (3-5)$$

## 4 Stochastische Überlagerung durch Mischungsverteilungen

Bei der stochastischen Überlagerung hat die Mischungsverteilung die Aufgabe, die Anonymisierung in der Weise zu verstärken, daß möglichst wenige Realisationen den Originalwert - annähernd - unverändert lassen. Anders gesagt: Die anonymisierten Werte sollten nicht in der Nähe der Originalwerte liegen. Deshalb sollten im Fall der additiven Überlagerung die Realisationen der Überlagerungsvariablen möglichst weit entfernt vom Wert Null liegen, im Fall der multiplikativen Überlagerung möglichst weit entfernt vom Wert Eins.<sup>10</sup> Dies läßt sich am besten bzw. einfachsten mit dem in Abschnitt 2.1 betrachteten Spezialfall von **zwei** (stetigen) Mischungskomponenten (d.h.  $k = 2$ ) erreichen. Andererseits wird man zum Zwecke der Anonymisierung beide Mischungskomponenten symmetrisch wählen, insbesondere die Gewichte und die Varianzen gleich groß wählen.

Dann ergibt sich für die additive Überlagerung

$$\begin{aligned} \text{cov}(u_t, u_s) &= \sum_i \alpha_i (\mu_i(t) \mu_i(s)) - \left( \sum_i \alpha_i \mu_i(t) \right) \left( \sum_i \alpha_i \mu_i(s) \right) \\ &= 0.5(\mu^2 + (-\mu)^2) - (0.5\mu - 0.5\mu)^2 \\ &= \mu^2 \end{aligned}$$

und für die multiplikative Überlagerung

$$\begin{aligned} \text{cov}(u_t, u_s) &= \sum_i \alpha_i (\mu_i(t) \mu_i(s)) - \left( \sum_i \alpha_i \mu_i(t) \right) \left( \sum_i \alpha_i \mu_i(s) \right) \\ &= 0.5((1 + \delta)^2 + (1 - \delta)^2) - (0.5(1 + \delta) + 0.5(1 - \delta))^2 \\ &= \delta^2 \end{aligned}$$

Dies bedeutet, daß auch in diesem speziellen Fall durch die Mischungsverteilung eine **positive** Korrelation erzeugt wird.

<sup>10</sup>Die Idee der Überlagerung mittels Mischungsverteilung und insbesondere der multiplikativen Überlagerung wurde ebenfalls von Massell, Zayatz und Funk(2006) vorgeschlagen. Diese Autoren verwenden eine "gesplittete Dreiecksverteilung", was einer Mischungsverteilung entspricht, bei der zwei deutlich separierte, symmetrisch um 1 positionierte, Intervalle positive Wahrscheinlichkeits-Masse besitzen.

Wir behandeln zunächst - ausführlich - den eindimensionalen Fall und dann eher kursorisch den mehrdimensionalen Fall. In beiden Fällen wird zunächst die additive und dann die multiplikative Überlagerung betrachtet.

## 4.1 Eindimensionaler Fall

Im folgenden beschreiben wir die speziellen Parametrisierungen im Fall der Überlagerung und wenden dann die Ergebnisse aus Abschnitt 2.1 an, um die ersten und zweiten Momente zu bestimmen.

### 4.1.1 Additive Überlagerung

Im Fall der additiven Überlagerung wählen wir<sup>11</sup>

$$\mu_1 = \mu_m, \mu_2 = -\mu_m \quad (4-1)$$

und

$$\sigma_1^2 = \sigma_2^2 = \sigma_m^2 \quad (4-2)$$

sowie

$$\alpha_1 = \alpha_2 = \frac{1}{2} \quad (4-3)$$

Dann ergibt sich aus den allgemeinen Formeln (2-3) und (2-4)

$$E[U] = 0$$

und

$$V[U] = \sigma_m^2 + \mu_m^2,$$

d.h.  $\sigma_u^2 = \sigma_m^2 + \mu_m^2$  in der alternativen Schreibweise.

### 4.1.2 Multiplikative Überlagerung

Im Fall der multiplikativen Überlagerung wählen wir die alternative Spezifikation

$$\mu_1 = 1 + \delta_m, \mu_2 = 1 - \delta_m \quad (4-4)$$

für die Erwartungswerte. Im Übrigen werden die Spezifikationen des additiven Falls beibehalten: Für die Varianzen gilt (4-2) und für die Gewichte gilt (4-3).

Dann ergibt sich aus den allgemeinen Formeln (2-3) und (2-4)

$$E[U] = 1$$

und

$$V[U] = \sigma_m^2 + \delta_m^2,$$

d.h.  $\sigma_u^2 = \sigma_m^2 + \delta_m^2$  in der alternativen Schreibweise.

---

<sup>11</sup>Der Index "m" steht für Mischung und bezeichnet Parameter, die die einzelnen Komponenten der Mischungsverteilung charakterisieren, wenn sie für alle Mischungskomponenten identisch gewählt werden.

## 4.2 Mehrdimensionaler Fall

Entsprechend dem Vorgehen im Abschnitt 4.1 beschreiben wir auch für mehrdimensionale Verteilungen die speziellen Parametrisierungen im Fall der Überlagerung und wenden dann die Ergebnisse aus Abschnitt 3 an, um die ersten und zweiten Momente zu bestimmen. Auch hier beschränken wir uns auf den Fall von  $k = 2$  Mischungskomponenten.

### 4.2.1 Additive Überlagerung

Im Fall der additiven Überlagerung wählen wir

$$\boldsymbol{\mu}_1 = \boldsymbol{\mu}_m, \boldsymbol{\mu}_2 = -\boldsymbol{\mu}_m \quad (4-5)$$

wobei  $\boldsymbol{\mu}$  nicht notwendigerweise identische Erwartungswerte für alle  $r$  Komponenten besitzt. Ferner soll

$$\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma}_m \quad (4-6)$$

sowie

$$\alpha_1 = \alpha_2 = \frac{1}{2} \quad (4-7)$$

gelten.

Dann ergibt sich aus den allgemeinen Formeln (3-2) und (3-3)

$$E[\mathbf{U}] = \mathbf{0}$$

und

$$\text{cov}[\mathbf{U}] = \boldsymbol{\Sigma}_m + \boldsymbol{\mu}_m \boldsymbol{\mu}_m' \quad (4-8)$$

Man beachte, daß die spezielle Parametrisierung der Mischungsverteilung mit  $\boldsymbol{\mu}_1 = -\boldsymbol{\mu}_2 = \boldsymbol{\mu}_m$  eine "Vergrößerung" der Kovarianzmatrix bewirkt!

### 4.2.2 Multiplikative Überlagerung

Im Fall der multiplikativen Überlagerung wählen wir die alternative Spezifikation

$$\boldsymbol{\mu}_1 = \boldsymbol{\nu} + \boldsymbol{\delta}_m, \boldsymbol{\mu}_2 = \boldsymbol{\nu} - \boldsymbol{\delta}_m \quad (4-9)$$

für die Erwartungswerte, wobei  $\boldsymbol{\delta}_m$  nicht notwendigerweise identische Werte für alle  $r$  Komponenten besitzt. Im Übrigen werden die Spezifikationen des additiven Falls beibehalten: Für die Varianzen gilt (4-2) und für die Gewichte gilt (4-3).

Dann ergibt sich aus den allgemeinen Formeln (2-3) und (2-4)

$$E[\mathbf{U}] = \boldsymbol{\nu}$$

und

$$\text{cov}[\mathbf{U}] = \boldsymbol{\Sigma}_m + \boldsymbol{\delta}_m \boldsymbol{\delta}_m' \quad (4-10)$$

Falls alle Komponenten von  $\boldsymbol{\delta}_m$  identisch sind, d.h. falls

$$\boldsymbol{\delta}_m = \delta_m \boldsymbol{\nu} \quad (4-11)$$



gilt, ergibt sich für die Kovarianzmatrix die spezielle Form

$$\text{cov}[\mathbf{U}] = \boldsymbol{\Sigma}_m + \delta_m^2 \boldsymbol{u}\boldsymbol{u}' .$$

Falls wir weiterhin noch

$$\boldsymbol{\Sigma}_m = \sigma_m^2 \mathbf{I} \tag{4-12}$$

verlangen, erhalten wir

$$\text{cov}[\mathbf{U}] = \sigma_m^2 \mathbf{I} + \delta_m^2 \boldsymbol{u}\boldsymbol{u}' .$$

Dieses spezielle Ergebnis wird uns später bei der Betrachtung des Höhne-Verfahrens wieder begegnen. Siehe insbesondere Abschnitt 5.2. Aber auch hier sollte bereits die Bemerkung gemacht werden, daß die Mischungsverteilung mit zwei Komponenten, die mittels des Parameters  $\delta_m$  symmetrische Zu- und Abschläge erzeugt, dazu führt, daß die  $r$  Komponenten miteinander korreliert sind. Denn man kann die Kovarianzmatrix auch wie folgt schreiben:

$$\text{cov}[\mathbf{U}] = \frac{1}{\sigma_m^2 + \delta_m^2} ((1 - \rho)\mathbf{I} + \rho \boldsymbol{u}\boldsymbol{u}')$$

Dabei steht  $\rho$  für den zuvor abgeleiteten Korrelationskoeffizienten

$$\rho = \frac{\delta_m^2}{\sigma_m^2 + \delta_m^2} .$$

Wohlgermerkt: Obwohl die Verteilungen der einzelnen Mischungskomponenten skalare Kovarianzmatrix besitzen, weist die Mischungsverteilung selbst positive Korrelation aus. Siehe (4-12) in Verbindung mit (4-2).

## 5 Das Höhne-Verfahren

Das Verfahren, das Jörg Höhne vorgeschlagen hat, ist vor allem mit der Absicht entwickelt worden, bei der Überlagerung von **verschiedenen** Merkmalen deren Beziehung zueinander, insbesondere das proportionale Verhältnis, einigermaßen zu erhalten. Hier geht es vor allem um die formale Darstellung des Verfahrens und um die Frage, ob dies Verfahren mit einer Mischungsverteilung äquivalent ist bzw. inwiefern es sich davon unterscheidet.

Wir betrachten zunächst das "Höhne-Verfahren" für ein einziges Merkmal, dann aber auch für mehrere Merkmale gemeinsam.

### 5.1 Ein einziges Merkmal

#### 5.1.1 Additiver Fall

Wir betrachten das Modell

$$U = \mu D + \varepsilon , \tag{5-1}$$

wobei  $\mu$  ein Parameter mit beliebigem Wert ist. Für die diskrete Zufallsvariable  $D$  gilt

$$D = \begin{cases} +1 & \text{mit Wahrscheinlichkeit } \alpha \\ -1 & \text{mit Wahrscheinlichkeit } 1 - \alpha \end{cases} \tag{5-2}$$

und die stetige Zufallsvariable  $\varepsilon$  ist wie folgt spezifiziert:

$$E[\varepsilon] = 0 \quad , \quad V[\varepsilon] = \sigma_\varepsilon^2 \quad (5-3)$$

wobei die Annahme der Normalverteilung üblicherweise hinzutritt. Ferner wird angenommen, daß  $D$  und  $\varepsilon$  stochastisch unabhängig sind.

Erwartungswert und Varianz von  $D$  sind durch

$$E[D] = 2\alpha - 1 \quad , \quad V[D] = 4\alpha(1 - \alpha)$$

gegeben. Wenn  $\alpha = 1/2$  gilt, ist der Erwartungswert gleich 0 und die Varianz gleich 1. Außerdem ist die Wahrscheinlichkeitsdichtefunktion von  $D$  durch

$$h(d) = \alpha^{\frac{1+d}{2}} (1 - \alpha)^{\frac{1-d}{2}} \quad (5-4)$$

für  $d \in \{+1, -1\}$  gegeben.

Für Erwartungswert und Varianz von  $U$  gemäß (5-1) ergibt sich dann

$$E[U] = \mu(2\alpha - 1) \quad (5-5)$$

sowie

$$V[U] = \mu^2 V[D] + V[\varepsilon] = 4\alpha(1 - \alpha)\mu^2 + \sigma_\varepsilon^2 \quad (5-6)$$

Insbesondere erhalten wir für den "symmetrischen" Fall, d.h. wenn  $\alpha = 1/2$  gilt,

$$E[U] = 0 \quad , \quad V[U] = \mu^2 + \sigma_\varepsilon^2 \quad ,$$

was dem Resultat für die Mischungsverteilung entspricht, wenn dort ebenfalls der "symmetrische" Fall betrachtet wird.

### 5.1.2 Multiplikativer Fall

Im Fall der multiplikativen Überlagerung verwenden wir statt (5-1) den Ansatz

$$U = (1 + \delta D) + \varepsilon \quad (5-7)$$

und erhalten in diesem Fall für Erwartungswert und Varianz

$$E[U] = 1 + \delta(2\alpha - 1) \quad (5-8)$$

sowie

$$V[U] = \delta^2 V[D] + V[\varepsilon] = 4\alpha(1 - \alpha)\delta^2 + \sigma_\varepsilon^2 \quad (5-9)$$

Für  $\alpha = 1 - \alpha = 1/2$  erhalten wir das Resultat

$$E[U] = 1 \quad , \quad V[U] = \delta^2 + \sigma_\varepsilon^2 \quad ,$$

was dem Resultat für die Mischungsverteilung (im multiplikativen Fall) entspricht, wenn dort ebenfalls der "symmetrische" Fall betrachtet wird.

### 5.1.3 Dichtefunktion im Höhne-Ansatz

Obwohl die Gleichheit der ersten und zweiten Momente dafür spricht, daß Mischungsverteilung und Höhne-Verfahren äquivalent sind, ist dies erst nachgewiesen, wenn die Gleichheit der Dichtefunktionen gezeigt ist. Dazu gehen wir wie folgt vor: Wir betrachten zunächst die bedingte Dichte von  $U$  gegeben  $D$ , multiplizieren diese Dichte mit der Randdichte von  $D$ , um die gemeinsame Dichte zu bestimmen und "integrieren" dann die Zufallsvariable  $D$  aus, um die Randdichte von  $U$  zu erhalten. Dabei ist in diesem Fall allerdings wegen der Diskretheit von  $D$  eine Summation vorzunehmen. Wir unterstellen im folgenden zusätzlich

$$\varepsilon \sim N(0, \sigma_\varepsilon^2) \quad . \quad (5-10)$$

**Additiver Fall:** Für gegebenes  $D$  ist  $U$  in (5-1) normalverteilt, d.h.

$$U|D = d \sim N(\mu d, \sigma_\varepsilon^2) \quad ,$$

wobei  $N$  die Dichtefunktion der Normalverteilung symbolisieren soll. Für die gemeinsame Dichte von  $U$  und  $D$  erhalten wir dann

$$g(u, d) = h(d) \cdot N(\mu d, \sigma_\varepsilon^2)$$

mit  $h(d)$  aus (5-4). Dann erhalten wir die Randdichte von  $U$  durch folgende Formel:

$$f(u) = \sum_{d \in \{+1, -1\}} g(u, d) = \sum_{d \in \{+1, -1\}} \alpha^{\frac{1+D}{2}} (1 - \alpha)^{\frac{1-D}{2}} \cdot N(\mu d, \sigma_\varepsilon^2)$$

oder

$$f(u) = \alpha N(\mu_1, \sigma_\varepsilon^2) + (1 - \alpha) N(\mu_2, \sigma_\varepsilon^2) \quad . \quad (5-11)$$

Dies entspricht exakt der Dichte der Mischungsverteilung in (2-2), allerdings für den Spezialfall identischer Varianzen.

**Multiplikativer Fall:** Für gegebenes  $D$  ist  $U$  in (5-7) normalverteilt, d.h.

$$U|D = d \sim N(1 + \delta d, \sigma_\varepsilon^2) \quad .$$

Wie im additiven Fall erhalten wir für die gemeinsame Dichte von  $U$  und  $D$  dann

$$g(u, d) = h(d) \cdot N(1 + \delta d, \sigma_\varepsilon^2)$$

mit  $h(d)$  aus (5-4) und für die Randdichte von  $U$  ergibt sich

$$f(u) = \sum_{d \in \{+1, -1\}} g(u, d) = \sum_{d \in \{+1, -1\}} \alpha^{\frac{1+D}{2}} (1 - \alpha)^{\frac{1-D}{2}} \cdot N(1 + \delta d, \sigma_\varepsilon^2)$$

oder

$$f(u) = \alpha N(1 + \delta, \sigma_\varepsilon^2) + (1 - \alpha) N(1 - \delta, \sigma_\varepsilon^2) \quad . \quad (5-12)$$

Dies entspricht exakt der Dichte der Mischungsverteilung in (2-2), allerdings für den Spezialfall identischer Varianzen.

## 5.2 Mehrere Merkmale gemeinsam

Höhne hat vorgeschlagen, alle Merkmale mit demselben (multiplikativen) Zu- bzw. Abschlag  $\delta$  zu versehen und dann bei den einzelnen Merkmalen noch eine (merkmals-spezifische) Überlagerung durchzuführen. Dahinter steht die Idee, daß damit die proportionalen Beziehungen zwischen den einzelnen Merkmalen bzw. die daraus erzeugten Verhältniszahlen annähernd gleich bleiben.<sup>12</sup> Dies soll im folgenden dargestellt werden, sodann soll nachgewiesen werden, daß auch in diesem Fall Äquivalenz zum statistischen Modell der Mischungsverteilung besteht.

Da Höhnes Vorschlag nur die **multiplikative** Alternative betrachtet, werden wir uns im Text auf diese Variante beschränken. Appendix A stellt ergänzend auch die additive Variante kurz dar, weil sie in Abschnitt 8 ebenfalls benötigt wird.

### 5.2.1 Formale Darstellung des Verfahrens

Es wird der  $r$ -dimensionale Vektor  $\mathbf{U}$  betrachtet, für dessen  $j$ -te Komponente gelten soll:

$$U_j = (1 + \delta D) + \varepsilon_j, \quad j = 1, \dots, r, \quad (5-13)$$

Für alle  $r$  Komponenten gemeinsam kann man das auch schreiben als

$$\mathbf{U} = (1 + \delta D)\boldsymbol{\iota} + \boldsymbol{\varepsilon}, \quad (5-14)$$

wobei  $\delta$  ein Parameter mit beliebigem Wert ist. Genau wie in Abschnitt 5.1.1 ist die diskrete Zufallsvariable  $D$  durch

$$D = \begin{cases} +1 & \text{mit Wahrscheinlichkeit } \alpha \\ -1 & \text{mit Wahrscheinlichkeit } 1 - \alpha \end{cases}$$

gegeben (siehe (5-2)) und der  $r$ -dimensionale stetige Zufallsvektor  $\boldsymbol{\varepsilon}$  ist entsprechend (5-3) spezifiziert:

$$E[\boldsymbol{\varepsilon}] = \mathbf{0}, \quad \text{cov}[\boldsymbol{\varepsilon}] = \sigma_\varepsilon^2 \mathbf{I} \quad (5-15)$$

wobei die Annahme der Normalverteilung üblicherweise hinzutritt. Genau wie in Abschnitt 5.1.1 wird die stochastische Unabhängigkeit von  $D$  und  $\boldsymbol{\varepsilon}$  unterstellt.

Für den Erwartungswert der Zufallskomponente  $U_j$  aus dem Zufallsvektors  $\mathbf{U}$  ergibt sich (wie im Fall der Betrachtung eines einzigen Merkmals)

$$E[U_j] = 1 + \delta(2\alpha - 1), \quad j = 1, \dots, r \quad (5-16)$$

oder auch kompakter

$$E[\mathbf{U}] = (1 + \delta(2\alpha - 1))\boldsymbol{\iota}. \quad (5-17)$$

Im 'symmetrischen' Fall ( $\alpha = 1/2$ ) ergibt sich

$$E[\mathbf{U}] = \boldsymbol{\iota}.$$

Bei der Bestimmung der Kovarianzmatrix ist folgende Beobachtung wesentlich: Da alle  $r$  Komponenten die Zufallsvariable  $D$  enthalten, sind die Komponenten von  $\mathbf{U}$  miteinander

---

<sup>12</sup>Darauf werden wir in Abschnitt 7 näher eingehen.

korreliert, selbst wenn die Komponenten  $\varepsilon_j$  des Zufallsvektors  $\varepsilon$  unkorreliert sind!! Um die Kovarianz zu bestimmen, betrachten wir zunächst den Erwartungswert des Produktes  $U_i U_j$ . Dafür erhalten wir<sup>13</sup>

$$E_D[E_{\varepsilon|D}[U_i U_j] | D] = E_D[(1 + \delta D)^2] = E_D[1 + 2\delta D + \delta^2 D^2] = 1 + 2\delta(2\alpha - 1) + \delta^2$$

und für die Kovarianz selbst erhalten wir

$$\text{cov}[U_i U_j] = 1 + 2\delta(2\alpha - 1) + \delta^2 - (1 + \delta(2\alpha - 1))^2 = \delta^2 - \delta^2(2\alpha - 1)^2 \geq 0.$$

Diese Kovarianz ist also gleich Null nur für  $\alpha = 0$  bzw.  $\alpha = 1$ , d.h. wenn nur eine einzige Mischungskomponente verwendet wird. Ansonsten ist diese Kovarianz (und damit auch die entsprechende Korrelation) **stets positiv!** Im hier besonders interessierenden 'symmetrischen' Fall ( $\alpha = 1/2$ ) ist die Kovarianz gleich  $\delta^2$ .

Für die Varianz von  $U_j$  gilt ferner (siehe die Ergebnisse in Abschnitt 5.1.2)

$$V[U_j] = 4\alpha(1 - \alpha)\delta^2 + \sigma_\varepsilon^2,$$

und speziell für den symmetrischen Fall ( $\alpha = 1 - \alpha = 1/2$ ) erhalten wir

$$V[U] = \delta^2 + \sigma_\varepsilon^2.$$

Damit erhalten wir in diesem speziellen Fall für den Korrelationskoeffizienten eine besonders einfache Struktur:

$$\text{corr}[U_i U_j] = \frac{\delta^2}{\delta^2 + \sigma_\varepsilon^2} \quad \text{falls } \alpha = \frac{1}{2}.$$

Die durch das Verfahren erzeugte Korrelation ist also umso größer, je größer der gemeinsame Zuschlag  $\delta$  und je kleiner die Überlagerungsvarianz  $\sigma_\varepsilon^2$  ist!

Als Kovarianzmatrix erhalten wir aus obigen Ergebnissen (beispielsweise für  $r = 4$  Merkmale)

$$\text{cov}[\mathbf{U}] = \begin{pmatrix} 4\alpha(1 - \alpha)\delta^2 + \sigma_\varepsilon^2 & \delta^2 - \delta^2(2\alpha - 1)^2 & \delta^2 - \delta^2(2\alpha - 1)^2 & \delta^2 - \delta^2(2\alpha - 1)^2 \\ \delta^2 - \delta^2(2\alpha - 1)^2 & 4\alpha(1 - \alpha)\delta^2 + \sigma_\varepsilon^2 & \delta^2 - \delta^2(2\alpha - 1)^2 & \delta^2 - \delta^2(2\alpha - 1)^2 \\ \delta^2 - \delta^2(2\alpha - 1)^2 & \delta^2 - \delta^2(2\alpha - 1)^2 & 4\alpha(1 - \alpha)\delta^2 + \sigma_\varepsilon^2 & \delta^2 - \delta^2(2\alpha - 1)^2 \\ \delta^2 - \delta^2(2\alpha - 1)^2 & \delta^2 - \delta^2(2\alpha - 1)^2 & \delta^2 - \delta^2(2\alpha - 1)^2 & 4\alpha(1 - \alpha)\delta^2 + \sigma_\varepsilon^2 \end{pmatrix},$$

die sich im 'symmetrischen' Fall auf die einfache Form

$$\text{cov}[\mathbf{U}] = \sigma_\varepsilon^2 \mathbf{I} + \delta^2 \mathbf{u}' = \frac{1}{\sigma_\varepsilon^2 + \delta^2} ((1 - \rho) \mathbf{I} + \rho \mathbf{u}') \quad (5-18)$$

reduziert, wobei  $\rho$  für den zuvor abgeleiteten Korrelationskoeffizienten

$$\rho = \frac{\delta^2}{\delta^2 + \sigma_\varepsilon^2}$$

steht. Man beachte, daß diese Korrelation stets positiv ist!

<sup>13</sup>Wir verwenden

$$E[D^2] = V[D] + (E[D])^2 = 4\alpha(1 - \alpha) + (2\alpha - 1)^2 = 1.$$

## 5.2.2 Ableitung der Dichtefunktion

Wir können festhalten, daß sich für Erwartungswert und Varianz unter bestimmten Bedingungen dieselben Resultate ergeben wie bei der Mischungsverteilung: In beiden Fällen gilt

$$E[\mathbf{U}] = \boldsymbol{\nu} \quad \text{und} \quad \text{cov}[\mathbf{U}] = \sigma_\varepsilon^2 \mathbf{I} + \delta^2 \boldsymbol{\nu} \boldsymbol{\nu}' \quad (5-19)$$

Dabei wird dieses Ergebnis in der Mischungsverteilung erreicht, wenn für die Mittelwertvektoren und die Kovarianzmatrizen in allen Mischungskomponenten folgende sehr spezielle Annahme getroffen wird:

$$\boldsymbol{\mu}_j = \delta \boldsymbol{\nu} \quad \text{und} \quad \boldsymbol{\Sigma}_j = \sigma_\varepsilon^2 \mathbf{I} \quad (5-20)$$

und der 'symmetrische Fall'

$$\alpha = \frac{1}{2}$$

sowohl bei der Mischungsverteilung als auch im Höhenverfahren unterstellt wird.

Es bleibt zu prüfen, ob auch im mehrdimensionalen Fall die Dichtefunktion des Höhenverfahrens mit der Definition der Mischungsverteilung gemäß (3-1) übereinstimmt. Bei dieser Analyse gehen wir wie im eindimensionalen Fall vor. Siehe dazu Abschnitt 5.1.3. Wieder beschränken wir uns auf den hier relevanten Fall der **multiplikativen** Überlagerung.

Wir betrachten zunächst die bedingte Dichte von  $\mathbf{U}$  gegeben  $D$ , multiplizieren diese Dichte mit der Randdichte von  $D$ , um die gemeinsame Dichte zu bestimmen und "integrieren" dann die Zufallsvariable  $D$  aus, um die Randdichte von  $\mathbf{U}$  zu erhalten. Dabei unterstellen wir im folgenden wieder zusätzlich Normalverteilung:

$$\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma_\varepsilon^2 \mathbf{I}) \quad . \quad (5-21)$$

Für gegebenes  $D$  ist  $\mathbf{U}$  in (5-14) normalverteilt, d.h.

$$\mathbf{U} | (D = d) \sim N((1 + \delta d) \boldsymbol{\nu}, \sigma_\varepsilon^2 \mathbf{I}) \quad .$$

Für die gemeinsame Dichte von  $\mathbf{U}$  und  $D$  ergibt sich

$$g(\mathbf{u}, d) = h(d) \cdot N((1 + \delta d) \boldsymbol{\nu}, \sigma_\varepsilon^2 \mathbf{I})$$

mit  $h(d)$  aus (5-4) und für die Randdichte von  $\mathbf{U}$  ergibt sich

$$f(\mathbf{u}) = \sum_{d \in \{+1, -1\}} g(\mathbf{u}, d) = \sum_{d \in \{+1, -1\}} \alpha^{\frac{1+d}{2}} (1 - \alpha)^{\frac{1-d}{2}} \cdot N((1 + \delta d) \boldsymbol{\nu}, \sigma_\varepsilon^2 \mathbf{I})$$

oder

$$f(\mathbf{u}) = \alpha N((1 + \delta) \boldsymbol{\nu}, \sigma_\varepsilon^2 \mathbf{I}) + (1 - \alpha) N((1 - \delta) \boldsymbol{\nu}, \sigma_\varepsilon^2 \mathbf{I}) \quad . \quad (5-22)$$

Dies entspricht exakt der Dichte der Mischungsverteilung in (3-1), allerdings für den Spezialfall identischer (skalarer) Kovarianzmatrizen.

## 6 Anonymisierung mittels stochastischer Überlagerung

Im folgenden bezeichnet  $Y$  die originale Variable und  $Y^a$  die anonymisierte Variable, entsprechend bezeichnen  $\mathbf{Y}$  und  $\mathbf{Y}^a$  die betreffenden Vektoren im multivariaten Fall. Im folgenden gehen wir davon aus, daß alle Zufallsvariablen stetig sind.

Es soll gelten

$$E[Y] = \mu_y \text{ sowie } V[Y] = \sigma_y^2 \quad (6-1)$$

bzw. im multivariaten Fall für  $k$ -dimensionalen Vektor  $\mathbf{Y}$

$$E[\mathbf{Y}] = \boldsymbol{\mu}_y \text{ sowie } cov[\mathbf{Y}] = \boldsymbol{\Sigma}_{yy} \quad (6-2)$$

### 6.1 Additive Überlagerung

Im eindimensionalen Fall gilt

$$Y^a = Y + U \quad (6-3)$$

wobei  $U$  eine stetige Zufallsvariable mit Erwartungswert 0 und Varianz  $\sigma_u^2 > 0$  ist. Im multivariaten Fall erhalten wir

$$\mathbf{Y}^a = \mathbf{Y} + \mathbf{U} \quad (6-4)$$

wobei der Zufallsvektor  $\mathbf{U}$  Erwartungswert  $\mathbf{0}$  und Kovarianzmatrix  $cov[\mathbf{U}]$  besitzt.

Für Erwartungswert und Varianz bzw. Kovarianzmatrix ergibt sich

$$E[Y^a] = \mu_y \text{ und } V[Y^a] = \sigma_y^2 + \sigma_u^2 \quad (\text{eindimensional})$$

und

$$E[\mathbf{Y}^a] = \boldsymbol{\mu}_y \text{ und } cov[\mathbf{Y}^a] = cov[\mathbf{Y}] + cov[\mathbf{U}] \quad (\text{mehrdimensional}) \quad .$$

Falls die Überlagerung mittels der Mischungsverteilung aus Abschnitt 4.1.1 (bzw. mit dem Höhne-Verfahren) erfolgt, ergibt sich - im eindimensionalen Fall - für die Varianz der überlagerten Variablen

$$V[Y^a] = \sigma_y^2 + \sigma_m^2 + \mu_m^2 \quad (6-5)$$

wobei  $\mu_m$  und  $\sigma_m^2$  die Parameter aus der Mischungsverteilung sind, siehe (4-1).

Im multivariaten Fall ergibt sich aus (4-8) als Kovarianzmatrix der überlagerten Variablen

$$cov[\mathbf{Y}^a] = cov[\mathbf{Y}] + \boldsymbol{\Sigma}_m + \boldsymbol{\mu}_m \boldsymbol{\mu}_m' \quad , \quad (6-6)$$

wobei  $\boldsymbol{\mu}_m$  und  $\boldsymbol{\Sigma}_m$  die Parameter aus der Mischungsverteilung bezeichnen. Siehe Abschnitt 4.2.1.

## 6.2 Multiplikative Überlagerung in univariaten Verteilungen

Üblicherweise wird man bei der multiplikativen Überlagerung unterstellen, daß sowohl die zu überlagernde Variable  $Y$  als auch die Überlagerungsvariable  $U$  nichtnegativ sind, um zu garantieren, daß die anonymisierte Variable  $Y^a$  ebenfalls nichtnegativ ist. Dieser Aspekt wird hier nicht im Einzelnen berücksichtigt.

Für den eindimensionalen Fall ergibt sich hier

$$Y^a = Y U \quad (6-7)$$

wobei  $U$  eine **positive** stetige Zufallsvariable mit Erwartungswert 1 und Varianz  $\sigma_u^2 > 0$  ist, die stochastisch unabhängig von  $Y$  erzeugt wird. Alternativ könnte man schreiben:

$$E[U|Y] = 1 \quad \text{und} \quad V[U|Y] = \sigma_u^2 .$$

Es ergibt sich aus (6-7) für den bedingten Erwartungswert

$$E[Y^a|Y] = 1 \cdot Y$$

und damit für den unbedingten Erwartungswert

$$E[Y^a] = \mu_y$$

Ferner erhalten wir für die bedingte Varianz

$$V[Y^a|Y] = Y^2 V[U|Y] = Y^2 \sigma_u^2$$

und damit für die unbedingte Varianz<sup>14</sup>

$$V[Y^a] = \sigma_y^2 + \sigma_u^2 (\mu_y^2 + \sigma_y^2) . \quad (6-8)$$

Falls die Überlagerung mittels der Mischungsverteilung aus Abschnitt 4.1.2 (bzw. mit dem Höhe-Verfahren) erfolgt, ergibt sich - im eindimensionalen Fall - für die Varianz der überlagerten Variablen

$$V[Y^a] = \sigma_y^2 + (\sigma_m^2 + \delta_m^2) (\mu_y^2 + \sigma_y^2) \quad (6-9)$$

wobei  $\delta_m$  der Mittelwertparameter und  $\sigma_m^2$  der Varianzparameter aus der Mischungsverteilung sind, siehe (4-4) und (4-2).

## 6.3 Multiplikative Überlagerung in multivariaten Verteilungen

Im multivariaten Fall sind der sogenannte 'skalare' bzw. 'konstante' Fall und der Fall der 'eigentlichen multivariaten' Überlagerung zu unterscheiden. Im ersten Fall werden alle

---

<sup>14</sup>Es wird die Formel für bedingte und unbedingte Varianzen verwendet:

$$V[Y^a] = E_y\{V[Y^a|Y]\} + V_y\{E[Y^a|Y]\} .$$



Komponenten des Merkmals-Vektors mit **demselden Faktor** multiplikativ überlagert. Dies hat den Vorteil, daß bei der Bildung von Quotienten dieser Faktor tendenziell (siehe jedoch Genaueres in Abschnitt 7) eliminiert wird, d.h. daß die Quotienten durch diese Art der multiplikativen Überlagerung nicht (stark) verändert werden!! Formal ergibt sich in diesem Fall

$$\mathbf{Y}^a = U \mathbf{Y} \quad (\text{konstante multiplikative Überlagerung}) \quad (6-10)$$

Dabei soll für die (nichtnegative) Störvariable  $U$  gelten:

$$E[U|\mathbf{Y}] = 1 \quad \text{und} \quad V[U|\mathbf{Y}] = \sigma_u^2.$$

Im Fall der eigentlichen multivariaten Überlagerung erhalten wir dagegen

$$\mathbf{Y}^a = \mathcal{U} \mathbf{Y} \quad (\text{eigentliche multiplikative Überlagerung}) \quad (6-11)$$

wobei  $\mathcal{U}$  eine **positive Matrix** mit  $r$  Spalten und  $r$  Zeilen ist, die im allgemeinen<sup>15</sup> die Komponenten der Störterme auf der Diagonalen haben wird, d.h.

$$\mathcal{U} = \begin{pmatrix} U_1 & & & & \\ & U_2 & & & \\ & & \ddots & & \\ & & & U_{k-1} & \\ & & & & U_k \end{pmatrix}.$$

In beiden Fällen ist wieder die stochastische Unabhängigkeit der Störvariablen von den Originalvariablen zu unterstellen.

Eine spezielle Formulierung der multiplikativen Überlagerung, bei der jede Variable jeweils mit einer individuellen Störvariablen überlagert wird, läßt sich mittels des Hadamard-Produktes erreichen. Dieses Produkt formuliert die elementweise Multiplikation der Vektoren  $\mathbf{Y}$  und  $\mathbf{U}$  bzw. der Matrizen  $\mathcal{Y}$  und  $\mathcal{U}$ , wobei beide Vektoren bzw. beide Matrizen identische Dimensionen besitzen.<sup>16</sup> In diesem Fall gilt

$$\mathbf{Y}^a = \mathbf{U} \odot \mathbf{Y} \quad (\text{variablenspezifische multiplikative Überlagerung bei Vektoren}) \quad (6-12)$$

bzw.

$$\mathcal{Y}^a = \mathcal{U} \odot \mathcal{Y} \quad (\text{variablenspezifische multiplikative Überlagerung bei Matrizen}) \quad (6-13)$$

### 6.3.1 Überlagerung mit Hilfe eines konstanten Faktors

Im "konstanten" Fall ergibt sich aus (6-10) für den bedingten Erwartungswert

$$E[\mathbf{Y}^a|\mathbf{Y}] = 1 \cdot \mathbf{Y}$$

<sup>15</sup>Es sind aber auch allgemeine Spezifikationen denkbar. Insbesondere lassen sich anonymisierte Variablen als Linearkombinationen von überlagerten Originalvariablen bilden. Beispielsweise kann man

$$\mathcal{U} \mathbf{Y} = \begin{pmatrix} U_{11}Y_1 + U_{12}Y_2 + \dots + U_{1r}Y_r \\ U_{21}Y_1 + U_{22}Y_2 + \dots + U_{2r}Y_r \\ \vdots \\ U_{r1}Y_1 + U_{r2}Y_2 + \dots + U_{rr}Y_r \end{pmatrix}$$

betrachten, wobei  $\mathcal{U}$  eine nichtnegative stochastische  $(r \times r)$ -Matrix ist.

<sup>16</sup>Siehe Marshall und Olkin (1979 Seite 257-8) für einige Eigenschaften.

und damit für den unbedingten Erwartungswert

$$E[\mathbf{Y}^a] = \boldsymbol{\mu}_y$$

Ferner erhalten wir für die bedingte Varianz-Kovarianz-Matrix folgendes: Für die Varianz der einzelnen Komponenten ergibt sich entsprechend dem univariaten Fall

$$V[Y_k^a | \mathbf{Y}] = \sigma_u^2 Y_k^2 \text{ für alle } k$$

sowie

$$\text{cov}[Y_j^a, Y_k^a | \mathbf{Y}] = \sigma_u^2 Y_j Y_k \text{ für alle } j, k \text{ } j \neq k .$$

Insgesamt gilt also für die bedingte Kovarianzmatrix

$$\text{cov}[\mathbf{Y}^a | \mathbf{Y}] = \sigma_u^2 \mathbf{Y} \mathbf{Y}' .$$

Zur Berechnung der unbedingten Kovarianzmatrix verwenden wir die Formel die Formel für bedingte Erwartungswerte und Kovarianzen und erhalten

$$\begin{aligned} \text{cov}[\mathbf{Y}^a] &= E[\text{cov}[\mathbf{Y}^a | \mathbf{Y}]] + \text{cov}[E[\mathbf{Y}^a | \mathbf{Y}]] = \sigma_u^2 E[\mathbf{Y} \mathbf{Y}'] + \text{cov}[\mathbf{Y}] \\ &= \sigma_u^2 \{ \text{cov}[\mathbf{Y}] + \boldsymbol{\mu}_y \boldsymbol{\mu}_y' \} + \text{cov}[\mathbf{Y}] \end{aligned} \quad (6-14)$$

mit

$$\sigma_u^2 = \sigma_m^2 + \delta_m^2 .$$

Man beachte die Entsprechung zu (6-8) für den eindimensionalen Fall.

### 6.3.2 Überlagerung mit Hilfe unkorrelierter Störvariablen

Eine komplexere Formel ergibt sich für den Fall der "eigentlichen multiplikativen" Überlagerung. Dabei spielt offensichtlich auch die Annahme über die Varianzen und Kovarianzen der Matrix  $\mathcal{U}$  eine Rolle. Wir untersuchen zunächst den einfachsten Fall, in dem die Matrix  $\mathcal{U}$  diagonal ist und alle  $U_k$  Erwartungswert 1 und identische Varianz  $\sigma_u^2$  besitzen, ferner miteinander unkorreliert sind. Man beachte, daß jetzt

$$Y_j^a = U_j Y_j \quad , \quad j = 1, \dots, r ,$$

gilt.<sup>17</sup> Aus (6-11) ergibt sich für den bedingten Erwartungswert

$$E[\mathbf{Y}^a | \mathbf{Y}] = \mathbf{1} \cdot \mathbf{Y}$$

und damit für den unbedingten Erwartungswert

$$E[\mathbf{Y}^a] = \boldsymbol{\mu}_y$$

Ferner erhalten wir für die bedingte Varianz-Kovarianz-Matrix folgendes: Für die Varianz der einzelnen Komponenten ergibt sich entsprechend dem univariaten Fall

$$V[Y_k | \mathbf{Y}] = \sigma_u^2 Y_k^2 \text{ für alle } k$$

---

<sup>17</sup>Diese Operation würde man auch durch das Hadamard-Produkt (6-12) der beiden Vektoren  $\mathbf{Y}$  und  $\mathbf{U}$  erreichen.

sowie

$$\text{cov}[Y_j^a, Y_k^a | \mathbf{Y}] = \text{cov}[U_j Y_j, U_k Y_k | \mathbf{Y}] = \text{cov}[U_j, U_k] Y_j Y_k = 0 \text{ für alle } j, k, j \neq k,$$

weil die  $U_j$  unkorreliert sind. Insgesamt gilt also für die bedingte Kovarianzmatrix eine Diagonalmatrix (im Unterschied zum konstanten Fall!!):

$$\text{cov}[\mathbf{Y}^a | \mathbf{Y}] = \sigma_u^2 \begin{pmatrix} Y_1^2 & & & & \\ & Y_2^2 & & & \\ & & \ddots & & \\ & & & Y_{k-1}^2 & \\ & & & & Y_k^2 \end{pmatrix}.$$

Zur Berechnung der unbedingten Kovarianzmatrix verwenden wir wieder

$$\text{cov}[\mathbf{Y}^a] = E[\text{cov}[\mathbf{Y}^a | \mathbf{Y}]] + \text{cov}[E[\mathbf{Y}^a | \mathbf{Y}]]$$

und erhalten in diesem Fall

$$\begin{aligned} \text{cov}[\mathbf{Y}^a] &= \sigma_u^2 E \left[ \begin{pmatrix} Y_1^2 & & & & \\ & Y_2^2 & & & \\ & & \ddots & & \\ & & & Y_{k-1}^2 & \\ & & & & Y_k^2 \end{pmatrix} \right] + \text{cov}[\mathbf{Y}] \\ &= \sigma_u^2 \begin{pmatrix} \sigma_{y(1)}^2 + \mu_{y(1)}^2 & & & & \\ & \sigma_{y(2)}^2 + \mu_{y(2)}^2 & & & \\ & & \ddots & & \\ & & & \sigma_{y(k-1)}^2 + \mu_{y(k-1)}^2 & \\ & & & & \sigma_{y(k)}^2 + \mu_{y(k)}^2 \end{pmatrix} \\ &\quad + \text{cov}[\mathbf{Y}] \end{aligned} \tag{6-15}$$

mit

$$\sigma_u^2 = \sigma_m^2 + \delta_m^2.$$

Man beachte auch hier die Entsprechung zu (6-8) im eindimensionalen Fall.

Demnach stimmen die Varianzen des "eentlichen" Falls mit den Varianzen des "konstanten" Falls überein, nicht aber die Kovarianzen. Für diese läßt sich eine Abschätzung wie folgt vornehmen: Falls alle  $r$  Merkmale positiv miteinander korreliert sind und alle Erwartungswerte positiv sind (was bei nichtnegativen Merkmalen automatisch der Fall ist), ergeben sich für den eigentlichen Fall **kleinere** Kovarianzen als im konstanten Fall. Man vergleiche (6-14) mit (6-15). Wegen der identischen Varianzen bedeutet dies eine geringere Korrelation im Fall der "eentlichen" Überlagerung.

### 6.3.3 Überlagerung mit Hilfe der Mischungsverteilung

Wir haben in Abschnitt 4.2 und speziell für multiplikative Überlagerung im Unterabschnitt 4.2.2 festgestellt, daß die Annahme der Unkorreliertheit nicht mit der Mischungsverteilung

verträglich ist, wenn man sie wie dort beschrieben für die Überlagerung einsetzt. Deshalb müssen wir, wenn wir eine Mischungsverteilung für die stochastische Überlagerung unterstellen, die Annahme aufgeben, daß die Überlagerungsfaktoren miteinander unkorreliert sind. Andererseits interessiert uns vor allem die Auswirkung der speziellen Parametrisierung, die wir als Höhenverfahren bezeichnen: Alle Merkmale werden mit demselben Zuschlag versehen, d.h. gemäß (5-13) soll gelten:

$$U_j = 1 + \delta D + \varepsilon_j \quad , j = 1, \dots, r,$$

Ferner sollen die Störterme  $\varepsilon_j$  den "klassischen" Annahmen genügen:

$$E[\varepsilon] = \mathbf{0} \quad , \quad cov[\varepsilon] = \sigma_\varepsilon^2 \mathbf{I}$$

Siehe dazu (5-3). Dies führt zu folgender Kovarianzmatrix:

$$cov[\mathbf{U}] = \sigma_\varepsilon^2 \mathbf{I} + \delta^2 \boldsymbol{\mu}' = \frac{1}{\sigma_\varepsilon^2 + \delta^2} ((1 - \rho) \mathbf{I} + \rho \boldsymbol{\mu}' \boldsymbol{\mu}'),$$

wobei  $\rho$  durch

$$\rho = \frac{\delta^2}{\delta^2 + \sigma_\varepsilon^2}$$

gegeben ist. Siehe (5-18). Demnach gilt für Varianzen und Kovarianzen:

$$V[U_j] = \sigma_\varepsilon^2 + \delta^2 \quad \text{und} \quad cov[U_j, U_k] = \delta^2 \quad . \quad (6-16)$$

Dies entspricht der speziellen multiplikativen Überlagerung, wie sie in Abschnitt 4.2.2 beschrieben wurde.<sup>18</sup> Für diesen speziellen Fall wollen wir deshalb im folgenden Erwartungsvektor und Kovarianzmatrix des Vektors  $\mathbf{Y}^a$  bestimmen.

Es gilt wieder

$$Y_j^a = U_j Y_j \quad , j = 1, \dots, r,$$

Deshalb ergibt sich aus (6-11) für den bedingten Erwartungswert

$$E[\mathbf{Y}^a | \mathbf{Y}] = \mathbf{1} \cdot \mathbf{Y}$$

und damit für den unbedingten Erwartungswert

$$E[\mathbf{Y}^a] = \boldsymbol{\mu}_y$$

Ferner erhalten wir für die bedingte Varianz-Kovarianz-Matrix folgendes: Für die bedingte Varianz der einzelnen Komponenten ergibt sich

$$V[Y_j | \mathbf{Y}] = V[U_j] \cdot Y_j^2 \quad \text{für alle } j$$

und für die bedingten Kovarianzen erhalten wir

$$cov[Y_j^a, Y_k^a | \mathbf{Y}] = cov[U_j Y_j, U_k Y_k | \mathbf{Y}] = cov[U_j, U_k] \cdot Y_j Y_k \quad \text{für alle } j, k \quad j \neq k,$$

Insgesamt ergibt sich also für die bedingte Kovarianzmatrix :

$$cov[\mathbf{Y}^a | \mathbf{Y}] = cov[\mathbf{U}] \odot \mathbf{Y} \mathbf{Y}' = (\sigma_\varepsilon^2 \mathbf{I} + \delta^2 \boldsymbol{\mu}' \boldsymbol{\mu}') \odot \mathbf{Y} \mathbf{Y}'$$

<sup>18</sup>Dort wurden die etwas abweichenden Bezeichnungen  $\sigma_m^2$  und  $\delta_m$  gewählt.

Zur Berechnung der unbedingten Kovarianzmatrix verwenden wir wieder

$$\text{cov}[\mathbf{Y}^a] = E[\text{cov}[\mathbf{Y}^a|\mathbf{Y}]] + \text{cov}[E[\mathbf{Y}^a|\mathbf{Y}]]$$

und erhalten in diesem Fall

$$\text{cov}[\mathbf{Y}^a] = (\sigma_\varepsilon^2 \mathbf{I} + \delta^2 \boldsymbol{\mu}'\boldsymbol{\mu}) \odot (\text{cov}[\mathbf{Y}] + \boldsymbol{\mu}_y \boldsymbol{\mu}_y') + \text{cov}[\mathbf{Y}] \quad (6-17)$$

Alternativ könnte man in dieser Formel die Parameter  $\sigma_m^2$  und  $\delta_m$  verwenden. Die Entsprechung zu (6-8) im eindimensionalen Fall ist hier nur noch bedingt gegeben.

### 6.3.4 Zusammenfassung

Die Ergebnisse der drei Unterabschnitte von Abschnitt 6.3 lassen sich wie folgt zusammenfassen: In allen drei Fällen wird auch mittels  $\mathbf{Y}^a$  der Mittelwertvektor erwartungstreu geschätzt. Für die Kovarianzmatrix von  $\mathbf{Y}^a$  ergibt sich in allen drei Fällen eine Kovarianzmatrix mit additiver Struktur: Zu der Kovarianzmatrix  $\text{cov}[\mathbf{Y}]$  wird eine zweite Matrix addiert, in der die ersten und zweiten Momente von  $\mathbf{Y}$  sowie die Parameter der Überlagerungsvariablen eine Rolle spielen. In den ersten beiden Fällen ist dies nur die Varianz von  $U_j$ , im Fall der Mischungsverteilung treten auch die Kovarianzen hinzu. In allen drei Fällen tritt also eine zweite positiv semi-definite Matrix zu der Matrix  $\text{cov}[\mathbf{Y}]$ , was den (unterschiedlichen) Effizienzverlust charakterisiert.

## 6.4 Korrelation bei multiplikativer Überlagerung

Es soll nun die Auswirkung der Überlagerung auf den Korrelationskoeffizienten eingehender untersucht werden. Dabei konzentrieren wir uns auf die Resultate aus Abschnitt 6.3.3, in dem Mischungsverteilungen vom Höhne-Typ verwendet werden. Insbesondere ist hier die daraus resultierende Formel (6-17) für die Kovarianzmatrix relevant. Aus dieser Formel ergibt sich (unter der Annahme konstanter Korrelation  $\rho$  für alle Paare)

$$\begin{aligned} V[Y_j^a] &= (\delta^2 + \sigma_\varepsilon^2) (\sigma_j^2 + \mu_j^2) + \sigma_j^2 \\ \text{cov}[Y_j^a, Y_k^a] &= \delta^2 (\rho \sigma_j \sigma_k + \mu_j \mu_k) + \rho \sigma_j \sigma_k \end{aligned} \quad (6-18)$$

sowie

$$\text{corr}[Y_j^a, Y_k^a] = \frac{\rho (1 + \delta^2) \sigma_j \sigma_k + \delta^2 \mu_j \mu_k}{\sqrt{\{(\delta^2 + \sigma_\varepsilon^2) (\sigma_j^2 + \mu_j^2) + \sigma_j^2\} \{(\delta^2 + \sigma_\varepsilon^2) (\sigma_k^2 + \mu_k^2) + \sigma_k^2\}}} \quad (6-19)$$

Aus dieser Formel folgt, daß die Korrelation von  $Y_j^a$  und  $Y_k^a$  eine monotone Beziehung zur Korrelation der entsprechenden Originalvariablen aufweist: Je größer der Wert des Korrelationskoeffizienten  $\rho$  auf der Skala zwischen -1 und +1 ist, desto größer ist auch der Wert des Korrelationskoeffizienten für die überlagerten Variablen. Allerdings ist im Fall  $\rho = 0$  (Originalvariablen unkorreliert) die Korrelation der überlagerten Variablen im Allgemeinen ungleich Null und hängt von den Vorzeichen der beiden Erwartungswerte  $\mu_j$  und  $\mu_k$  ab. Setzt man andererseits den Zählerausdruck von (6-19) gleich Null, so ergibt sich daraus

$$\rho = - \frac{\delta^2 \mu_j \mu_k}{(1 + \delta^2) \sigma_j \sigma_k} ,$$

d.h. die Originalvariablen werden nur dann ebenfalls Null-Korrelation aufweisen, wenn mindestens einer der Erwartungswerte identisch Null ist, was sicher ein singuläres Ereignis ist.

Aus der Formel (6-19) sieht man auch, daß nur die Veränderung des **Verhältnisses** der Erwartungswerte und Varianzen für die Korrelation der anonymisierten Variablen eine Rolle spielt. Denn man kann dafür auch schreiben:

$$\text{corr}[Y_j^a, Y_k^a] = \frac{\rho(1 + \delta^2)\gamma_j\gamma_k + \delta^2}{\sqrt{\left\{(\delta^2 + \sigma_\varepsilon^2)(\gamma_j^2 + 1) + \gamma_j^2\right\} \left\{(\delta^2 + \sigma_\varepsilon^2)(\gamma_k^2 + 1) + \gamma_k^2\right\}}} \quad (6-20)$$

mit

$$\gamma_j \equiv \frac{\sigma_j}{\mu_j} .$$

Falls  $\mu_j > 0$  gilt, bezeichnet man  $\gamma_j$  als Variationskoeffizienten.

Da es für das Vorzeichen der Korrelation nur auf den Zähler ankommt, können wir schreiben:

$$\rho > -\frac{\delta^2}{(1 + \delta^2)} \frac{\mu_j}{\sigma_j} \frac{\mu_k}{\sigma_k} \implies \rho^a > 0 ,$$

d.h. es hängt von nur von  $\delta$  sowie  $\gamma_j$  und  $\gamma_k$  ab, ob beide Korrelationskoeffizienten das gleiche oder unterschiedliche Vorzeichen besitzen. Da  $\delta$  in der Anonymisierungspraxis stets kleiner als 1 und üblicherweise einen Wert von etwa 0,10 aufweist (und damit  $\delta^2 = 0,0001$  gilt), ist die Abweichung im Vorzeichen zwischen den beiden Korrelationskoeffizienten allerdings in der Praxis nicht sehr häufig!

Ferner ergeben sich interessante Resultate für die Fälle, daß die Originalvariablen exakt positiv bzw. exakt negativ korreliert sind. Im Fall  $\rho = 1$  erhalten wir aus (6-19)

$$\text{corr}[Y_j^a, Y_k^a]_{\rho=1} = \frac{(1 + \delta^2)\sigma_j\sigma_k + \delta^2\mu_j\mu_k}{\sqrt{\left\{(\delta^2 + \sigma_\varepsilon^2)(\sigma_j^2 + \mu_j^2) + \sigma_j^2\right\} \left\{(\delta^2 + \sigma_\varepsilon^2)(\sigma_k^2 + \mu_k^2) + \sigma_k^2\right\}}} \quad (6-21)$$

und im Fall  $\rho = -1$

$$\text{corr}[Y_j^a, Y_k^a]_{\rho=-1} = \frac{\delta^2\mu_j\mu_k - (1 + \delta^2)\sigma_j\sigma_k}{\sqrt{\left\{(\delta^2 + \sigma_\varepsilon^2)(\sigma_j^2 + \mu_j^2) + \sigma_j^2\right\} \left\{(\delta^2 + \sigma_\varepsilon^2)(\sigma_k^2 + \mu_k^2) + \sigma_k^2\right\}}} \quad (6-22)$$

Vor allem die Formel (6-22) besagt, daß die Korrelation der überlagerten Variablen durchaus positiv sein kann, obwohl die Originalvariablen eine exakte negative Korrelation aufweisen! Dies wird tendenziell dann der Fall sein, wenn die Erwartungswerte im Verhältnis zu den Standardabweichungen groß sind.

Ferner wird bei exakter **positiver** Korrelation der Originalvariablen die Korrelation der überlagerten Variablen stets kleiner als 1 sein. Dies sieht man deutlicher, wenn man

$$\mu_k = \alpha\mu_j \quad \text{und} \quad \sigma_k = \beta\sigma_j$$

setzt und dann die rechte Seite von (6-21) wie folgt schreibt:

$$\frac{(1 + \delta^2)\beta\sigma_j^2 + \delta^2\alpha\mu_j^2}{\sqrt{\left\{(1 + \delta^2 + \sigma_\varepsilon^2)\sigma_j^2 + (\delta^2 + \sigma_\varepsilon^2)\mu_j^2\right\} \left\{(1 + \delta^2 + \sigma_\varepsilon^2)\beta^2\sigma_j^2 + (\delta^2 + \sigma_\varepsilon^2)\alpha^2\mu_j^2\right\}}}$$

Für  $\alpha = \beta = 1$  ist unmittelbar zu sehen, daß für  $\sigma_\varepsilon^2 > 0$  der Ausdruck kleiner als 1 ist.<sup>19</sup>

## 7 Einfluß der Überlagerung auf Quotienten

### 7.1 Einleitende Bemerkungen

Das Verfahren, das Höhne vorschlug, soll die - sinnvolle - gemeinsame Anonymisierung mehrerer Merkmale ermöglichen. Das bedeutet insbesondere, daß die (proportionale) Beziehung zwischen verschiedenen Merkmalen nicht zu sehr verändert wird. Statistisch gesehen bedeutet dies, daß der Erwartungswert der Zufallsvariablen

$$Z = \frac{Y_1}{Y_2} \quad \text{mit } W\{Y_2 > 0\} = 1$$

nicht zu sehr vom Erwartungswert der Zufallsvariablen

$$Z^a = \frac{Y_1^a}{Y_2^a} \quad \text{mit } W\{Y_2^a > 0\} = 1$$

abweicht.<sup>20</sup>

Dabei ist zu beachten, daß ganz allgemein der Erwartungswert von  $Z$  nicht gleich dem Verhältnis der Erwartungswerte der beiden Zufallsvariablen ist, d.h. es gilt ganz allgemein

$$E[Z] \neq \frac{E[Y_1]}{E[Y_2]}$$

und nur im Fall der **stochastischen Unabhängigkeit** von  $Y_1$  und  $Y_2$  läßt sich der Erwartungswert wie folgt schreiben:

$$E[Z] = E[Y_1] E\left[\frac{1}{Y_2}\right],$$

was allerdings meist auch nicht sehr gut analysierbar ist.<sup>21</sup>

Allerdings läßt sich - für den allgemeinen Fall - folgende Approximation angeben:<sup>22</sup>

$$E\left[\frac{Y_1}{Y_2}\right] \approx \frac{E[Y_1]}{E[Y_2]} - \frac{1}{(E[Y_2])^2} \text{cov}[Y_1, Y_2] + \frac{E[Y_1]}{(E[Y_2])^3} V[Y_2] \quad (7-1)$$

Demnach ist der Erwartungswert von  $Z$  nur annähernd gleich dem Verhältnis der beiden merkmals-spezifischen Erwartungswerte! Die Abweichung hängt einerseits davon ab, ob die beiden Merkmale miteinander korreliert sind. Man beachte, daß dieser Term mit negativem

<sup>19</sup>Beweis für den allgemeinen Fall steht noch aus!

<sup>20</sup>Man könnte auch die stärkere Forderung aufstellen, daß die Verteilung von  $Z$  nicht zu sehr von der Verteilung von  $Z^a$  abweichen soll.

<sup>21</sup>Von Nutzen ist oft die Jensensche Ungleichung, nach der ganz allgemein für eine nichtnegative Zufallsvariable  $W$  gilt:

$$E\left[\frac{1}{W}\right] \geq \frac{1}{E[W]}.$$

Siehe z.B. Mood Graybill Boes 1974 S. 72.

<sup>22</sup>Mood Graybill Boes (1974 S. 181). Eine umfassende Analyse der Verteilung von Quotienten liefert das Buch von Springer (1979). Diesen Hinweis verdanke ich Robert Jung.

Vorzeichen eingeht. Das bedeutet, daß eine positive Korrelation sich reduzierend auf den Erwartungswert von  $Z$  auswirkt. Falls die Korrelation Null ist, fällt der zweite Term auf der rechten Seite weg! Außerdem sind Größenordnung von Erwartungswert und Varianz der Nenner-Variablen  $Y_2$  für das Ausmaß der Abweichung verantwortlich!

Wir haben in Abschnitt 5 gezeigt, daß der Ansatz von Höhne mit der Verwendung der Mischungsverteilung äquivalent ist. Deshalb werde ich im folgenden die Formulierung aus dem genannten Abschnitt bei der Beantwortung der Frage verwenden, welche Beziehung zwischen  $Z$  und  $Z^a$  bei stochastischer Überlagerung mit einer Mischungsverteilung besteht, weil sie anschaulicher ist als die direkte Darstellung der Mischungsverteilung in den Abschnitten 2 und 3. Dabei betrachten wir im folgenden nur den für unser Projekt relevanten Fall der **multiplikativen Überlagerung**.

## 7.2 Das Verhältnis von $Y_1$ zu $Y_2$ bei multiplikativer Überlagerung

Im Folgenden betrachten wir die beiden Zufallsvariablen bzw. Merkmale  $Y_1$  und  $Y_2$ , die jeweils multiplikativ mit den Faktoren

$$U_j = (1 + \delta_j D_j) + \varepsilon_j, \quad j = 1, 2, \quad (7-2)$$

überlagert werden, d.h. die anonymisierten Variablen sind durch

$$Y_j^a = Y_j \cdot U_j = Y_j \cdot (1 + \delta_j D_j + \varepsilon_j), \quad j = 1, 2, \quad (7-3)$$

gegeben.

Der Vorschlag von Höhne, der bereits im ersten Anonymisierungsprojekt realisiert wurde<sup>23</sup>, läuft darauf hinaus, daß die folgenden Restriktionen eingeführt werden:

$$\delta_1 = \delta_2 = \delta \quad (7-4)$$

und

$$D_1 = D_2 \quad \text{mit Wahrscheinlichkeit } 1. \quad (7-5)$$

Dann ergibt sich für das Verhältnis der anonymisierten Variablen

$$Z^a = \frac{Y_1 \cdot ((1 + \delta D) + \varepsilon_1)}{Y_2 \cdot ((1 + \delta D) + \varepsilon_2)} \quad (7-6)$$

Für den Erwartungswert eines Verhältnisses von Zufallsvariablen, in dem Fall von

$$E \left[ \frac{Y_1 \cdot ((1 + \delta D) + \varepsilon_1)}{Y_2 \cdot ((1 + \delta D) + \varepsilon_2)} \right],$$

ist, wie oben gezeigt, die Analyse komplex. Wir beschränken uns im folgenden auf den Fall, daß  $Y_1$  und  $Y_2$  stochastisch unabhängig sind. Da jedoch die Zufallsvariable  $D$  im Zähler und Nenner von (7-6) steht, sind auch unter dieser Unabhängigkeitsannahme die beiden Terme nicht stochastisch unabhängig. Jedoch können wir den **bedingten Erwartungswert** für gegebenes  $D$  wie folgt bestimmen:

$$\begin{aligned} E[Z^a | D] &= E[Y_1] \cdot ((1 + \delta D)) E \left[ \frac{1}{Y_2 \cdot ((1 + \delta D) + \varepsilon_2)} \right] \\ &= E[Y_1] \cdot ((1 + \delta D)) E \left[ \frac{1}{Y_2} \right] \cdot E \left[ \frac{1}{((1 + \delta D) + \varepsilon_2)} \right]. \quad (7-7) \end{aligned}$$

---

<sup>23</sup>Siehe Ronning et al (2005).



Für den zweiten und dritten Erwartungswert auf der rechten Seite kann man folgende Approximation angeben<sup>24</sup>:

$$E\left[\frac{1}{Y_2}\right] \approx \frac{1}{E[Y_2]} + \frac{1}{(E[Y_2])^3}V[Y_2]$$

sowie

$$E\left[\frac{1}{((1 + \delta D) + \varepsilon_2)}\right] \approx \frac{1}{(1 + \delta D)} + \frac{1}{(1 + \delta D)^3}\sigma_\varepsilon^2$$

Dabei hängt die Verzerrung vom Erwartungswert sowie von der Varianz der jeweiligen Zufallsvariablen ab! Bei der Variablen  $Y_2$  ist dies eine empirische Frage; die Störvariable  $\varepsilon$  wird im allgemeinen bei der Anonymisierung nur eine kleine Varianz aufweisen, was eine nur unbedeutende Verzerrung impliziert.

Für den Erwartungswert in (7-7) ergibt sich damit

$$E[Z^a|D] \approx (1 + \delta D)E[Y_1] \left( \frac{1}{E[Y_2]} + \frac{1}{(E[Y_2])^3}V[Y_2] \right) \left( \frac{1}{(1 + \delta D)} + \frac{1}{(1 + \delta D)^3}\sigma_\varepsilon^2 \right) \quad (7-8)$$

Dafür können wir in grober Approximation auch schreiben:

$$\begin{aligned} E[Z^a|D] &\approx (1 + \delta D)E[Y_1] \left( \frac{1}{E[Y_2]} \right) \left( \frac{1}{(1 + \delta D)} \right) + \mathbf{R} \\ &= \frac{E[Y_1]}{E[Y_2]} + \mathbf{R}, \end{aligned} \quad (7-9)$$

wobei im allgemeinen für den Restterm

$$\mathbf{R} = (1 + \delta D)E[Y_1] \left( \frac{1}{E[Y_2]} \frac{1}{(1 + \delta D)^3}\sigma_\varepsilon^2 + \frac{1}{(E[Y_2])^3}V[Y_2] \frac{1}{(1 + \delta D)} + \frac{1}{(E[Y_2])^3}V[Y_2] \frac{1}{(1 + \delta D)^3}\sigma_\varepsilon^2 \right)$$

unabhängig davon, ob  $D = 1$  oder  $D = -1$  gilt, die Ungleichung

$$\mathbf{R} > 0$$

gelten wird, d.h. der berechnete bedingte Erwartungswert von  $Z^a$  wird im allgemeinen das Verhältnis

$$\frac{E[Y_1]}{E[Y_2]}$$

**überschätzen.** Andererseits kürzt sich der Faktor  $(1 + \delta D)$  umso exakter heraus (und spielt in  $\mathbf{R}$  keine Rolle), je kleiner die Überlagerungs(rest)varianz  $\sigma_\varepsilon^2$  gewählt wird. Siehe (7-7).

Wenn sich der Faktor  $(1 + \delta D)$  herauskürzt, dann ist der bedingte Erwartungswert  $E[Z^a|D]$  nicht mehr von  $D$  abhängig, d.h. dann ist der bedingte Erwartungswert mit dem unbedingten Erwartungswert  $E[Z^a]$  identisch. Die hier präsentierten Überlegungen gelten jedoch nur für den Fall, daß  $Y_1$  und  $Y_2$  stochastisch unabhängig sind. In der Simulationsstudie, die im folgenden Unterabschnitt präsentiert wird, werden wir auch den Fall betrachten, daß die beiden Variablen korreliert sind.

<sup>24</sup>Siehe Mood Graybill Boes 1974 S. 181 sowie Formel (7-1) weiter oben. Man beachte, daß wir im Zähler des Ausdrucks die Konstante 1 stehen haben. Somit ist die Kovarianz zwischen  $Y_2$  und 1 gleich Null!

### 7.3 Simulationsergebnisse

In diesem Abschnitt sollen die zuvor theoretisch analysierten Sachverhalte durch einige numerische - simulierte - Beispiele illustriert werden. Dazu betrachte ich zwei Zufallsvariable  $X$  und  $Y$  sowie deren Verhältnis  $Z$ . Ferner betrachte ich die multiplikativ überlagerten Variablen  $X^a$  und  $Y^a$  sowie das daraus resultierende Verhältnis  $Z^a$ . Siehe dazu die Ausführungen in Abschnitt 7.1.

Für die beiden Ausgangsvariablen nehme ich eine gemeinsame Normalverteilung wie folgt an:

$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim N \left[ \begin{pmatrix} \mu_x \\ \mu_y \end{pmatrix}, \begin{pmatrix} \sigma_x^2 & \rho \sigma_x \sigma_y \\ \rho \sigma_x \sigma_y & \sigma_y^2 \end{pmatrix} \right]$$

Zwecks Anonymisierung werden diese beiden Variablen wie folgt überlagert:

$$\begin{aligned} X^a &= (1 + \delta D + \varepsilon_x) X \\ Y^a &= (1 + \delta D + \varepsilon_y) Y \end{aligned}$$

mit

$$\begin{pmatrix} \varepsilon_x \\ \varepsilon_y \end{pmatrix} \sim N(\mathbf{0}, \sigma_\varepsilon^2 \mathbf{I})$$

Die beiden Annahmen der Normalverteilung sind insofern problematisch, als dadurch nicht garantiert ist, daß (a) die Ausgangsvariablen stets nichtnegativ und (b) die Überlagerungen stets positiv sind. Ich habe dies dadurch vermieden, daß ich die Erwartungswerte  $\mu_x$  und  $\mu_y$  relativ groß gewählt habe.<sup>25</sup>

In den folgenden Beispielen wurde stets

$$\delta = 0.10 \quad \text{und} \quad \sigma_\varepsilon = 0.03$$

gewählt. Dabei werden drei verschiedene Ausgangsszenarios zugrunde gelegt:

**A**  $\mu_x = \mu_y$  ;  $\sigma_x^2 = \sigma_y^2$ , Variationskoeffizienten klein

**B**  $\mu_x = \mu_y$  ;  $\sigma_x^2 = \sigma_y^2$ , Variationskoeffizienten sehr klein

**C**  $\mu_x < \mu_y$  ;  $\sigma_x^2 < \sigma_y^2$ ,

Für alle drei Szenarios wurde die Korrelation wie folgt variiert:  $\rho \in \{-0.999, 0, +0.999\}$ . Alle Beispiele wurden mit  $n = 100$  Beobachtungen und 50 Wiederholungen simuliert. Die Ergebnisse sind in der folgenden Tabelle zusammengefaßt.<sup>26</sup>

<sup>25</sup>Eine formal bessere Lösung wäre, sowohl für die Ausgangsvariablen als auch für die Überlagerungsreste Lognormalverteilungen zu verwenden. Dabei sollten die Ausgangsvariablen im Intervall  $[0, \infty)$  und die Überlagerungen im Intervall  $[-(1 + \delta D), \infty)$  variieren.

<sup>26</sup>Alle Ergebnisse wurden mit den GAUSS-Programmen QUOT01.prg, QUOT02.prg und QUOT03.prg (Laptop Ronning, Februar 2007) erzeugt.

$\mu_x$	$\mu_y$	$\sigma_x$	$\sigma_y$	$\rho$	$\rho^a$	$r^a$	$\frac{\mu_x}{\mu_y}$	$\bar{z}$	$\sigma_{\bar{z}}$	$\bar{z}^a$	$\sigma_{\bar{z}^a}$	bias
Szenario A												
10.00	10.00	2.00	2.00	-0.999	-0.5914	-0.5836	1.0000	1.0961	0.0454	1.0959	0.0453	0.0780
10.00	10.00	2.00	2.00	0.000	0.1948	0.1933	1.0000	1.0385	0.0335	1.0391	0.0337	0.0400
10.00	10.00	2.00	2.00	0.999	0.9810	0.9806	1.0000	1.0001	0.0010	1.0010	0.0039	0.0000
Szenario B												
25.00	25.00	2.00	2.00	-0.999	0.2039	0.2212	1.0000	1.0152	0.0130	1.0177	0.0142	0.0128
25.00	25.00	2.00	2.00	0.000	0.5757	0.5732	1.0000	1.0071	0.0111	1.0074	0.0123	0.0064
25.00	25.00	2.00	2.00	0.999	0.9475	0.9464	1.0000	1.0000	0.0004	1.0021	0.0045	0.0000
Szenario C												
10.00	20.00	2.00	4.00	-0.999	-0.5914	-0.6022	0.5000	0.5461	0.0266	0.5461	0.0264	0.0340
10.00	20.00	2.00	4.00	0.000	0.1948	0.2072	0.5000	0.5257	0.0160	0.5268	0.0160	0.0200
10.00	20.00	2.00	4.00	0.999	0.9810	0.9804	0.5000	0.5001	0.0006	0.5011	0.0020	0.0000
Bemerkungen												
$\rho^a$ gibt den theoretischen Korrelationskoeffizienten gemäß (6-19) an.												
$r^a$ gibt den empirischen Korrelationskoeffizienten über alle Wiederholungen gemittelt an.												
$\bar{z}$ gibt den geschätzten Wert von $Z$ über alle Wiederholungen gemittelt an.												
$\sigma_{\bar{z}}$ gibt die entsprechende Standardabweichung an.												
$\bar{z}^a$ gibt den geschätzten Wert von $Z^a$ über alle Wiederholungen gemittelt an.												
$\sigma_{\bar{z}^a}$ gibt die entsprechende Standardabweichung an.												
"bias" gibt die Summe aus dem zweiten und dritten Summanden auf der rechten Seite von (7-1) an.												

Als erstes fällt auf, daß bei allen drei Szenarios die Korrelation  $\rho^a$  der überlagerten Variablen stark von derjenigen für die Originalvariablen abweicht. Dies ist in Szenario B besonders deutlich; dort ergibt sich ein deutlich positiver Wert für  $\rho^a$ , während für die Originalvariablen  $\rho = -0.999$  gilt! Wenn  $\rho$  positiv ist, nimmt die Diskrepanz ab, ist aber immer noch deutlich sichtbar. Im übrigen werden die Ergebnisse für  $\rho^a$  durch das empirische Maß  $r^a$  bestätigt. Bemerkenswert ist die Gleichheit von  $\rho^a$  in den Szenarios A und C. Hier wirkt sich aus, daß die Variationskoeffizienten für  $X$  und  $Y$  in beiden Fällen identisch sind. Aus (6-19) ergibt sich, daß dann das Maß  $\rho^a$  konstant bleibt.

Wesentlich ist ferner die Beobachtung, daß bereits der aus den Originalvariablen gebildete Quotient  $Z$  eine Verzerrung aufweist, die dann stärker ist, wenn die Originalvariablen sehr hohe **negative** Korrelation aufweisen. Diese Verzerrung wird durch die Bias-Formel (7-1) sehr gut abgebildet. Bezüglich des aus den überlagerten Variablen erzeugten Quotienten ergibt sich, daß dieser kaum von dem Quotienten, der aus den Originalvariablen gebildet wird, abweicht. Dies gilt praktisch unabhängig davon, wie groß die Diskrepanz zwischen den Korrelationskoeffizienten  $\rho$  und  $\rho^a$  ist. Die ursprüngliche Motivation für die Verwendung des "Höhne"-Verfahrens wird dadurch insoweit "legitimiert", daß der Quotienten-Bias der anonymisierten Variablen mit dem der Originalvariablen harmoniert. Andere Auswirkungen, vor allem die stark abweichende Korrelation sowie die Konsequenzen für die Schätzung von linearen Modellen (siehe dazu den folgenden Abschnitt), sind allerdings ebenfalls zu bewerten.

## 8 Lineare Modelle mit Fehler in den Variablen aus Mischungsverteilungen

### 8.1 Einleitung

Im Folgenden soll untersucht werden, wie sich fehlerbehaftete Variablen auf die Schätzung von linearen Modellen auswirken. Gegenüber der bisherigen Literatur neu ist die Annahme von kontemporär korrelierten Fehler- bzw. Überlagerungsvariablen.

Wir betrachten das (klassische) multiple lineare Regressionsmodell, das wir wie folgt schreiben:

$$\mathbf{y} = \beta_0 \boldsymbol{\iota} + \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\eta} \quad (8-1)$$

Dabei ist  $\boldsymbol{\iota}$  ein Einsvektor und  $\mathbf{X}$  eine  $(n \times K)$ -Matrix, d.h. wir haben  $n$  Beobachtungen (im Querschnitt) und  $K$  (echte) Regressoren. Für die ersten und zweiten Momente von  $\boldsymbol{\eta}$  soll gelten:

$$E[\boldsymbol{\eta}] = \mathbf{0} \quad \text{und} \quad \text{cov}[\boldsymbol{\eta}] = \sigma_\eta^2 \mathbf{I}. \quad (8-2)$$

Der OLS-Schätzer des Vektors  $\boldsymbol{\beta}$  lautet

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{M}_\iota\mathbf{X})^{-1} \mathbf{X}'\mathbf{M}_\iota\mathbf{y} = \boldsymbol{\beta} + (\mathbf{X}'\mathbf{M}_\iota\mathbf{X})^{-1} \mathbf{X}'\mathbf{M}_\iota\boldsymbol{\eta} \quad (8-3)$$

und seine Kovarianzmatrix ist durch

$$\text{cov}[\hat{\boldsymbol{\beta}}] = \sigma_\eta^2 (\mathbf{X}'\mathbf{M}_\iota\mathbf{X})^{-1} \quad (8-4)$$

gegeben. Dabei ist die idempotente Matrix  $\mathbf{M}_\iota$  durch

$$\mathbf{M}_\iota = \mathbf{I}_n - \frac{1}{n} \boldsymbol{\iota}\boldsymbol{\iota}'$$

gegeben.

Im folgenden sollen die Auswirkungen einer additiven bzw. multiplikativen Überlagerung der Regressormatrix  $\mathbf{X}$  sowie des Vektors  $\mathbf{y}$  auf die Schätzung des Parametervektors  $\boldsymbol{\beta}$  untersucht werden. Dabei beschränken wir uns auf die Auswirkungen auf den naiven Schätzer

$$\hat{\boldsymbol{\beta}}^a = (\mathbf{X}^{a'}\mathbf{M}_\iota\mathbf{X}^a)^{-1} \mathbf{X}^{a'}\mathbf{M}_\iota\mathbf{y}^a \quad (8-5)$$

Für den Fall **unkorrelierter** Fehlervariablen können wir auf Ergebnisse aus Rosemann (2006 Dissertation) zurückgreifen. Dies gilt vor allem für den bisher wenig behandelten multiplikativen Fall. Für additive Überlagerung sind die Ergebnisse ohnehin seit längerem bekannt und können in jedem Ökonometrie-Lehrbuch nachgelesen werden. Beim Fall der korrelierten Fehlervariablen interessiert vor allem die spezielle Überlagerungsstruktur, die wir im Höhne-Verfahren betrachtet haben (siehe Abschnitt 5.2) und die wir im folgenden mit einer flexibleren Spezifikation vergleichen. Allerdings ist diese Spezifikation, die im einzelnen weiter unten dargestellt wird, nicht äquivalent zu einer Mischungsverteilung im Sinne von Abschnitt 3.

Im Fall der additiven Überlagerung gilt

$$\mathbf{y}^a = \mathbf{y} + \mathbf{u} \quad \text{und} \quad \mathbf{X}^a = \mathbf{X} + \mathbf{U} \quad (8-6)$$

im Fall der multiplikativen Überlagerung gilt

$$\mathbf{y}^a = \mathbf{y} \odot \mathbf{u} \quad \text{und} \quad \mathbf{X}^a = \mathbf{X} \odot \mathbf{U} \quad (8-7)$$

Dabei ist  $\mathbf{u}$  ein Zufallsvektor und  $\mathbf{U}$  eine Zufallsmatrix. Man beachte, daß im Gegensatz zu den Abschnitten 1 bis 6 hier eine andere Symbolik verwendet wird.<sup>27</sup> Für (8-7) können wir auch schreiben:

$$\mathbf{y}^a = \mathbf{y} \odot \mathbf{u}_y \quad \text{und} \quad (\mathbf{x}_1^a, \dots, \mathbf{x}_K^a) = (\mathbf{x}_1 \odot \mathbf{u}_1, \dots, \mathbf{x}_K \odot \mathbf{u}_K), \quad (8-8)$$

womit die separate Überlagerung jedes Regressor  $\mathbf{x}_k$  besonders gut analysiert werden kann. Es wird sich allerdings herausstellen, daß eine zeilenweise d.h. beobachtungspunkt-spezifische Darstellung mehr weiterhilft, indem wir (8-6) wie folgt schreiben:

$$\begin{aligned} (\mathbf{y}^a)' &= (y\langle 1 \rangle \cdot u\langle 1 \rangle, y\langle 2 \rangle \cdot u\langle 2 \rangle, \dots, y\langle n-1 \rangle \cdot u\langle n-1 \rangle, y\langle n \rangle \cdot u\langle n \rangle) \\ &\quad \text{und} \\ (\mathbf{X}^a)' &= (\mathbf{x}^a\langle 1 \rangle, \dots, \mathbf{x}^a\langle K \rangle) = (\mathbf{x}\langle 1 \rangle \odot \mathbf{u}\langle 1 \rangle, \dots, \mathbf{x}\langle n \rangle \odot \mathbf{u}\langle n \rangle), \end{aligned} \quad (8-9)$$

Dabei bezeichnet der Index in eckigen Klammern einen bestimmten Beobachtungspunkt. Der Vektor  $\mathbf{x}\langle i \rangle$  ist  $K$ -dimensional und steht - gestürzt - für die  $i$ -te **Zeile (!)** der Matrix  $\mathbf{X}$ . Auch die Vektoren  $\mathbf{u}\langle i \rangle$  sind  $K$ -dimensional und stehen - gestürzt - für die  $i$ -te Zeile der Matrix  $\mathbf{U}$ .

In bestimmten Fällen benötigen wir noch eine zusätzliche Darstellung für die "gemeinsame" (multiplikative) Überlagerung von Regressoren und Regressand: Dabei enthalten die einzelnen Zufallsvektoren  $K$  Elemente, wenn nur die Regressoren überlagert werden und  $K + 1$  Elemente, wenn zusätzlich die abhängige Variable überlagert wird. Dies soll durch die Bezeichnungen  $\mathbf{u}_x$  und  $u_y$  kenntlich machen, die im letzteren Fall gemeinsam den  $(K + 1)$ -dimensionalen Vektor  $\mathbf{u}$  bilden:

$$\mathbf{u} = \begin{pmatrix} \mathbf{u}_x \\ u_y \end{pmatrix}$$

## 8.2 Spezielle und flexible Höhne-Spezifikation

Im folgenden werden wir die spezielle Höhne-Spezifikation (5-13) bzw. (5-14), die einen identischen Zuschlag  $\delta$  für alle  $r = K$  Merkmale (bei ausschließlicher Überlagerung der Regressoren bzw.  $r = K + 1$  bei zusätzlicher Überlagerung auch der abhängigen Variablen) fordert, betrachten, die wir hier für einen bestimmten Beobachtungspunkt  $i$  wie folgt schreiben:

$$\mathbf{u}\langle i \rangle = (1 + \delta D\langle i \rangle) \boldsymbol{\iota} + \boldsymbol{\varepsilon}\langle i \rangle, \quad i = 1, \dots, n. \quad (8-10)$$

wobei die stochastischen Spezifikationen der Zufallsvariablen  $D$  und des Zufallsvektors  $\boldsymbol{\varepsilon}$  für jedes  $i$  durch (5-2) und (5-3) gegeben sind.

Im Gegensatz dazu geht die "flexiblere" Spezifikation davon aus, daß für jedes der  $r = K$  (bzw.  $r = K + 1$ ) Merkmale das Vorzeichen des Zuschlags, nicht aber der Zuschlag selbst über die Merkmale variiert wird. Dies schreiben wir als

$$\mathbf{u}\langle i \rangle = (\boldsymbol{\iota} + \delta \mathbf{D}\langle i \rangle) + \boldsymbol{\varepsilon}\langle i \rangle, \quad i = 1, \dots, n. \quad (8-11)$$

<sup>27</sup>In Abschnitt 3 haben wir  $\mathbf{U}$  als Zufallsvektor eingeführt. Jetzt müssen wir die Überlagerung für alle  $n$  Beobachtungen modellieren, deshalb die geänderte Symbolik.

wobei  $\mathbf{D}$  jetzt ein  $r$ -dimensionaler Zufallsvektor ist, dessen Komponenten jeweils die Spezifikation (5-2) erfüllen und die stochastisch voneinander unabhängig sind. Man kann zeigen, daß Erwartungswerte und Varianzen dieser Spezifikation denen von (8-10) bzw. (5-14) entsprechen. Dagegen sind alle Kovarianzen gleich Null.<sup>28</sup>: Insbesondere gilt für den "symmetrischen" Fall

$$E[u_j] = 1, V[u_j] = \delta^2 + \sigma_\varepsilon^2, cov[u_j, u_k] = 0 \text{ für alle } j, k, j \neq k \quad .$$

### 8.3 Additive Überlagerung im linearen Regressionsmodell

Obwohl die Ergebnisse für den Fall der additiven Überlagerung seit langem bekannt sind, sollen sie hier der Vollständigkeit und auch zum Vergleich mit den Ergebnissen für den multiplikativen Fall kurz vorgestellt werden. Dabei nehme ich wiederholt Bezug auf Ergebnisse in Rosemann (Dissertation 2006), ohne dies an allen Stellen ausdrücklich kenntlich zu machen. Auch übergehe ich die einzelnen Annahmen, unter denen diese (asymptotischen) Ergebnisse abgeleitet werden. Des öfteren wird von der Annahme Gebrauch gemacht, daß

$$\text{plim} \frac{1}{n} \mathbf{X}'\mathbf{M}_l\mathbf{X} = \mathbf{Q}$$

eine feste und reguläre d.h. invertierbare Matrix ist. Damit weiche ich von der üblichen Annahme ab, die die Momentenmatrix  $\mathbf{X}'\mathbf{X}$  betrachtet. Der Unterschied besteht allerdings nur darin, daß ich statt der Originalwerte die Abweichungen vom Mittel betrachte, also zentrierte statt der unzentrierten Momente.

Im Fall der ausschließlichen (additiven) Überlagerung der Regressoren (d.h.  $\mathbf{y}^a = \mathbf{y}$ ) erhalten wir

$$\text{plim} \hat{\boldsymbol{\beta}}^a = (cov[\mathbf{u}] + \mathbf{Q})^{-1} \mathbf{Q} \boldsymbol{\beta} \quad (8-12)$$

wobei  $cov[\mathbf{u}]$  durch (6-6) gegeben ist. Der "naive" Schätzer "unter"schätzt also in diesem all den wahren Wert derart, daß die Norm von  $\text{plim} \hat{\boldsymbol{\beta}}^a$  kleiner ist als die von  $\boldsymbol{\beta}$ .<sup>29</sup> Dies entspricht dem bekannten Ergebnis im Fall der Einfachregression.

Vorausgesetzt daß  $\mathbf{Q}$  und  $cov[\mathbf{u}]$  bekannt sind oder konsistent geschätzt werden können, läßt sich aus diesem Ergebnis ein konsistenter "Korrektur-Schätzer" wie folgt definieren:

$$\hat{\boldsymbol{\beta}}^{a, \text{korr}} = \mathbf{Q}^{-1} (cov[\mathbf{u}] + \mathbf{Q}) \hat{\boldsymbol{\beta}}^a \quad .$$

Dabei ist  $\hat{\boldsymbol{\beta}}^a$  durch (8-5) gegeben. Im Fall der Anonymisierung ist allerdings nur die Regressormatrix  $\mathbf{X}^a$  bekannt, die Matrix  $\mathbf{Q}$  also unbekannt. Da jedoch

$$\text{plim} \frac{1}{n} \mathbf{X}^{a'}\mathbf{M}_l\mathbf{X}^a = cov[\mathbf{x}] + cov[\mathbf{u}] = \mathbf{Q} + cov[\mathbf{u}]$$

gilt, kann man - für bekannte Kovarianzmatrix  $cov[\mathbf{u}]$  - folgende operationale Form des Korrekturschätzers verwenden:

$$\hat{\boldsymbol{\beta}}^{a, \text{korr}} = \left( \frac{1}{n} \mathbf{X}^{a'}\mathbf{M}_l\mathbf{X}^a - cov[\mathbf{u}] \right)^{-1} \left( \frac{1}{n} \mathbf{X}^{a'}\mathbf{M}_l\mathbf{X}^a \right) \hat{\boldsymbol{\beta}}^a \quad . \quad (8-13)$$

<sup>28</sup>Die stochastischen Eigenschaften dieser Spezifikation werden in Appendix B beschrieben! Insbesondere läßt sich zeigen, daß in diesem Fall keine Äquivalenz zu einer Mischungsverteilung besteht.

<sup>29</sup>Man kann zeigen, daß die Eigenwerte von  $(\mathbf{A} + \mathbf{B})^{-1}\mathbf{A}$ ,  $\mathbf{A}$  und  $\mathbf{B}$  positiv definit, stets kleiner als 1 sind, was zu einer "Schrumpfung" (englisch "shrinkage") des Vektors  $\mathbf{x}$  führt, d.h.  $\|(\mathbf{A} + \mathbf{B})^{-1}\mathbf{A} \mathbf{x}\| \leq \|\mathbf{x}\|$ . Siehe z.B. Ronning(1977). Für entsprechende Ergebnisse bezüglich der Matrix  $(\mathbf{A} + \mathbf{B})^{-1} \odot \mathbf{A}$  siehe Marshall und Olkin (1979 S. ).

was der Formel (16.77) bei Rosemann (2006) entspricht.

Es wird im Einzelnen noch zu untersuchen sein, wie sich die beiden unterschiedlichen Kovarianzmatrizen, die aus der "speziellen" bzw. "flexiblen" Spezifikation folgen, auf die Inkonsistenz des "naiven" Schätzers auswirken.<sup>30</sup> Für den speziellen Fall erhalten wir aus Appendix A

$$cov[\mathbf{u}] = \sigma_\varepsilon^2 \mathbf{I} + \mu^2 \boldsymbol{\nu} \boldsymbol{\nu}' = \frac{1}{\sigma_\varepsilon^2 + \mu^2} ((1 - \rho) \mathbf{I} + \rho \boldsymbol{\nu} \boldsymbol{\nu}')$$

wobei  $\rho$  durch

$$\rho = \frac{\mu^2}{\mu^2 + \sigma_\varepsilon^2}$$

gegeben ist. Man beachte, daß diese Korrelation stets positiv sein wird!

Dagegen gilt im Fall der "flexiblen" Spezifikation

$$cov[\mathbf{u}] = (\mu^2 + \sigma_\varepsilon^2) \mathbf{I} .$$

Siehe (B-2) für den "symmetrischen" Spezialfall ( $\alpha = 0, 5$ ). In diesem Fall besteht also keine Korrelation zwischen den einzelnen Komponenten!

Derselbe Wahrscheinlichkeitsgrenzwert ergibt sich, wenn auch die abhängige Variable (additiv) überlagert wird und zwischen den Überlagerungsvariablen keine Korrelation besteht.<sup>31</sup> Rosemann (Dissertation 2006) betrachtet jedoch auch den Fall, in dem Korrelation zwischen dem Vektor  $\mathbf{u}_x$  und der Überlagerung  $u_y$  für die abhängige Variable besteht. Dieser Fall wird in der Lehrbuchliteratur als nicht sehr bedeutsam angesehen. Im Fall der Überlagerung durch multivariate Mischungsverteilungen ist er allerdings wichtig, weil gerade im "speziellen" Höhenverfahren diese Korrelation besteht, wie bereits oben demonstriert. In diesem Fall ergibt sich folgender Wahrscheinlichkeitsgrenzwert:

$$plim \hat{\boldsymbol{\beta}}^a = (cov[\mathbf{u}] + \mathbf{Q})^{-1} (\mathbf{Q} \boldsymbol{\beta} + cov[\mathbf{u}_x, u_y]) \quad (8-14)$$

Dabei ist  $cov[\mathbf{u}_x, u_y]$  ein **Vektor**, der die Kovarianzen von  $\mathbf{u}_x$  mit  $u_y$  enthält. Im "speziellen" Fall hat dieser Vektor folgende Struktur:

$$cov[\mathbf{u}_x, u_y] = \mu^2 \boldsymbol{\nu} . \quad (8-15)$$

Siehe Appendix A. Dagegen ist im Fall der "Flexiblen" Spezifikation dieser Vektor gleich dem Nullvektor. Konsequenz dieses Ergebnisses ist, daß sich unterschiedliche Verzerrungen für beide Spezifikationen ergeben.

Für den Fall, daß Korrelation zwischen den Meßfehlern der Regressoren und des Regressanden bestehen, ergibt sich aus (8-14) folgender Korrektorschätzer:

$$\hat{\boldsymbol{\beta}}^{a, \text{korr}} = \mathbf{Q}^{-1} \left\{ (cov[\mathbf{u}] + \mathbf{Q}) \hat{\boldsymbol{\beta}}^a - cov[\mathbf{u}_x, u_y] \right\} .$$

Dabei ist  $\hat{\boldsymbol{\beta}}^a$  durch (8-5) gegeben. Entsprechend (8-13) erhalten wir in diesem Fall als operationalen Korrektorschätzer

$$\hat{\boldsymbol{\beta}}^{a, \text{korr}} = \left( \frac{1}{n} \mathbf{X}^{\mathbf{a}'} \mathbf{M}_l \mathbf{X}^{\mathbf{a}} - cov[\mathbf{u}] \right)^{-1} \left( \frac{1}{n} \mathbf{X}^{\mathbf{a}'} \mathbf{M}_l \mathbf{X}^{\mathbf{a}} \hat{\boldsymbol{\beta}}^a + cov[\mathbf{u}_x, u_y] \right) , \quad (8-16)$$

<sup>30</sup>Vergleiche Norm von  $(\mathbf{A} + \mathbf{B})^{-1} \mathbf{A}$  mit Norm von  $(\mathbf{A} + \mathbf{D})^{-1} \mathbf{A}$  wobei  $\mathbf{D}$  eine positive Diagonalmatrix ist und alle drei Matrizen nichtnegativ definit sind.

<sup>31</sup>Allerdings steigt die Varianz des Schätzers.

was Formel (16.81) bei Rosemann (2006) entspricht, wobei dieser noch ausnutzt, daß

$$\frac{1}{n} \mathbf{X}^{\mathbf{a}'} \mathbf{M}_l \mathbf{X}^{\mathbf{a}} \hat{\boldsymbol{\beta}}^{\mathbf{a}} = \frac{1}{n} \mathbf{X}^{\mathbf{a}'} \mathbf{M}_l \mathbf{y}^{\mathbf{a}}$$

gilt.

Abschließend sei zur Vollständigkeit vermerkt, daß sich nur im Fall, daß **ausschließlich die abhängige Variable** (additiv) überlagert wird, eine konsistente Schätzung ergibt! Siehe z.B. Rosemann (2006 Seite 158-163). Auf Spezialfälle, in denen einzelne Regressoren überlagert werden, gehe ich nicht ein. Auch wird die nicht unwichtige Schätzung der Restvarianz bzw. der Kovarianzmatrix des naiven Schätzers nicht behandelt.

## 8.4 Multiplikative Überlagerung im linearen Regressionsmodell

Wir unterstellen, daß die einzelnen  $n$  Beobachtungspunkte stochastisch unabhängig voneinander überlagert werden, bzw. daß die Meßfehler zwar "kontemporär" (= beobachtungspunktspezifisch) korreliert sein können, nicht aber über verschiedene Beobachtungspunkte hinweg.<sup>32</sup> In diesem Fall ist die spezielle Formulierung der multiplikativen Überlagerung der Regressormatrix aus (8-8) besonders geeignet:

$$(\mathbf{X}^{\mathbf{a}})' = (\mathbf{x}^{\mathbf{a}}\langle 1 \rangle, \dots, \mathbf{x}^{\mathbf{a}}\langle K \rangle) = (\mathbf{x}\langle 1 \rangle \odot \mathbf{u}\langle 1 \rangle, \dots, \mathbf{x}\langle n \rangle \odot \mathbf{u}\langle n \rangle),$$

Die obige Annahme kann dann wie folgt präzisiert werden: Die  $n$  Zufallsvektoren

$$\mathbf{u}\langle 1 \rangle, \mathbf{u}\langle 2 \rangle, \dots, \mathbf{u}\langle n \rangle,$$

sind stochastisch voneinander unabhängig. Dabei enthalten die einzelnen Zufallsvektoren  $K$  Elemente, wenn nur die Regressoren überlagert werden und  $K + 1$  Elemente, wenn zusätzlich die abhängige Variable überlagert wird. Dies werden wir durch die Symbolik  $\mathbf{u}_x$  und  $u_y$  kenntlich machen, die im letzteren Fall gemeinsam den  $(K + 1)$ -dimensionalen Vektor  $\mathbf{u}$  bilden:

$$\mathbf{u} = \begin{pmatrix} \mathbf{u}_x \\ u_y \end{pmatrix} .$$

Bei Rosemann (Dissertation 2006) wird im Einzelnen ausgeführt, wie die Wahrscheinlichkeitsgrenzwerte im Fall der multiplikativen Überlagerung abgeleitet werden. Wir beschränken uns hier darauf, nur die Ergebnisse aufzuführen. Im Fall der ausschließlichen Überlagerung der Regressoren ergibt sich

$$\text{plim } \hat{\boldsymbol{\beta}}^{\mathbf{a}} = (\mathbf{M} \odot \mathbf{Q})^{-1} \mathbf{Q} \boldsymbol{\beta}, \quad (8-17)$$

und im Fall der zusätzlichen Überlagerung auch der abhängigen Variablen ergibt sich

$$\text{plim } \hat{\boldsymbol{\beta}}^{\mathbf{a}} = (\mathbf{M} \odot \mathbf{Q})^{-1} \mathbf{Q} \boldsymbol{\beta} \odot \mathbf{k} . \quad (8-18)$$

Dabei gilt für die beiden Ausdrücke  $\mathbf{M}$  und  $\mathbf{k}$ :

$$\mathbf{M} = E[\mathbf{u}_x \mathbf{u}_x'] = \text{cov}[\mathbf{u}_x] + (E[\mathbf{u}_x])(E[\mathbf{u}_x])' = \sigma_\varepsilon^2 \mathbf{I} + (1 + \delta^2) \boldsymbol{\mu} \boldsymbol{\mu}' \quad (8-19)$$

<sup>32</sup>Diese Annahme wird natürlich auch bei der additiven Überlagerung gemacht, wurde dort aber nicht speziell problematisiert.



und

$$\mathbf{k} = E[u_y \mathbf{u}_x] \quad (8-20)$$

$$= cov[\mathbf{u}_x, u_y] + (E[u_y])(E[\mathbf{u}_x]) \quad (8-21)$$

$$= \begin{cases} (1 + \delta^2) \boldsymbol{\iota} & \text{für die spezielle Spezifikation} \\ \boldsymbol{\iota} & \text{für die flexible Spezifikation} \end{cases} \quad (8-22)$$

Vorausgesetzt daß  $\mathbf{Q}$  und  $cov[\mathbf{u}]$  bekannt sind oder konsistent geschätzt werden können, läßt sich aus dem Ergebnis (8-17) für den Fall der ausschließlichen Überlagerung der Regressoren ein konsistenter "Korrektur-Schätzer" wie folgt definieren:

$$\hat{\boldsymbol{\beta}}^{a, \text{korr}} = \mathbf{Q}^{-1} (cov[\mathbf{u}] \odot \mathbf{Q}) \hat{\boldsymbol{\beta}}^a .$$

Dagegen ergibt sich im Fall, den (8-18) betrachtet, ein konsistenter Korrektorschätzer wie folgt:

$$\hat{\boldsymbol{\beta}}^{a, \text{korr}} = \mathbf{Q}^{-1} (cov[\mathbf{u}] \odot \mathbf{Q}) \mathbf{D}_k^{-1} \hat{\boldsymbol{\beta}}^a .$$

Dabei ist  $\hat{\boldsymbol{\beta}}^a$  jeweils durch (8-5) gegeben und  $\mathbf{D}_k$  ist eine Diagonalmatrix mit den Elementen des Vektors  $\mathbf{k}$  auf der Diagonalen.<sup>33</sup>

Im Fall der Anonymisierung ist allerdings nur die Regressormatrix  $\mathbf{X}^a$  bekannt, die Matrix  $\mathbf{Q}$  also unbekannt. Da jedoch im multiplikativen Fall

$$\text{plim} \frac{1}{n} \mathbf{X}^{a'} \mathbf{M}_l \mathbf{X}^a = cov[\mathbf{u}] \odot \mathbf{Q}$$

(siehe Hwang 1986) und somit

$$\text{plim} \left( \frac{1}{n} \mathbf{X}^{a'} \mathbf{M}_l \mathbf{X}^a \div cov[\mathbf{u}] \right) = \mathbf{Q}$$

und außerdem

$$\text{plim} \frac{1}{n} \mathbf{X}^{a'} \mathbf{M}_l \mathbf{y}^a = \mathbf{Q} \boldsymbol{\beta}$$

(siehe Hwang 1986) gilt, erhalten wir für den Fall, daß nur die Regressoren überlagert sind, einen konsistenten Korrektur-Schätzer als

$$\hat{\boldsymbol{\beta}}^{a, \text{korr}} = (\mathbf{X}^{a'} \mathbf{M}_l \mathbf{X}^a \div cov[\mathbf{u}])^{-1} \mathbf{X}^{a'} \mathbf{M}_l \mathbf{y}^a , \quad (8-23)$$

was der Formel (16.105) bei Rosemann (2006) entspricht. Falls auch die abhängige Variable überlagert wird, ergibt sich

$$\hat{\boldsymbol{\beta}}^{a, \text{korr}} = (\mathbf{X}^{a'} \mathbf{M}_l \mathbf{X}^a \div cov[\mathbf{u}])^{-1} \mathbf{D}_k^{-1} \mathbf{X}^{a'} \mathbf{M}_l \mathbf{y}^a , \quad (8-24)$$

was der Formel (16.134) bei Rosemann (2006) entspricht, wobei zu beachten ist, daß bei der hier betrachteten Überlagerung mittels der speziellen Höhne-Spezifikation eine Korrelation zwischen den Überlagerungen der Regressoren und denen der abhängigen Variablen **immer** gegeben ist.

<sup>33</sup>Hier wird ausgenutzt, daß für zwei Vektoren  $\mathbf{a}$  und  $\mathbf{b}$  gilt:

$$\mathbf{a} \odot \mathbf{b} = \mathbf{D}_a \mathbf{b} = \mathbf{D}_b \mathbf{a} .$$

Im obigen Fall gilt  $\mathbf{a} = \boldsymbol{\beta}$  und  $\mathbf{b} = \mathbf{k}$ , d.h.  $\boldsymbol{\beta} \odot \mathbf{k} = \mathbf{D}_k \boldsymbol{\beta}$ .

## 9 Ergänzende Überlegungen für Paneldaten

### 9.1 Allgemeines

#### 9.1.1 Überlagerungsstrategien

Im Fall von Paneldaten treten an die Stelle von  $r$  verschiedenen Merkmalen bzw. Zufallsvariablen

$$Y_1, Y_2, \dots, Y_{r-1}, Y_r$$

im Querschnitt die  $T \cdot r$  Variablen  $Y_{jt}$  bzw. ausführlich

$$\begin{array}{cccccc} Y_{11} & Y_{12} & \dots & Y_{1,r-1} & Y_{1r} \\ Y_{21} & Y_{22} & \dots & Y_{2,r-1} & Y_{2r} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ Y_{T1} & Y_{T2} & \dots & Y_{T,r-1} & Y_{Tr} \end{array}$$

für  $r$  Merkmale in  $T$  Zeitpunkten. Diese Zufallsvariablen lassen sich auch als  $T$  verschiedene  $r$ -dimensionale Zufallsvektoren

$$\mathbf{Y}_1 \mathbf{Y}_2 \dots, \mathbf{Y}_{T-1}, \mathbf{Y}_T$$

interpretieren.

Es stellt sich die Frage, in welcher Weise die Überlagerung nun vorgenommen werden soll. Wir beschränken uns hier auf die Spezifikation<sup>34</sup>

$$U_{jt} = 1 + \delta D + \varepsilon_{jt} \quad (\text{"alle Merkmale und Zeitpunkte gemeinsam"}) \quad , \quad (9-1)$$

die den multiplikativen Fall betrifft. Wir werden jedoch auch kurz auf den entsprechenden additiven Fall eingehen.

#### 9.1.2 Das einfache lineare Panelmodell

Entsprechend dem Vorgehen von Biørn (1996) (und abweichend von der Darstellung im Abschnitt 8.4 für Querschnittsdaten betrachten wir hier das **einfache** Regressionsmodell

$$y_{it} = \alpha + \beta x_{it} + \eta_{it} \quad i = 1, \dots, n \quad t = 1, \dots, T \quad (9-2)$$

mit einem einzigen Regressor.

---

<sup>34</sup>Denkbar sind folgende Varianten:

$$\begin{array}{ll} U_{jt} = 1 + \delta_{jt} D_{jt} + \varepsilon_{jt} & \text{alle Merkmale und Zeitpunkte getrennt} \\ U_{jt} = 1 + \delta_j D_j + \varepsilon_{jt} & \text{alle Merkmale getrennt} \\ U_{jt} = 1 + \delta_t D_t + \varepsilon_{jt} & \text{alle Zeitpunkte getrennt} \\ U_{jt} = 1 + \delta D + \varepsilon_{jt} & \text{alle Merkmale und Zeitpunkte gemeinsam} \end{array}$$

Weitere Varianten, in denen  $\delta$  und  $D$  nicht identisch behandelt werden, sind denkbar. Die Effekte der einzelnen Überlagerungsvarianten hängen auch davon ab, ob man eine Zeitreihenstruktur unterlegt, beispielsweise einen  $r$ -dimensionalen autoregressiven Prozeß.

Wir werden zunächst wieder die ausschließliche Überlagerung der Regressoren (in diesem Fall des einen Regressors) betrachten und später dann untersuchen, was sich ändert, wenn  $x$  und  $y$  gemeinsam überlagert werden oder nur  $y$  überlagert wird. Siehe dazu die Abschnitte 9.4.5, 9.4.6, 9.4.7 und 9.4.8. Wie wir gesehen haben, hat der zweite Fall **im Querschnitt** sowohl im additiven als auch im multiplikativen Überlagerung Auswirkung auf die Verzerrung. Allerdings müssen wir dafür zu allererst die Überlagerung von Paneldaten (Abschnitt 9.2) darstellen. Dabei werden wir uns wieder auf die multiplikative Überlagerung und dort auf die "spezielle" Höhne-Variante konzentrieren. Ferner werden wir in Abschnitt 9.3 das einfache lineare Panelmodell mit Individualeffekten einführen.

## 9.2 Überlagerung von Paneldaten

### 9.2.1 Symbolik

Bei gemeinsamer Betrachtung von  $r$  Merkmalen im Fall von Paneldaten benötigen wir folgende Symbolik:

$$y_{itj} \quad i = 1, \dots, n, \quad t = 1, \dots, T, \quad j = 1, \dots, r, \quad (9-3)$$

bezeichnet das Merkmal<sup>35</sup>  $y_j$  für Beobachtungseinheit  $i$  im Zeitpunkt  $t$ , wobei insgesamt  $r$  Merkmale betrachtet werden.

### 9.2.2 Additive Überlagerung

Wenn wir die additive Entsprechung zu (9-1) für die Überlagerung verwenden, d.h.

$$u_{itj} = y_{itj} + \mu D_i + \varepsilon_{itj} \quad , \quad (9-4)$$

so ergibt sich für die anonymisierten Merkmale

$$y_{itj}^a = y_{itj} + u_{itj} = y_{itj} + \mu D_i + \varepsilon_{itj} \quad (9-5)$$

Dabei soll  $\varepsilon_{itj}$  konstante Varianz besitzen, ferner sollen alle  $(itj)$ -Kombinationen unkorreliert sein:

$$\begin{aligned} E[\varepsilon_{itj}] &= 0 \\ \text{var}[\varepsilon_{itj}] &= \sigma_\varepsilon^2 \\ \text{cov}[\varepsilon_{itj}, \varepsilon_{hsk}] &= 0, \quad i \neq h, t \neq s, j \neq k \end{aligned} \quad \left\{ \begin{array}{l} i = 1, \dots, n, \\ t = 1, \dots, T, \\ j = 1, \dots, r. \end{array} \right. \quad (9-6)$$

Gleichwohl wird über die Zufallsvariable  $D_i$  Korrelation sowohl über die Merkmale als auch über die Zeitpunkte erzeugt bzw. diese Korrelation wird verstärkt, soweit die Originalvariablen zeitliche Korrelation aufweisen. Entsprechend dem Vorgehen in Abschnitt 6 (Anonymisierung mittels stochastischer Überlagerung) für Querschnittsdaten läßt sich zeigen, daß

$$\text{var}[y_{itj}^a] = \sigma_{tj}^2 + \mu^2 + \sigma_\varepsilon^2 \quad (9-7)$$

<sup>35</sup>Das Symbol  $y$  steht hier für jedes beliebige Merkmal und nicht für die abhängige Variable. Dies entspricht dem Vorgehen in den Abschnitten \*\*\*.

und

$$\text{cov}[y_{itj}^a, y_{hsk}^a] = \begin{cases} \sigma_{jk} + \mu^2 & \text{falls } i = h, t = s, j \neq k \\ \sigma_{ts} + \mu^2 & \text{falls } i = h, t \neq s, j = k \\ 0 & \text{falls } i \neq h \end{cases} \quad (9-8)$$

für alle  $(j, k, t, s)$ -Kombinationen gilt. Dies entspricht (6-6) im Abschnitt 6 (Anonymisierung mittels stochastischer Überlagerung) für Querschnittsdaten. In (9-7) wird jetzt zugelassen, daß die Varianz für Merkmal  $j$  über die Zeit hin variiert und in (9-8) bezeichnet  $\sigma_{jk}$  die (kontemporäre) Kovarianz zwischen den Merkmalen  $j$  und  $k$ , während  $\sigma_{ts}$  die Kovarianz zwischen den Zeitpunkten  $t$  und  $s$  für das  $j$ -te Merkmal bezeichnet.

Durch die Art der Überlagerung ist jetzt zugelassen, daß – für eine bestimmte Beobachtungseinheit  $i$  – Korrelation auch über verschiedene Zeitpunkte ("Wellen") hinweg besteht, selbst wenn die zeitliche Korrelation der Originalvariablen gleich Null ist, d.h. falls  $\sigma_{ts} = 0$ . Formal gilt

$$\text{corr}[y_{itj}^a, y_{hsk}^a] = \begin{cases} \frac{\sigma_{jk} + \mu^2}{\sqrt{(\sigma_{tj}^2 + \mu^2 + \sigma_\varepsilon^2)(\sigma_{tk}^2 + \mu^2 + \sigma_\varepsilon^2)}} & \text{falls } i = h, t = s, j \neq k \\ \frac{\sigma_{ts} + \mu^2}{\sqrt{(\sigma_{tj}^2 + \mu^2 + \sigma_\varepsilon^2)(\sigma_{sj}^2 + \mu^2 + \sigma_\varepsilon^2)}} & \text{falls } i = h, t \neq s, j = k \\ 0 & \text{falls } i \neq h \end{cases} \quad (9-9)$$

Man beachte, daß dabei unterstellt wird, daß zwischen den einzelnen Beobachtungseinheiten keine Korrelation, beispielsweise in Form von Cluster-Effekten, besteht. Das wurde aber beim Querschnittsfall auch so unterstellt.

### 9.2.3 Multiplikative Überlagerung

Im multiplikativen Fall verwenden wir statt (9-5) die Spezifikation

$$y_{itj}^a = y_{itj} (1 + \delta D_i + \varepsilon_{itj}) \quad (9-10)$$

Dabei soll  $\varepsilon_{itj}$  wieder (9-6) erfüllen.

Wegen der Symmetrie bezüglich der Merkmalskomponente  $j$  einerseits und der Zeitkomponente  $t$  andererseits kann man auch hier die Ergebnisse der Auswirkung der stochastischen Überlagerung in Anlehnung an Abschnitt 6 (Anonymisierung mittels stochastischer Überlagerung) für Querschnittsdaten ableiten. Dabei beschränken wir uns auf die Überlagerung mithilfe der Mischungsverteilung, wie sie Abschnitt 6.3.3 (spezielle Höhe-Spezifikation) betrachtet wurde.

Es läßt sich zeigen, daß

$$\text{var}[y_{itj}^a] = \sigma_{tj}^2 + (\delta^2 + \sigma_\varepsilon^2) (\sigma_{tj}^2 + \mu_{tj}^2) \quad (9-11)$$

und

$$\text{cov}[y_{itj}^a, y_{hsk}^a] = \begin{cases} \sigma_{jk} + \delta^2 (\sigma_{jk} + \mu_{tj} \mu_{tk}) & \text{falls } i = h, t = s, j \neq k \\ \sigma_{ts} + \delta^2 (\sigma_{ts} + \mu_{tj} \mu_{sj}) & \text{falls } i = h, t \neq s, j = k \\ 0 & \text{falls } i \neq h \end{cases} \quad (9-12)$$

gilt. Dabei bezeichnet  $\mu_{tj}$  den Erwartungswert für Merkmal  $j$  im Zeitpunkt  $t$ . Varianzen und Kovarianzen entsprechen exakt der Struktur, wie sie in (6-17) für Querschnittsdaten angegeben wurde. Neben die Merkmalskomponenten tritt allerdings nun die Zeitkomponente.

Für die Korrelation zwischen den Variablen als auch zwischen den einzelnen Zeitpunkten erhalten wir jetzt folgende Formel:

$$corr[y_{itj}^a, y_{hsk}^a] = \begin{cases} \frac{\sigma_{jk} + \delta^2 (\sigma_{jk} + \mu_{tj} \mu_{tk})}{\sqrt{\{\sigma_{tj}^2 + (\delta^2 + \sigma_\varepsilon^2) (\sigma_{tj}^2 + \mu_{tj}^2)\} \{\sigma_{tj}^2 + (\delta^2 + \sigma_\varepsilon^2) (\sigma_{tj}^2 + \mu_{tj}^2)\}}} & \text{falls } i = h, t = s, j \neq k \\ \frac{\sigma_{ts} + \delta^2 (\sigma_{ts} + \mu_{tj} \mu_{sj})}{\sqrt{\{\sigma_{tj}^2 + (\delta^2 + \sigma_\varepsilon^2) (\sigma_{tj}^2 + \mu_{tj}^2)\} \{\sigma_{tj}^2 + (\delta^2 + \sigma_\varepsilon^2) (\sigma_{tj}^2 + \mu_{tj}^2)\}}} & \text{falls } i = h, t \neq s, j = k \\ 0 & \text{falls } i \neq h \end{cases} \quad (9-13)$$

Ebenso wie beim additiven Fall wird auch hier unterstellt, daß zwischen den einzelnen Beobachtungseinheiten keine Korrelation, beispielsweise in Form von Cluster-Effekten, besteht. Das wurde aber auch so beim Querschnittsfall unterstellt. Es geht ja darum herauszuarbeiten, welche zusätzliche Wirkung bezüglich Korrelation die Überlagerung hat.

### 9.3 Das einfache lineare Panelmodell mit Individualeffekten

#### 9.3.1 Feste Individualeffekten

Das einfache lineare Panelmodell mit festen Individualeffekten schreiben wir wie folgt:

$$y_{it} = \alpha + \beta x_{it} + \sum_{h=1}^n \tau_h W_{ht} + \eta_{it} \quad , \quad i = 1, \dots, n \quad , \quad t = 1, \dots, T \quad (9-14)$$

wobei

$$W_{ht} = \begin{cases} 1 & \text{falls } i = h \\ 0 & \text{sonst} \end{cases}$$

die individuenspezifische Dummyvariable im Zeitpunkt  $t$  ist und  $\tau_h$  den entsprechenden Effekt bezeichnet. Aus Gründen der Identifikation setzen wir

$$\alpha = 0 \quad .$$

In Matrixschreibweise erhalten wir dann

$$\mathbf{y} = \mathbf{x}\beta + \mathbf{W}\boldsymbol{\tau} + \boldsymbol{\eta} \quad (9-15)$$

wobei  $\mathbf{x}$  ein Vektor und  $\beta$  ein Skalar ist. Als sogenannter "Within-Schätzer" (oder auch "fixed effects estimator") ergibt sich die Teilschätzung nach der Kleinstquadrat-Methode:

$$\hat{\beta}_W = \frac{\mathbf{x}' (\mathbf{I} - \mathbf{W}(\mathbf{W}'\mathbf{W})^{-1}\mathbf{W}) \mathbf{y}}{\mathbf{x}' (\mathbf{I} - \mathbf{W}(\mathbf{W}'\mathbf{W})^{-1}\mathbf{W}) \mathbf{x}} = \frac{\sum_{t=1}^T \sum_{i=1}^n (x_{it} - \bar{x}_{i\bullet})(y_{it} - \bar{y}_{i\bullet})}{\sum_{t=1}^T \sum_{i=1}^n (x_{it} - \bar{x}_{i\bullet})^2} \quad (9-16)$$

mit

$$\bar{x}_{i\bullet} = \frac{1}{T} \sum_{t=1}^T x_{it} \quad , \quad \bar{y}_{i\bullet} = \frac{1}{T} \sum_{t=1}^T y_{it} \quad .$$

Dieser Schätzer hat die üblichen "guten" Eigenschaften des klassischen Regressionsmodells!

### 9.3.2 Stochastische Individualeffekte

Wenn wir statt der festen Effekte zufällige Effekte annehmen, dann erhalten wir

$$y_{it} = \alpha + \beta x_{it} + \tau_i + \eta_{it} \quad , \quad i = 1, \dots, n, \quad t = 1, \dots, T \quad (9-17)$$

wobei die  $\tau_i$  nun Zufallsvariable sind. Weil daraus

$$\bar{y}_{i\bullet} = \bar{x}_{i\bullet} \beta + \tau_i + \bar{\eta}_{i\bullet}$$

und

$$y_{it} - \bar{y}_{i\bullet} = (x_{it} - \bar{x}_{i\bullet})\beta + \eta_{it} - \bar{\eta}_{i\bullet}$$

folgt, spielt der individuenspezifische (zufällige) Effekt bei der Schätzung mittels  $\hat{\beta}_W$  keine Rolle:

$$\hat{\beta}_W = \frac{\sum_{t=1}^T \sum_{i=1}^n (x_{it} - \bar{x}_{i\bullet}) \{(x_{it} - \bar{x}_{i\bullet})\beta + (\eta_{it} - \bar{\eta}_{i\bullet})\}}{\sum_{t=1}^T \sum_{i=1}^n (x_{it} - \bar{x}_{i\bullet})^2} \quad (9-18)$$

d.h.  $\hat{\beta}_W$  ist ein konsistenter Schätzer.<sup>36</sup>

## 9.4 Schätzung des linearen Panelmodells aus anonymisierten Daten

### 9.4.1 Der naive Panelschätzer

Im Fall von anonymisierten Daten verwenden wir statt (9-16) den folgenden "naiven" Schätzer (Within-Schätzer, FE-Schätzer):

$$\hat{\beta}_W^a = \frac{\mathbf{x}^{a'} (\mathbf{I} - \mathbf{W}(\mathbf{W}'\mathbf{W})^{-1}\mathbf{W}) \mathbf{y}^a}{\mathbf{x}^{a'} (\mathbf{I} - \mathbf{W}(\mathbf{W}'\mathbf{W})^{-1}\mathbf{W}) \mathbf{x}^a} = \frac{\sum_{t=1}^T \sum_{i=1}^n (x_{it}^a - \bar{x}_{i\bullet}^a)(y_{it}^a - \bar{y}_{i\bullet}^a)}{\sum_{t=1}^T \sum_{i=1}^n (x_{it}^a - \bar{x}_{i\bullet}^a)^2} \quad (9-19)$$

mit

$$\bar{x}_{i\bullet}^a = \frac{1}{T} \sum_{t=1}^T x_{it}^a \quad , \quad \bar{y}_{i\bullet}^a = \frac{1}{T} \sum_{t=1}^T y_{it}^a \quad .$$

### 9.4.2 Additive Meßfehler allgemein

Biørn (1996 Kapitel 10.2.2) betrachtet ein lineares Panelmodell, in dem die Regressorvariable  $x$  latent ist und nur die meßfehlerbehaftete Variable  $x^a$  betrachtet werden kann<sup>37</sup>, und leitet den Wahrscheinlichkeitsgrenzwert des naiven Panelschätzers (9-19) bei additiver Überlagerung ab. Das Fehlermodell lautet<sup>38</sup>

$$x_{it}^a = x_{it} + u_{it} \quad .$$

<sup>36</sup>Siehe beispielsweise Wooldridge (2002, Kap. 10.5.5), wo der allgemeinere Fall einer verallgemeinerten Schätzung behandelt wird. Wooldridge (2002 Kap. 10.7.2) geht auch darauf ein, daß die "Robustheit gegenüber der Korrelation von  $\tau_i$  und dem Störterm" dann problematisch ist, wenn man den Effekt von zeitlich nur wenig variierenden Regressoren schätzen will.

<sup>37</sup>Aus Gründen der Einheitlichkeit bleibe ich auch hier bei der Symbolik des Anonymisierungsfalls, verwende also  $x$  statt  $x^*$  und  $x^a$  statt  $x$ , also ein "Fehler-in-den-Variablen-Modell" vorliegt.

<sup>38</sup>Aus Gründen der Einheitlichkeit bleibe ich auch hier bei der Symbolik des Anonymisierungsfalls, verwende also  $x$  statt  $x^*$  und  $x^a$  statt  $x$ .

Dann ergibt sich für den Wahrscheinlichkeitsgrenzwert<sup>39</sup>

$$\begin{aligned} \text{plim}_{n \rightarrow \infty} \hat{\beta}^a &= \frac{\text{plim} \frac{1}{nT} \sum_{t=1}^T \sum_{i=1}^n (x_{it}^a - \bar{x}_{i\bullet}^a)(y_{it} - \bar{y}_{i\bullet})}{\text{plim} \frac{1}{nT} \sum_{t=1}^T \sum_{i=1}^n (x_{it}^a - \bar{x}_{i\bullet}^a)^2} \\ &= \frac{\text{plim} \frac{1}{nT} \sum_{t=1}^T \sum_{i=1}^n (x_{it} - \bar{x}_{i\bullet} + u_{it} - \bar{u}_{i\bullet})((x_{it} - \bar{x}_{i\bullet})\beta + \eta_{it} - \bar{\eta}_{i\bullet})}{\text{plim} \frac{1}{nT} \sum_{t=1}^T \sum_{i=1}^n ((x_{it} - \bar{x}_{i\bullet} + u_{it} - \bar{u}_{i\bullet}))^2} \end{aligned}$$

Da  $x$ ,  $u$  und  $\eta$  gemäß Annahme miteinander unkorreliert sind, ergibt sich<sup>40</sup>

$$\begin{aligned} \text{plim}_{n \rightarrow \infty} \hat{\beta}^a &= \frac{\text{plim} \frac{1}{nT} \sum_{t=1}^T \sum_{i=1}^n (x_{it} - \bar{x}_{i\bullet})^2 \beta}{\text{plim} \frac{1}{nT} \sum_{t=1}^T \sum_{i=1}^n (x_{it} - \bar{x}_{i\bullet})^2 + \text{plim} \frac{1}{nT} \sum_{t=1}^T \sum_{i=1}^n (u_{it} - \bar{u}_{i\bullet})^2} \\ &= \frac{\sigma_x^2}{\sigma_x^2 + (1 - \frac{1}{T}) \sigma_u^2} \beta \end{aligned} \quad (9-20)$$

Der Faktor  $1 - 1/T$  wird dann besonders bedeutsam sein, wenn  $T$  sehr groß ist, was bei Panel-Mikrodaten eher selten der Fall sein wird. Da dann der Nenner größer wird, wird die Verzerrung gegen Null verstärkt!

### 9.4.3 Einschub: Korrekter Wahrscheinlichkeitsgrenzwert im Panelfall ?

Das Ergebnis (9-20) bedarf allerdings eines Kommentars: Bei Biørn (1996 Formel (10-59), Symbolik leicht verändert) wird der Wahrscheinlichkeitsgrenzwert von  $\hat{\beta}^a$  wie folgt geschrieben:

$$\text{plim}_{n \rightarrow \infty} \hat{\beta}^a = \beta - \frac{(1 - \frac{1}{T}) \sigma_u^2 \beta}{Q_x + (1 - \frac{1}{T}) \sigma_u^2}$$

mit

$$Q_x = \text{plim} \frac{1}{nT} \sum_{t=1}^T \sum_{i=1}^n (x_{it} - \bar{x}_{i\bullet})^2$$

d.h. dieser Wahrscheinlichkeitsgrenzwert wird nicht explizit angegeben. In meinem obigen Ergebnis habe ich  $Q_x = \sigma_x^2$  gesetzt.

Entsprechend dem Vorgehen in Fußnote 40 würde sich aber

$$Q_x = \left(1 - \frac{1}{T}\right) \sigma_x^2$$

<sup>39</sup>Ich betrachte ausschließlich den Grenzwert für  $n \rightarrow \infty$ . Deshalb schreibe ich im folgenden oftmals nur  $\text{plim}$  statt  $\text{plim}_{n \rightarrow \infty}$ .

<sup>40</sup>Dabei wird ausgenutzt, daß

$$\sum_{t=1}^T \sum_{i=1}^n (u_{it} - \bar{u}_{i\bullet})^2 = \sum_{t=1}^T \sum_{i=1}^n (u_{it} - \bar{u})^2 - T \sum_{i=1}^n (\bar{u}_{i\bullet} - \bar{u})^2$$

und

$$\text{var}[\bar{u}_{i\bullet}] = \frac{\sigma_u^2}{T}$$

gilt. Damit erhalten wir

$$\text{plim} \frac{1}{nT} \sum_{t=1}^T \sum_{i=1}^n (u_{it} - \bar{u}_{i\bullet})^2 = \text{plim} \frac{1}{nT} \sum_{t=1}^T \sum_{i=1}^n (u_{it} - \bar{u})^2 - \text{plim} \frac{1}{nT} T \sum_{i=1}^n (\bar{u}_{i\bullet} - \bar{u})^2 = \sigma_u^2 - \frac{\sigma_u^2}{T}.$$

ergeben, so daß sich der Faktor  $1 - 1/T$  insgesamt herauskürzen würde. Entsprechende Bemerkungen gelten für die im folgenden verwendeten Wahrscheinlichkeitsgrenzwerte!! Auch die abgeleiteten Korrekturformeln sind unter diesem Vorbehalt zu sehen.

#### 9.4.4 Meßfehler mit Faktorstruktur

Biørn (1996 Kapitel 10.3) betrachtet außerdem ein lineares Panelmodell, in dem er für das Fehler-Modell eine Varianz-Komponenten-Struktur (error components structure) wie folgt unterstellt:

$$x_{it}^a = x_{it} + \tau_i + \phi_t + \varepsilon_{it} \quad (9-21)$$

Für den naiven Panelschätzer ergibt sich bei dieser Fehler-Struktur der folgende Wahrscheinlichkeitsgrenzwert:

$$\text{plim}_{n \rightarrow \infty} \hat{\beta}^a = \frac{\sigma_x^2}{\sigma_x^2 + \sigma_\phi^2 + (1 - \frac{1}{T})\sigma_\varepsilon^2} \beta \quad (9-22)$$

Siehe Biørn (1996), Formel (10-115), der allerdings auch hier wieder keinen expliziten Grenzwert bezüglich  $x$  angibt. Siehe meine Bemerkungen in Unterabschnitt 9.4.3.

Bemerkenswert ist, daß durch die Bildung von Abweichungen vom individuen-spezifischen Mittelwert der Effekt  $\tau_i$  bei der Schätzung keine Rolle spielt, im Gegensatz zum zeit-spezifischen Effekt  $\phi_t$ . Man beachte auch den engen Zusammenhang mit der additiven Überlagerung a la Hönne in (9-5), wo der zeitspezifische Effekt entfällt. Wir werden später sehen, daß deshalb bei der additiven Überlagerung der Bias des naiven Panelschätzers nur von  $\sigma_\varepsilon^2$ , nicht aber vom Parameter  $\mu$  abhängt. Siehe dazu Abschnitt 9.4.5. Entsprechende Ergebnisse für den multiplikativen Fall werden in den Abschnitten 9.4.6 sowie 9.4.7 besprochen. Dort beeinflußt allerdings auch der Zuschlagsparameter ( $\delta$ ) den Bias der Schätzung.

#### 9.4.5 Additive Überlagerung a la Hönne

Bei der additiven Überlagerung der Regressoren a la Hönne verwenden wir im Panelfall die Spezifikation (9-5), d.h.

$$x_{it}^a = x_{it} + \mu D_i + \varepsilon_{it}$$

und erhalten wegen

$$\bar{x}_{i\bullet}^a = \bar{x}_{i\bullet} + \mu D_i + \bar{\varepsilon}_{i\bullet}$$

und

$$x_{it}^a - \bar{x}_{i\bullet}^a = x_{it} - \bar{x}_{i\bullet} + \varepsilon_{it} - \bar{\varepsilon}_{i\bullet}$$

als Wahrscheinlichkeitsgrenzwert des "naiven" Panelschätzers

$$\text{plim}_{n \rightarrow \infty} \hat{\beta}^a = \frac{\sigma_x^2}{\sigma_x^2 + (1 - \frac{1}{T}) \sigma_\varepsilon^2} \beta, \quad (9-23)$$

d.h. der Bias hängt nur von der Varianz des Restterms  $\varepsilon$ , nicht aber von der Größenordnung des Zuschlag-Parameters  $\mu$  ab! Da der (quadrierte) Zuschlagsparameter deutlich größer ist als die Varianz  $\sigma_\varepsilon^2$ , verringert sich der Bias gegenüber der Schätzung im Querschnittsfall



(siehe Abschnitt 8.3) beträchtlich. Auch absolut gesehen wird der Bias äußerst gering sein, weil das Verhältnis  $\sigma_x/\sigma_\varepsilon$  meist sehr groß sein wird.

Ein Korrekturschätzer läßt sich wie folgt konstruieren: Wegen (6-5) gilt

$$\text{var}[x^a] = \sigma_x^2 + \mu^2 + \sigma_\varepsilon^2 .$$

Weil außer  $\sigma_x^2$  alle Größen bekannt sind bzw. geschätzt werden können, erhalten wir den Korrekturschätzer

$$\hat{\beta}^{a,korr} = \frac{\widehat{\text{var}}[x^a] - \mu^2 - \frac{1}{T}\sigma_\varepsilon^2}{\widehat{\text{var}}[x^a] - \mu^2 - \sigma_\varepsilon^2} \hat{\beta}_W^a \quad (9-24)$$

wobei  $\widehat{\text{var}}[x^a]$  eine (konsistente) Schätzung von  $\text{var}[x^a]$  ist. In der Korrekturschätzung taucht also auch der Parameter  $\mu$  auf !

Falls auch die abhängige Variable überlagert wird, erhalten wir entsprechend der obigen Analyse

$$x_{it}^a - \bar{x}_{i\bullet}^a = x_{it} - \bar{x}_{i\bullet} + \varepsilon_{itx} - \bar{\varepsilon}_{i\bullet x}$$

und

$$y_{it}^a - \bar{y}_{i\bullet}^a = y_{it} - \bar{y}_{i\bullet} + \varepsilon_{ity} - \bar{\varepsilon}_{i\bullet y}$$

wobei die beiden Überlagerungsvariablen nun angeben, ob sie zu  $x$  oder  $y$  gehören. Wie in Abschnitt 8.3 ausführlich dargestellt, gilt gemäß Annahme (!!)

$$\text{cov}[\varepsilon_x, \varepsilon_y] = 0 .$$

Demnach erhalten wir für den naiven Panelschätzer<sup>41</sup>

$$\text{plim}_{n \rightarrow \infty} \hat{\beta}^a = \frac{\sigma_x^2 \beta + (1 - \frac{1}{T}) \text{cov}[\varepsilon_x, \varepsilon_y]}{\sigma_x^2 + (1 - \frac{1}{T}) \sigma_\varepsilon^2} = \frac{\sigma_x^2 \beta}{\sigma_x^2 + (1 - \frac{1}{T}) \sigma_\varepsilon^2} , \quad (9-25)$$

Dies entspricht der Formel (9-23) für den Fall, daß nur der Regressor (additiv) überlagert wird. Demnach ist derselbe Korrekturschätzer auch hier verwendbar. Man beachte, daß dieses Ergebnis deutlich von dem im Fall von Querschnittsdaten abweicht. Dort ergab sich eine zusätzliche Verzerrung bei Überlagerung beider Variablen! Siehe Abschnitt 8.3.

<sup>41</sup>Hier wird die folgende Zerlegung verwendet: Für beliebige Zufallsvariablen  $\lambda_{itj}$  mit  $\text{cov}[\lambda_{itj}, \lambda_{itk}] = \kappa, j \neq k$  gilt

$$\sum_{t=1}^T \sum_{i=1}^n (\lambda_{itx} - \bar{\varepsilon}_{i\bullet x})(\lambda_{ity} - \bar{\varepsilon}_{i\bullet y}) = \sum_{t=1}^T \sum_{i=1}^n (\lambda_{itx} - \bar{\varepsilon}_x)(\lambda_{ity} - \bar{\varepsilon}_y) - T \sum_{i=1}^n (\bar{\varepsilon}_{i\bullet x} - \bar{\varepsilon}_x)(\bar{\varepsilon}_{i\bullet y} - \bar{\varepsilon}_y)$$

und

$$\text{cov}[\bar{\varepsilon}_{i\bullet x}, \bar{\varepsilon}_{i\bullet y}] = \frac{\kappa}{T}$$

gilt. Damit erhalten wir

$$\text{plim} \frac{1}{nT} \sum_{t=1}^T \sum_{i=1}^n (\lambda_{itx} - \bar{\varepsilon}_{i\bullet x})(\lambda_{ity} - \bar{\varepsilon}_{i\bullet y}) = \kappa - \frac{\kappa}{T} .$$

Allerdings ist im obigen Fall die Kovarianz gleich Null!

### 9.4.6 Multiplikative Überlagerung des Regressors

Im Fall der multiplikativen Überlagerung des Regressors im Fall von Paneldaten verwenden wir statt (9-5) die Spezifikation (9-10), d.h.

$$x_{it}^a = x_{it} (1 + \delta D_i + \varepsilon_{it}) \quad ,$$

die direkt aus (9-1) folgt. Dann ergibt sich für die Berechnung der Abweichung vom Mittelwert

$$\bar{x}_{i\bullet}^a = (1 + \delta D_i) \bar{x}_{i\bullet} + \bar{x}\varepsilon_{i\bullet}$$

und

$$x_{it}^a - \bar{x}_{i\bullet}^a = (1 + \delta D_i) (x_{it} - \bar{x}_{i\bullet}) + x_{it} \varepsilon_{it} - \bar{x}\varepsilon_{i\bullet} \quad (9-26)$$

wobei

$$\bar{x}\varepsilon_{i\bullet} = \frac{1}{T} \sum_t x_{it} \varepsilon_{it} \quad .$$

Im Gegensatz zur additiven Überlagerung verschwindet also der Zuschlagsparameter ( $\delta$ ) in diesem Fall nicht! Er wird also auch bei der Bestimmung des Bias eine Rolle spielen. Dies soll im folgenden abgeleitet werden.

Unter Verwendung der obigen Ergebnisse kann man den naiven Panelschätzer wie folgt schreiben:

$$\begin{aligned} \hat{\beta}_W^a &= \frac{\sum_{t=1}^T \sum_{i=1}^n (x_{it}^a - \bar{x}_{i\bullet}^a) (y_{it} - \bar{y}_{i\bullet})}{\sum_{t=1}^T \sum_{i=1}^n (x_{it}^a - \bar{x}_{i\bullet}^a)^2} \\ &= \frac{\sum_{t=1}^T \sum_{i=1}^n ((1 + \delta D_i) (x_{it} - \bar{x}_{i\bullet}) + x_{it} \varepsilon_{it} - \bar{x}\varepsilon_{i\bullet}) (y_{it} - \bar{y}_{i\bullet})}{\sum_{t=1}^T \sum_{i=1}^n ((1 + \delta D_i) (x_{it} - \bar{x}_{i\bullet}) + x_{it} \varepsilon_{it} - \bar{x}\varepsilon_{i\bullet})^2} \\ &= \frac{\sum_{t=1}^T \sum_{i=1}^n ((1 + \delta D_i) (x_{it} - \bar{x}_{i\bullet}) + x_{it} \varepsilon_{it} - \bar{x}\varepsilon_{i\bullet}) ((x_{it} - \bar{x}_{i\bullet}) \beta + \eta_{it} - \bar{\eta}_{i\bullet})}{\sum_{t=1}^T \sum_{i=1}^n ((1 + \delta D_i) (x_{it} - \bar{x}_{i\bullet}) + x_{it} \varepsilon_{it} - \bar{x}\varepsilon_{i\bullet})^2} \\ &= \frac{\sum_{t=1}^T \sum_{i=1}^n \left\{ (1 + \delta D_i) (x_{it} - \bar{x}_{i\bullet})^2 \beta + (1 + \delta D_i) (x_{it} - \bar{x}_{i\bullet}) (\eta_{it} - \bar{\eta}_{i\bullet}) \right\}}{\sum_{t=1}^T \sum_{i=1}^n \left\{ (1 + \delta D_i)^2 (x_{it} - \bar{x}_{i\bullet})^2 + 2(1 + \delta D_i) (x_{it} - \bar{x}_{i\bullet}) (x_{it} \varepsilon_{it} - \bar{x}\varepsilon_{i\bullet}) + (x_{it} \varepsilon_{it} - \bar{x}\varepsilon_{i\bullet})^2 \right\}} \\ &\quad + \frac{\sum_{t=1}^T \sum_{i=1}^n \left\{ (x_{it} \varepsilon_{it} - \bar{x}\varepsilon_{i\bullet}) (x_{it} - \bar{x}_{i\bullet}) \beta + (x_{it} \varepsilon_{it} - \bar{x}\varepsilon_{i\bullet}) (\eta_{it} - \bar{\eta}_{i\bullet}) \right\}}{\sum_{t=1}^T \sum_{i=1}^n \left\{ (1 + \delta D_i)^2 (x_{it} - \bar{x}_{i\bullet})^2 + 2(1 + \delta D_i) (x_{it} - \bar{x}_{i\bullet}) (x_{it} \varepsilon_{it} - \bar{x}\varepsilon_{i\bullet}) + (x_{it} \varepsilon_{it} - \bar{x}\varepsilon_{i\bullet})^2 \right\}} \end{aligned} \quad (9-27)$$

Zur Bestimmung des Wahrscheinlichkeitsgrenzwertes dieses Schätzers sind jeweils die den empirischen Momenten entsprechenden theoretischen Momente zu bestimmen, denn bei Berechnung des "plim" erhalten wir

$$\text{plim}(\hat{\beta}_W^a) = \frac{\frac{1}{T} \sum_{t=1}^T \{A_t + B_t\}}{\frac{1}{T} \sum_{t=1}^T \{\Phi_t + \Pi_t + \Psi_t\}} + \frac{\frac{1}{T} \sum_{t=1}^T \{\Gamma_t + \Delta_t\}}{\frac{1}{T} \sum_{t=1}^T \{\Phi_t + \Pi_t + \Psi_t\}} \quad (9-28)$$

wobei

$$A_t = E[(1 + \delta D)(X_t - E[X_t])^2 \beta] \quad (9-29)$$

$$B_t = E[(1 + \delta D)(X_t - E[X_t])(\eta_t - E[\eta_t])] \quad (9-30)$$

$$\Gamma_t = E[(X_t \varepsilon_t - E[X_t \varepsilon_t])(X_t - E[X_t])\beta] \quad (9-31)$$

$$\Delta_t = E[(X_t \varepsilon_t - E[X_t \varepsilon_t])(\eta_t - E[\eta_t])] \quad (9-32)$$

$$\Phi_t = E[(1 + \delta D)^2 (X_t - E[X_t])^2] \quad (9-33)$$

$$\Pi_t = 2 E[(1 + \delta D)(X_t - E[X_t])(X_t \varepsilon_t - E[X_t \varepsilon_t])] \quad (9-34)$$

$$\Psi_t = E[(X_t \varepsilon_t - E[X_t \varepsilon_t])^2] \quad (9-35)$$

gilt. Da alle Variablen als stationär verteilt angenommen werden, ist der Index  $t$  de facto zu vernachlässigen. Nicht berücksichtigt wurde, daß die empirischen Momente die individuellen spezifischen Mittelwerte enthalten. Siehe dazu meine Bemerkung in Unterabschnitt 9.4.3.

Für die einzelnen Ausdrücke ergibt sich folgendes: Wegen der Unabhängigkeit von  $D$  und  $X_t$  erhalten wir für (9-29)

$$A_t = \sigma_x^2 \beta,$$

wobei wir  $E[D] = 0$  verwendet haben (symmetrischer Fall). Da  $X_t$  und  $\eta_t$  unabhängig sind, gilt für (9-30)

$$B_t = 0 \quad .$$

In (9-31) ist die Kovarianz zwischen  $X_t \varepsilon_t$  und  $X_t$  zu bestimmen. Wegen der stochastischen Unabhängigkeit dieser beiden Zufallsvariablen sowie wegen  $E[\varepsilon_t] = 0$  ergibt sich

$$E[(X_t \varepsilon_t - E[X_t \varepsilon_t])(X_t - E[X_t])] = E[X_t^2] E[\varepsilon_t] - (E[X_t])^2 E[\varepsilon_t] = E[\varepsilon_t] \sigma_x^2 = 0$$

und damit

$$\Gamma_t = 0 \quad .$$

Für (9-32) ergibt sich

$$\Delta_t = 0 \quad ,$$

weil  $\eta_t$  stochastisch unabhängig von  $X_t$  und  $\varepsilon_t$  ist. In (9-33) kann wieder die Unabhängigkeit zwischen  $D$  und  $X_t$  ausgenutzt werden. Weil

$$E[(1 + \delta D)^2] = 1 + \delta^2$$

gilt, erhalten wir

$$\Phi_t = (1 + \delta^2) \sigma_x^2 \quad .$$

Für (9-34) erhalten wir

$$\Pi_t = 0 \quad .$$

Vergleiche dazu die Berechnung zu (9-30) oben. Außerdem kann wieder die Unabhängigkeit zwischen  $D$  und  $X_t$  ausgenutzt werden. In (9-35) schließlich ist die Varianz des Produktes  $X_t \varepsilon_t$  zu bestimmen. Deshalb erhalten wir

$$\Psi_t = \sigma_\varepsilon^2 (\sigma_x^2 + \mu_x^2) \quad .$$

Damit ergibt sich für den Wahrscheinlichkeitsgrenzwert von  $\hat{\beta}^a$  bei ausschließlicher Überlagerung des Regressor mit der multiplikativen "speziellen" Höhe-Variante folgender Ausdruck:

$$\text{plim}(\hat{\beta}_W^a) = \frac{\sigma_x^2 \beta}{(1 + \delta^2) \sigma_x^2 + \sigma_\varepsilon^2 (\sigma_x^2 + \mu_x^2)} \quad . \quad (9-36)$$

Demnach verschwindet der Bias bei der multiplikativen Überlagerung nur dann, wenn sowohl der Zuschlag als auch die Restkomponente nicht auf die Variable  $X$  einwirken, d.h. wenn

$$\delta = 0 \quad \text{und} \quad \sigma_\varepsilon^2 = 0$$

gilt.

#### 9.4.7 Multiplikative Überlagerung beider (aller) Variablen

Falls auch die abhängige Variable  $y$  nach dem Höhenverfahren multiplikativ überlagert wird, gilt für diese überlagerte Variable<sup>42</sup>

$$\begin{aligned} y_{it}^a &= y_{it} (1 + \delta D_i + \varepsilon_{ity}) \\ &= (1 + \delta D_i + \varepsilon_{ity})(\beta x_{it} + \tau_i + \eta_{it}) \\ &= (1 + \delta D_i) \tau_i + (1 + \delta D_i)(\beta x_{it} + \eta_{it}) + \tau_i \varepsilon_{ity} + \varepsilon_{ity}(\beta x_{it} + \eta_{it}) \end{aligned}$$

wobei jetzt  $\varepsilon_{ity}$  anzeigt, daß es sich um die Störvariable für die Überlagerung von  $y$  handelt. Für den Mittelwert erhalten wir

$$\overline{y^a}_{i\bullet} = (1 + \delta D_i) \tau_i + (1 + \delta D_i) (\beta \overline{x}_{i\bullet} + \overline{\eta}_{i\bullet}) + \tau_i \overline{\varepsilon}_{i\bullet y} + \beta \overline{x \varepsilon}_{i\bullet y} + \overline{\varepsilon \eta}_{i\bullet}$$

wobei

$$\overline{x \varepsilon}_{i\bullet y} = \frac{1}{T} \sum_t x_{it} \varepsilon_{itx}$$

und

$$\overline{\varepsilon \eta}_{i\bullet y} = \frac{1}{T} \sum_t \varepsilon_{ity} \eta_{it}$$

Dann erhalten wir für die Abweichung vom Mittelwert

$$\begin{aligned} y_{it}^a - \overline{y^a}_{i\bullet} &= (1 + \delta D_i) (x_{it} - \overline{x}_{i\bullet}) \beta + (1 + \delta D_i) (\eta_{it} - \overline{\eta}_{i\bullet}) + \tau_i (\varepsilon_{ity} - \overline{\varepsilon}_{i\bullet y}) \\ &\quad + (x_{it} \varepsilon_{ity} - \overline{x \varepsilon}_{i\bullet y}) \beta + (\varepsilon_{ity} \eta_{it} - \overline{\varepsilon \eta}_{i\bullet y}) \end{aligned} \quad (9-37)$$

Um den Wahrscheinlichkeitsgrenzwert von  $\hat{\beta}_W^a$  zu bestimmen, ist der Grenzwert des Zählers von (9-19), d.h.

$$\text{plim}_{n \rightarrow \infty} \frac{1}{nT} \sum_{t=1}^T \sum_{i=1}^n (x_{it}^a - \overline{x}_{i\bullet}^a) (y_{it}^a - \overline{y^a}_{i\bullet}) \quad ,$$

unter Verwendung von (9-26) und (9-37) neu zu bestimmen, während für den Nenner die Ergebnisse aus dem Abschnitt 9.4.6 übernommen werden können. Dabei verwenden wir jetzt für die Überlagerung von  $x$  ebenfalls die Symbolik  $\varepsilon_{itx}$  für die hier relevante Störvariable und schreiben deshalb (9-26) und daraus folgende Ergebnis wie folgt:

$$\begin{aligned} x_{it}^a &= x_{it} (1 + \delta D_i + \varepsilon_{itx}) \quad , \\ \overline{x^a}_{i\bullet} &= (1 + \delta D_i) \overline{x}_{i\bullet} + \overline{x \varepsilon}_{i\bullet x} \end{aligned}$$

<sup>42</sup>Wieder unterstellen wir, daß das lineare Modell mich **stochastischen Effekten** gilt. Siehe Abschnitt \*\*\*.

und

$$x_{it}^a - \bar{x}_{i\bullet}^a = (1 + \delta D_i)(x_{it} - \bar{x}_{i\bullet}) + x_{it} \varepsilon_{itx} - \bar{x} \bar{\varepsilon}_{i\bullet x} \quad (9-38)$$

wobei

$$\bar{x} \bar{\varepsilon}_{i\bullet x} = \frac{1}{T} \sum_t x_{it} \varepsilon_{itx}$$

gilt. Im Folgenden wird die bisher nicht ausdrücklich formulierte Annahme benutzt, daß  $\varepsilon_{itx}$  und  $\varepsilon_{ity}$  für alle  $i$  und  $t$  unkorreliert sind.

Im folgenden werden die einzelnen Terme zur Berechnung des Grenzwertes des Zählers von (9-19), d.h. von

$$\text{plim}_{n \rightarrow \infty} \frac{1}{nT} \sum_{t=1}^T \sum_{i=1}^n (x_{it}^a - \bar{x}_{i\bullet}^a)(y_{it}^a - \bar{y}_{i\bullet}^a) \quad ,$$

unter Verwendung von (9-38) und (9-37) systematisch dargestellt.

Summanden von $y_{it}^a - \bar{y}_{i\bullet}^a$	Summanden von $x_{it}^a - \bar{x}_{i\bullet}^a$	
	$(1 + \delta D_i)(x_{it} - \bar{x}_{i\bullet})$	$x_{it} \varepsilon_{itx} - \bar{x} \bar{\varepsilon}_{i\bullet x}$
$(1 + \delta D_i)(x_{it} - \bar{x}_{i\bullet}) \beta$	$\Phi_t \beta$	0
$(1 + \delta D_i)(\eta_{it} - \bar{\eta}_{i\bullet})$	0	0
$\tau_i(\varepsilon_{it} - \bar{\varepsilon}_{i\bullet})$	0	0
$(x_{it} \varepsilon_{it} - \bar{x} \bar{\varepsilon}_{i\bullet}) \beta$	0	0
$(\varepsilon_{it} \eta_{it} - \bar{\varepsilon}_{i\bullet} \bar{\eta}_{i\bullet})$	0	0

Hinweis: In den einzelnen Zellen wird der jeweilige Grenzwert angegeben.

Eine Analyse der einzelnen Kombinationen, die sich als Produkte der einzelnen Summanden ergeben, zeigt, daß nur die Kombination in der linken oberen Zelle, die zum Ausdruck

$$(1 + \delta D_i)^2 (x_{it} - \bar{x}_{i\bullet})^2 \beta$$

führt, einen von Null verschiedenen Grenzwert ergibt. Wie bereits in Unterabschnitt 9.4.6 gezeigt, erhalten wir

$$\text{plim}_{n \rightarrow \infty} \frac{1}{n} \sum_i ((1 + \delta D_i)^2 (x_{it} - \bar{x}_{i\bullet})^2) \beta = \Phi_t \beta = (1 + \delta^2) \sigma_x^2 \beta$$

Damit ergibt sich für den Wahrscheinlichkeitsgrenzwert von  $\hat{\beta}^a$  bei Überlagerung sowohl des Regressors als auch der abhängigen Variablen mit der multiplikativen "speziellen" Höhne-Variante folgender Ausdruck:

$$\text{plim}(\hat{\beta}_W^a) = \frac{(1 + \delta^2) \sigma_x^2 \beta}{(1 + \delta^2) \sigma_x^2 + \sigma_\varepsilon^2 (\sigma_x^2 + \mu_x^2)} \quad . \quad (9-39)$$

Ein Vergleich mit dem in (9-36) gegebenen Grenzwert bei ausschließlicher Überlagerung des Regressors zeigt, daß die zusätzliche Überlagerung der abhängigen Variablen den Bias um den Faktor  $1 + \delta^2$  erhöht. Auch in diesem Fall verschwindet der Bias nur dann, wenn sowohl der Zuschlag als auch die Restkomponente nicht auf die Variable  $X$  einwirken, d.h. wenn

$$\delta = 0 \quad \text{und} \quad \sigma_\varepsilon^2 = 0$$

gilt.

### 9.4.8 Multiplikative Überlagerung der abhängigen Variablen

Der Vollständigkeit halber sei noch angemerkt, daß sich bei ausschließlicher Überlagerung der abhängigen Variablen keine Verzerrung ergibt. Dies sieht man beispielsweise dadurch, daß man die Tabelle in Abschnitt 9.4.7 wie folgt modifiziert: Es gibt nur eine Spalte, und zwar mit dem Summanden  $x_{it} - \bar{x}_{i\bullet}$ , der jeweils mit den Zeilenausdrücken bezüglich der abhängigen Variablen zu kombinieren ist. Der einzige von Null verschiedene Grenzwert ergibt sich für die oberste Zeile mit  $A_t = \sigma_x^2 \beta$ . Da auch der Nenner von  $\hat{\beta}_W^a$  gegen  $\sigma_x^2$  strebt, erhalten wir als Grenzwert insgesamt

$$\frac{\sigma_x^2}{\sigma_x^2} \beta \quad .$$

### 9.5 Korrekturschätzer

Unter der Annahme, daß dem Datennutzer die Parameter der stochastischen Überlagerung, also  $\delta$  und  $\sigma_\varepsilon^2$ , bekannt sind, können aus den zuvor abgeleiteten Ergebnissen Korrekturschätzer für die einzelnen Situationen abgeleitet werden. Ich beschränke mich hier beispielhaft auf die Situation, in der sowohl  $x$  als auch  $y$  multiplikativ mit dem speziellen Höhne-Verfahren überlagert werden. Fas entsprechende Ergebnis für den "naiven" Panelschätzer zeigt (9-39). Daraus erhalten wir bei Auflösung nach  $\beta$  folgenden Korrekturschätzer:

$$\hat{\beta}_W^{a,korr} = \frac{(1 + \delta^2) \widehat{\sigma}_x^2 + \sigma_\varepsilon^2 (\widehat{\sigma}_x^2 + \widehat{\mu}_x^2)}{(1 + \delta^2) \widehat{\sigma}_x^2} \hat{\beta}_W^a \quad . \quad (9-40)$$

Dabei ergibt sich unter Verwendung von (9-11) folgende Schätzung für  $\sigma_x^2$ :

$$\widehat{\sigma}_x^2 = \frac{s_{x^a}^2 - (\delta^2 + \sigma_\varepsilon^2) \bar{x}^{a2}}{1 + \delta^2 + \sigma_\varepsilon^2} \quad , \quad (9-41)$$

wobei

$$\bar{x}^a = \frac{1}{nT} \sum_t \sum_i x_{it}^a$$

und

$$s_{x^a}^2 = \frac{1}{nT} \sum_t \sum_i (x_{it}^a - \bar{x}^a)^2$$

verwendet werden sollten. Ferner sollte

$$\widehat{\mu}_x = \bar{x}^a$$

benutzt werden.

## 10 Literatur

- Biørn, Erik(1996) "Panel data with measurement errors". in: Matyas, L. and P. Sevestre: The Econometrics of Panel Data: A Handbook of the Theory with Applications, Kluwer, \*\*\*\*. Zweite revidierte Auflage, 236-279.
- Evans,T., L. Zayatz und J. Slanta (1998). Using Noise for disclosure limitation of establishment tabular data. Journal of Official Statistics \*\*\*\*\*
- Hwang, J. (1986). Multiplicative errors-in-variables models with applications to recent data released by the U.S. Department of Energy. Journal of the American Statistical Association 81, 680-688.
- J.J. Kim und W. E. Winkler : "Masking Microdata Files." *American Statistical Association. Proceedings of the Section on Survey Research Methods* 1995 , 114-119.
- Lin, A.(1989). Estimation of multiplicative Measurement error models and some simulation results. Economics letters 31, 13-20
- Marshall und Olkin(1979). Inequalities.Theory of Majorization and Its Application. Academic Press: New York.
- Massell,P., L. Zayatz und J. Funk (2006). Protecting the confidentiality of survey tabular data by adding noise to the underlying micro data: Application to the Commodity Flow Survey. In: J. Domingo und L. Franconi (Herausgeber): Privacy in Statistical Data Bases. CENEX-SDC Project International Conference, PSD 2006, Rome, Italy, December 2006, Proceedings, Springer, Berlin, 304-317.
- G. McLachlan und D. Peel: *Finite Mixture Models*. Wiley, New York, 2000.
- Ronning, G.(1977). Pseudo-Bayessche Schätzer in der Ökonometrie. *Operations Research Verfahren* 26 , 861-871.
- Ronning, G. et al (2005). Handbuch zur Anonymisierung wirtschaftsstatistischer Mikrodaten. Statistisches Bundesamt, Wiesbaden , Reihe "Statistik und Wissenschaft", Band 4, 2005, (gemeinsam mit Roland Sturm, Jörg Höhne, Rainer Lenz, Martin Rosemann, Michael Scheffler und Daniel Vorgrimler)
- Rosemann, M.(2006). Auswirkungen datenverändernder Anonymisierungsverfahren auf die Analyse von Mikrodaten". IAW Tübingen.
- Gina Marie Roque. Masking Microdata Files with Mixtures of Multivariate Normal Distributions. Dissertation, June 2000, University of California, Riverside.
- Springer, M.D. (1979). The Algebra of Random Variables. Wiley: New York.
- Wooldridge, J.M. 2002). Econometric Analysis of Cross Section and Panel Data. MIT Press: Cambridge (MA).
- . E. Yancey, W. E. Winkler und R.H. Creecy: "Disclosure Risk Assessment in Perturbative Micro Data Protection." in: J. Domingo-Ferrer (Herausgeber): *Inference Control in Statistical Databases*. Springer: Berlin, 2002, 135-152.

## A Die (spezielle) Höhne-Spezifikation im additiven Fall

In Abschnitt 5.2 haben wir nur die multiplikative Variante des "speziellen" Höhne-Verfahrens betrachtet. Hier ergänzen wir diese Ergebnisse für den "additiven" Fall:

$$\mathbf{U} = \mu D \boldsymbol{\nu} + \boldsymbol{\varepsilon} \quad , \quad (\text{A-1})$$

wobei  $\mu$  ein Parameter mit beliebigem Wert ist. Genau wie in Abschnitt 5.2 ist die diskrete Zufallsvariable  $D$  durch

$$D = \begin{cases} +1 & \text{mit Wahrscheinlichkeit } \alpha \\ -1 & \text{mit Wahrscheinlichkeit } 1 - \alpha \end{cases}$$

gegeben (siehe (5-2)) und der  $r$ -dimensionale stetige Zufallsvektor  $\boldsymbol{\varepsilon}$  ist entsprechend (5-3) spezifiziert:

$$E[\boldsymbol{\varepsilon}] = \mathbf{0} \quad , \quad \text{cov}[\boldsymbol{\varepsilon}] = \sigma_{\varepsilon}^2 \mathbf{I} \quad (\text{A-2})$$

wobei die Annahme der Normalverteilung üblicherweise hinzutritt. Genau wie in Abschnitt 5.2 wird die stochastische Unabhängigkeit von  $D$  und  $\boldsymbol{\varepsilon}$  unterstellt.

Im symmetrischen Fall ergibt sich für den additiven Fall

$$E[\mathbf{U}] = \mathbf{0}$$

und

$$\text{cov}[\mathbf{U}] = \sigma_{\varepsilon}^2 \mathbf{I} + \mu^2 \boldsymbol{\nu} \boldsymbol{\nu}' = \frac{1}{\sigma_{\varepsilon}^2 + \mu^2} ((1 - \rho) \mathbf{I} + \rho \boldsymbol{\nu} \boldsymbol{\nu}') \quad (\text{A-3})$$

wobei

$$\rho = \frac{\mu^2}{\mu^2 + \sigma_{\varepsilon}^2}$$

gilt. Man beachte, daß diese Korrelation stets positiv ist!

## B Flexible Höhne-Spezifikation

### B.1 Multiplikativer Fall

Im Abschnitt 8, in dem die Auswirkungen der Überlagerung auf die Schätzung eines linearen Modells untersucht, betrachte ich (8-11), die als eine "flexible" Variante von (8-10) interpretiert werden kann und die hier ohne den Beobachtungs-Index  $\langle i \rangle$  geschrieben wird:

$$\mathbf{u} = (\boldsymbol{\nu} + \delta \mathbf{D}) + \boldsymbol{\varepsilon} \quad . \quad i = 1, \dots, n \quad (\text{B-1})$$

Dabei ist  $\mathbf{D}$  jetzt ein  $r$ -dimensionaler Zufallsvektor, dessen Komponenten jeweils die Spezifikation (5-2) erfüllen und die stochastisch voneinander unabhängig sind. Um die Eigenschaften der dadurch implizierten Verteilung sowie der ersten und zweiten Momente abzuleiten, gehe ich wie in den Abschnitten 5.2.1 und 5.2.2 vor.

Da die einzelnen Komponenten von  $\mathbf{U}$  wegen der Struktur

$$U_j = 1 + \delta D_j + \varepsilon_j$$



voneinander stochastisch unabhängig sind und für die Varianz von  $U_j$

$$\text{Var}[U_j] = \delta^2 \text{Var}[D_j] + \sigma_\varepsilon^2$$

gilt, ergibt sich für Erwartungswert und Kovarianzmatrix

$$E[\mathbf{U}] = (1 + \delta(2\alpha - 1))\boldsymbol{\iota} \quad , \quad \text{cov}[\mathbf{U}] = (4\alpha(1 - \alpha)\delta^2 + \sigma_\varepsilon^2) \mathbf{I} . \quad (\text{B-2})$$

Im "symmetrischen" Fall, also für  $\alpha = 0,5$ , erhalten wir

$$E[\mathbf{U}] = \boldsymbol{\iota} \quad , \quad \text{cov}[\mathbf{U}] = (\delta^2 + \sigma_\varepsilon^2) \mathbf{I} .$$

Diese Ergebnisse entsprechen für Erwartungswerte und Varianzen, aber **nicht für Kovarianzen** den Ergebnissen der spezielleren Höhne-Spezifikation (8-10). Der grundlegende Unterschied ist, daß bei der "flexiblen" Spezifikation alle Komponenten unkorreliert sind.

Für die Ableitung der Dichtefunktion von  $\mathbf{U}$  gehen wir wie folgt vor: Wir betrachten zunächst die bedingte Dichte von  $\mathbf{U}$  gegeben der Vektor (!!!)  $\mathbf{D}$ , multiplizieren diese Dichte mit der Randdichte von  $\mathbf{D}$ , um die gemeinsame Dichte zu bestimmen und "integrieren" dann den Zufallsvektor  $\mathbf{D}$  aus, um die Randdichte von  $\mathbf{U}$  zu erhalten. Dabei unterstellen wir im folgenden wieder zusätzlich Normalverteilung:

$$\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma_\varepsilon^2 \mathbf{I}) \quad . \quad (\text{B-3})$$

Für gegebenen Vektor  $\mathbf{D}$  ist jede Komponente von  $\mathbf{U}$  in (B-1) normalverteilt. Zunächst gilt:

$$U_j | D_j \sim N(1 + \delta d_j, \sigma_\varepsilon^2) \quad .$$

Allerdings gilt wegen der stochastischen Unabhängigkeit der einzelnen  $D_j$  auch

$$U_j | D_j, D_k, k \neq j \sim N(1 + \delta d_j, \sigma_\varepsilon^2) \quad .$$

Außerdem sind wegen der stochastischen Unabhängigkeit der  $\varepsilon_j$  die einzelnen  $U_j$  (gegeben  $D_j$ ) stochastisch voneinander unabhängig (bedingte Unabhängigkeit). Dann können wir für die multivariate bedingte Verteilung von  $\mathbf{U}$  gegeben  $\mathbf{D}$  schreiben:

$$\begin{pmatrix} U_1 | D_1 \\ \vdots \\ U_r | D_r \end{pmatrix} \sim N(\boldsymbol{\iota} + \delta \mathbf{D}, \sigma_\varepsilon^2 \mathbf{I}) = \prod_{j=1}^r N_j(1 + \delta D_j, \sigma_\varepsilon^2) \quad .$$

Dabei bezeichnet  $N_j$  die eindimensionale Dichte der Normalverteilung.

Außerdem gilt wegen der stochastischen Unabhängigkeit der Komponenten von  $\mathbf{D}$

$$h(\mathbf{d}) = h(d_1) \cdot \dots \cdot h(d_r) \quad .$$

Demnach ergibt sich für die gemeinsame Dichte von  $\mathbf{U}$  und  $\mathbf{D}$

$$g(\mathbf{u}, \mathbf{d}) = h(\mathbf{d}) \cdot N(\boldsymbol{\iota} + \delta \mathbf{d}, \sigma_\varepsilon^2 \mathbf{I}) = \prod_{j=1}^r \alpha^{\frac{1+d_j}{2}} (1 - \alpha)^{\frac{1-d_j}{2}} \cdot N(1 + \delta d_j, \sigma_\varepsilon^2) \quad .$$

Um die Randdichte des Vektors  $\mathbf{U}$  zu bestimmen, müssen die einzelnen Komponenten von  $\mathbf{D}$  'heraussummiert' werden. Allerdings bedeutet dies, daß alle möglichen Ereignisse

von  $\mathbf{D}$  betrachtet werden müssen. Ich illustriere dies für den einfachsten Fall von  $r = 2$  Merkmalen: In diesem Fall sind für den Vektor  $(D_1, D_2)$  folgende Ereignisse möglich:

$$\{(+1, +1), (+1, -1), (-1, +1), (-1, -1)\} \quad .$$

Dann erhalten wir bei Summation über diese Ereignisse

$$\begin{aligned} f(\mathbf{u}) &= \sum_{d_1, d_2 \in \{+1, -1\}} \prod_{j=1}^r \alpha^{\frac{1+d_j}{2}} (1-\alpha)^{\frac{1-d_j}{2}} \cdot N(1 + \delta d_j, \sigma_\varepsilon^2) \\ &= \{\alpha N(1 + \delta, \sigma_\varepsilon^2) + \alpha N(1 + \delta, \sigma_\varepsilon^2)\} \{\alpha N(1 + \delta, \sigma_\varepsilon^2) + (1-\alpha) N(1 - \delta, \sigma_\varepsilon^2)\} \times \\ &\quad \times \{(1-\alpha) N(1 - \delta, \sigma_\varepsilon^2) + \alpha N(1 + \delta, \sigma_\varepsilon^2)\} \{(1-\alpha) N(1 - \delta, \sigma_\varepsilon^2) + (1-\alpha) N(1 - \delta, \sigma_\varepsilon^2)\} \end{aligned}$$

Es ist evident, daß dies nicht der allgemeinen Formel der Mischungsverteilung in Abschnitt 3 entspricht.

Die allgemeine Formulierung der Dichtefunktion lautet demnach

$$f(\mathbf{u}) = \sum_{d_1, d_2, \dots, d_r \in \{+1, -1\}} \prod_{j=1}^r \alpha^{\frac{1+d_j}{2}} (1-\alpha)^{\frac{1-d_j}{2}} \cdot N(1 + \delta d_j, \sigma_\varepsilon^2) \quad . \quad (\text{B-4})$$

## B.2 Additiver Fall

Im additiven Fall schreiben wir die "flexible" Spezifikation als

$$\mathbf{u} = \mu \mathbf{D} + \varepsilon \quad . \quad i = 1, \dots, n \quad . \quad (\text{B-5})$$

Dabei ist  $\mathbf{D}$  jetzt ein  $r$ -dimensionaler Zufallsvektor, dessen Komponenten jeweils die Spezifikation (5-2) erfüllen und die stochastisch voneinander unabhängig sind.

Im symmetrischen Fall ergibt sich für den additiven Fall

$$E[\mathbf{U}] = \mathbf{0}$$

und

$$\text{cov}[\mathbf{U}] = (\mu^2 + \sigma_\varepsilon^2) \mathbf{I} \quad . \quad (\text{B-6})$$

## C Höhe-Verfahren und Paneldaten für mehrere Regressoren

### C.1 Allgemeines

Wir betrachten hier das **multiple** Panelmodell mit stochastischen Individualeffekten und mit  $K$  Regressoren

$$y_{it} = \sum_{k=1}^K \beta_k x_{itk} + \tau_i + \eta_{it} \quad i = 1, \dots, n \quad t = 1, \dots, T \quad , \quad (\text{C-1})$$

das zum einfachen Modell aus Abschnitt 9.3.2 korrespondiert. Für Mittelwert und Abweichung vom Mittelwert ergibt sich in diesem Fall

$$\bar{y}_{i\bullet} = \sum_{k=1}^K \bar{x}_{i\bullet k} \beta_k + \tau_i + \bar{\eta}_{i\bullet}$$

und

$$y_{it} - \bar{y}_{i\bullet} = \sum_{k=1}^K (x_{itk} - \bar{x}_{i\bullet k}) \beta_k + \eta_{it} - \bar{\eta}_{i\bullet}$$

Im folgenden verwenden wir auch die kompakte Schreibweise dieses Modells:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\tau} + \boldsymbol{\eta} \quad (\text{C-2})$$

Dabei hat die Regressormatrix folgende Struktur:

$$\mathbf{X} = \begin{pmatrix} \mathbf{X}\langle 1 \rangle \\ \mathbf{X}\langle 2 \rangle \\ \vdots \\ \mathbf{X}\langle n-1 \rangle \\ \mathbf{X}\langle n \rangle \end{pmatrix}$$

Der Index in eckigen Klammern bezeichnet einen bestimmten Beobachtungspunkt  $i$ . Für jede der  $n$  verschiedenen  $(T \times K)$ -Matrizen  $\mathbf{X}\langle i \rangle$  gilt

$$\mathbf{X}\langle i \rangle = \left( \mathbf{x}_1\langle i \rangle \quad \mathbf{x}_2\langle i \rangle \quad \dots \quad \mathbf{x}_{K-1}\langle i \rangle \quad \mathbf{x}_K\langle i \rangle \right) \quad , \quad i = 1, \dots, n \quad ,$$

wobei der  $T$ -dimensionale Vektor  $\mathbf{x}_k\langle i \rangle$  durch

$$\mathbf{x}_k\langle i \rangle = \begin{pmatrix} x_{i1k} \\ x_{i2k} \\ \vdots \\ x_{i,T-1,k} \\ x_{iT k} \end{pmatrix} \quad , \quad k = 1, \dots, K \quad ,$$

gegeben ist. Für die Matrix  $\mathbf{X}^a$  gilt die entsprechende Symbolik.

Die Matrix  $\mathbf{W}$  im Panelmodell mit fixen Effekten (siehe Abschnitt 9.3.1) hat folgende Struktur:

$$\mathbf{W} = \mathbf{I}_n \otimes \boldsymbol{\iota}_T$$

Daraus ergibt sich die idempotente Matrix

$$\mathbf{M}_W = \mathbf{I}_{nT} - \mathbf{W}(\mathbf{W}'\mathbf{W})^{-1}\mathbf{W}' = \mathbf{I}_n \otimes \left( \mathbf{I}_T - \frac{1}{T} \boldsymbol{\iota}_T \boldsymbol{\iota}_T' \right)$$

und für den "within"-Schätzer der Originaldaten ergibt sich

$$\hat{\boldsymbol{\beta}}_W = (\mathbf{X}'\mathbf{M}_W\mathbf{X})^{-1}\mathbf{X}'\mathbf{M}_W\mathbf{y} \quad . \quad (\text{C-3})$$

## C.2 Ausschließliche Überlagerung der Regressoren

Der "naive" Panelschätzer für anonymisierte Regressoren (und Originalwerte der abhängigen Variablen  $y$ ) ist durch

$$\hat{\beta}_W^a = (\mathbf{X}^{a'} \mathbf{M}_W \mathbf{X}^a)^{-1} \mathbf{X}^{a'} \mathbf{M}_W \mathbf{y} \quad . \quad (\text{C-4})$$

gegeben.

Wir untersuchen zunächst die Struktur von  $\mathbf{M}_W \mathbf{X}^a$ . Unter Verwendung der Ergebnisse von oben erhalten wir

$$\mathbf{M}_W \mathbf{X}^a = \begin{pmatrix} \left( \mathbf{I}_T - \frac{1}{T} \boldsymbol{\nu}_T \boldsymbol{\nu}_T' \right) \mathbf{X}^a \langle 1 \rangle \\ \left( \mathbf{I}_T - \frac{1}{T} \boldsymbol{\nu}_T \boldsymbol{\nu}_T' \right) \mathbf{X}^a \langle 2 \rangle \\ \vdots \\ \left( \mathbf{I}_T - \frac{1}{T} \boldsymbol{\nu}_T \boldsymbol{\nu}_T' \right) \mathbf{X}^a \langle n-1 \rangle \\ \left( \mathbf{I}_T - \frac{1}{T} \boldsymbol{\nu}_T \boldsymbol{\nu}_T' \right) \mathbf{X}^a \langle n \rangle \end{pmatrix}$$

und für die quadratische Form von  $\mathbf{X}^{a'} \mathbf{M}_W \mathbf{X}^a$  ergibt sich

$$\begin{aligned} \mathbf{X}^{a'} \mathbf{M}_W \mathbf{X}^a &= \\ &= \sum_{i=1}^n \mathbf{X}^{a'} \langle i \rangle \left( \mathbf{I}_T - \frac{1}{T} \boldsymbol{\nu}_T \boldsymbol{\nu}_T' \right) \mathbf{X}^a \langle i \rangle \\ &= \sum_{i=1}^n \begin{pmatrix} \left( \mathbf{x}_1^a \langle i \rangle - \overline{\mathbf{x}}_1^a \langle i \bullet \rangle \right)^2 & \left( \mathbf{x}_1^a \langle i \rangle - \overline{\mathbf{x}}_1^a \langle i \bullet \rangle \right)' \left( \mathbf{x}_2^a \langle i \rangle - \overline{\mathbf{x}}_2^a \langle i \bullet \rangle \right) & \dots & \left( \mathbf{x}_1^a \langle i \rangle - \overline{\mathbf{x}}_1^a \langle i \bullet \rangle \right)' \left( \mathbf{x}_K^a \langle i \rangle - \overline{\mathbf{x}}_K^a \langle i \bullet \rangle \right) \\ \left( \mathbf{x}_2^a \langle i \rangle - \overline{\mathbf{x}}_2^a \langle i \bullet \rangle \right)' \left( \mathbf{x}_1^a \langle i \rangle - \overline{\mathbf{x}}_1^a \langle i \bullet \rangle \right) & \left( \mathbf{x}_2^a \langle i \rangle - \overline{\mathbf{x}}_2^a \langle i \bullet \rangle \right)^2 & & \left( \mathbf{x}_2^a \langle i \rangle - \overline{\mathbf{x}}_2^a \langle i \bullet \rangle \right)' \left( \mathbf{x}_K^a \langle i \rangle - \overline{\mathbf{x}}_K^a \langle i \bullet \rangle \right) \\ \vdots & \vdots & \ddots & \vdots \\ \left( \mathbf{x}_K^a \langle i \rangle - \overline{\mathbf{x}}_K^a \langle i \bullet \rangle \right)' \left( \mathbf{x}_1^a \langle i \rangle - \overline{\mathbf{x}}_1^a \langle i \bullet \rangle \right) & \left( \mathbf{x}_K^a \langle i \rangle - \overline{\mathbf{x}}_K^a \langle i \bullet \rangle \right)' \left( \mathbf{x}_2^a \langle i \rangle - \overline{\mathbf{x}}_2^a \langle i \bullet \rangle \right) & \dots & \left( \mathbf{x}_K^a \langle i \rangle - \overline{\mathbf{x}}_K^a \langle i \bullet \rangle \right)^2 \end{pmatrix} \\ &= \sum_{i=1}^n \sum_{t=1}^T \begin{pmatrix} \left( x_{it1}^a - \overline{x}_{i\bullet 1}^a \right)^2 & \left( x_{it1}^a - \overline{x}_{i\bullet 1}^a \right) \left( x_{it2}^a - \overline{x}_{i\bullet 2}^a \right) & \dots & \left( x_{it1}^a - \overline{x}_{i\bullet 1}^a \right) \left( x_{itK}^a - \overline{x}_{i\bullet K}^a \right) \\ \left( x_{it1}^a - \overline{x}_{i\bullet 1}^a \right) \left( x_{it2}^a - \overline{x}_{i\bullet 2}^a \right) & \left( x_{it2}^a - \overline{x}_{i\bullet 2}^a \right)^2 & \dots & \left( x_{it2}^a - \overline{x}_{i\bullet 2}^a \right) \left( x_{itK}^a - \overline{x}_{i\bullet K}^a \right) \\ \vdots & \vdots & \ddots & \vdots \\ \left( x_{itK}^a - \overline{x}_{i\bullet K}^a \right) \left( x_{it1}^a - \overline{x}_{i\bullet 1}^a \right) & \left( x_{itK}^a - \overline{x}_{i\bullet K}^a \right) \left( x_{it2}^a - \overline{x}_{i\bullet 2}^a \right) & \dots & \left( x_{itK}^a - \overline{x}_{i\bullet K}^a \right)^2 \end{pmatrix} \end{aligned}$$

wobei die  $K$ -dimensionalen Vektoren der individuen-spezifischen Mittelwerte wie folgt definiert sind:

$$\overline{\mathbf{x}}_k^a \langle i \bullet \rangle = \frac{1}{T} \boldsymbol{\nu}_T \boldsymbol{\nu}_T' \mathbf{x}_k^a \langle i \rangle = \frac{1}{T} \sum_{t=1}^T \begin{pmatrix} x_{it1}^a \\ x_{it2}^a \\ \vdots \\ x_{itK}^a \end{pmatrix}$$

Ich greife nun auf die Ergebnisse aus Abschnitt 9.4.6 zurück, in dem die Überlagerung eines einzigen Regressors betrachtet wurde. Daraus ergibt sich bei Überlagerung mittels (9-1) für den  $k$ -ten Regressor

$$x_{itk}^a = x_{itk} (1 + \delta D_i + \varepsilon_{itk}) \quad ,$$

und

$$\overline{x}_{i\bullet k}^a = (1 + \delta D_i) \overline{x}_{i\bullet k} + \overline{x \varepsilon}_{i\bullet k}$$

wobei

$$\overline{x \varepsilon}_{i\bullet k} = \frac{1}{T} \sum_t x_{itk} \varepsilon_{itk} \quad .$$

gilt. Für die Abweichung vom Mittelwert ergibt sich daraus

$$x_{itk}^a - \bar{x}_{i\bullet k}^a = (1 + \delta D_i)(x_{itk} - \bar{x}_{i\bullet k}) + x_{itk} \varepsilon_{itk} - \bar{x} \varepsilon_{i\bullet k} \quad (C-5)$$

Wie bereits in Abschnitt 9.4.6 bemerkt, verschwindet im Gegensatz zur additiven Überlagerung der Zuschlagsparameter ( $\delta$ ) in diesem Fall nicht! Ich habe im erwähnten Abschnitt auch bereits gezeigt, daß

$$\text{plim}_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n (x_{itk}^a - \bar{x}_{i\bullet k}^a)^2 = (1 + \delta^2) \sigma_k^2 + \sigma_\varepsilon^2 (\sigma_k^2 + \mu_k^2) \quad (C-6)$$

gilt. Siehe die Ergebnisse zu (9-33) bis (9-35). Allerdings schreiben wir dieses Ergebnis jetzt für den  $k$ -ten Regressor mit Varianz  $\sigma_k^2$  und Erwartungswert  $\mu_k$ .

Zusätzlich benötigen wir jetzt ein Ergebnis für

$$\text{plim}_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n (x_{itk}^a - \bar{x}_{i\bullet k}^a) (x_{ith}^a - \bar{x}_{i\bullet h}^a) \quad , k \neq h \quad .$$

Entsprechend dem Vorgehen im erwähnten Abschnitt zerlegen wir wieder den Summenausdruck:

$$\begin{aligned} & \sum_{i=1}^n (x_{itk}^a - \bar{x}_{i\bullet k}^a) (x_{ith}^a - \bar{x}_{i\bullet h}^a) \\ &= \sum_{i=1}^n \{(1 + \delta D_i)(x_{itk} - \bar{x}_{i\bullet k}) + (x_{itk} \varepsilon_{itk} - \bar{x} \varepsilon_{i\bullet k})\} \{(1 + \delta D_i)(x_{ith} - \bar{x}_{i\bullet h}) + (x_{ith} \varepsilon_{ith} - \bar{x} \varepsilon_{i\bullet h})\} \\ &= \sum_{i=1}^n \{(1 + \delta D_i)^2 (x_{itk} - \bar{x}_{i\bullet k})(x_{ith} - \bar{x}_{i\bullet h}) + (x_{itk} \varepsilon_{itk} - \bar{x} \varepsilon_{i\bullet k})(x_{ith} \varepsilon_{ith} - \bar{x} \varepsilon_{i\bullet h})\} \\ & \quad + \sum_{i=1}^n \{(1 + \delta D_i)(x_{itk} - \bar{x}_{i\bullet k})(x_{ith} \varepsilon_{ith} - \bar{x} \varepsilon_{i\bullet h}) + (1 + \delta D_i)(x_{itk} - \bar{x}_{i\bullet k})(x_{ith} \varepsilon_{ith} - \bar{x} \varepsilon_{i\bullet h})\} \end{aligned}$$

Aus dieser Zerlegung erkennt man, daß

$$\text{plim}_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n (x_{itk}^a - \bar{x}_{i\bullet k}^a) (x_{ith}^a - \bar{x}_{i\bullet h}^a) = (1 + \delta^2) \sigma_{kh} \quad (C-7)$$

gilt, wobei  $\sigma_{kh}$  die Kovarianz der Regressoren  $k$  und  $h$  bezeichnet. Man beachte insbesondere, daß im Gegensatz zu (9-35) hier statt der Varianz von  $X_{tk} \varepsilon_{tk}$  die Kovarianz zwischen  $X_{tk} \varepsilon_{tk}$  und  $X_{th} \varepsilon_{th}$  betrachtet wird, die Null ist.

Damit erhalten wir insgesamt

$$\begin{aligned} n \xrightarrow{\text{plim}} \frac{1}{n} \mathbf{X}^{\mathbf{a}'} \mathbf{M}_W \mathbf{X}^{\mathbf{a}} &= n \xrightarrow{\text{plim}} \frac{1}{n} \sum_{i=1}^n \mathbf{X}^{\mathbf{a}'(i)} \left( \mathbf{I}_T - \frac{1}{T} \iota_T \iota_T' \right) \mathbf{X}^{\mathbf{a}(i)} \\ &= \begin{pmatrix} (1 + \delta^2) \sigma_1^2 + \sigma_\varepsilon^2 (\sigma_1^2 + \mu_1^2) & (1 + \delta^2) \sigma_{12} & \dots & (1 + \delta^2) \sigma_{1K} \\ (1 + \delta^2) \sigma_{21} & (1 + \delta^2) \sigma_2^2 + \sigma_\varepsilon^2 (\sigma_2^2 + \mu_2^2) & \dots & (1 + \delta^2) \sigma_{2K} \\ \vdots & \vdots & \ddots & \vdots \\ (1 + \delta^2) \sigma_{K1} & (1 + \delta^2) \sigma_{K2} & \dots & (1 + \delta^2) \sigma_K^2 + \sigma_\varepsilon^2 (\sigma_K^2 + \mu_K^2) \end{pmatrix} \end{aligned} \quad (C-8)$$

Im Gegensatz zur additiven Überlagerung stehen hier nicht zwei Matrizen additiv nebeneinander. Bemerkenswert ist, daß die Diagonalelemente stärker erhöht werden als die Nichtdiagonalelemente!

Um den Wahrscheinlichkeitsgrenzwert für den "naiven" Panelschätzer zu bestimmen, benötigen wir ferner

$$\text{plim}_{n \rightarrow \infty} \frac{1}{n} \mathbf{X}^{\mathbf{a}'} \mathbf{M}_W \mathbf{y}$$

Dazu betrachten wir

$$\begin{aligned}
\mathbf{X}^{\mathbf{a}'}\mathbf{M}_W\mathbf{y} &= \sum_{i=1}^n \mathbf{X}^{\mathbf{a}'}\langle i \rangle \left( \mathbf{I}_T - \frac{1}{T} \boldsymbol{\iota}_T \boldsymbol{\iota}_T' \right) \mathbf{y}\langle i \rangle \\
&= \sum_{i=1}^n \begin{pmatrix} (\mathbf{x}_1^a\langle i \rangle - \overline{\mathbf{x}}_1^a\langle i \bullet \rangle)' (\mathbf{y}\langle i \rangle - \overline{\mathbf{y}}\langle i \bullet \rangle) \\ (\mathbf{x}_1^a\langle i \rangle - \overline{\mathbf{x}}_1^a\langle i \bullet \rangle)' (\mathbf{y}\langle i \rangle - \overline{\mathbf{y}}\langle i \bullet \rangle) \\ \vdots \\ (\mathbf{x}_1^a\langle i \rangle - \overline{\mathbf{x}}_1^a\langle i \bullet \rangle)' (\mathbf{y}\langle i \rangle - \overline{\mathbf{y}}\langle i \bullet \rangle) \end{pmatrix} \\
&= \sum_{i=1}^n \sum_{t=1}^T \begin{pmatrix} (x_{it1}^a - \overline{x}_{i\bullet 1}^a) (y_{it} - \overline{y}_{i\bullet}) \\ (x_{it2}^a - \overline{x}_{i\bullet 2}^a) (y_{it} - \overline{y}_{i\bullet}) \\ \vdots \\ (x_{itK}^a - \overline{x}_{i\bullet K}^a) (y_{it} - \overline{y}_{i\bullet}) \end{pmatrix}
\end{aligned}$$

Auch hier können die Ergebnisse aus Abschnitt 9.4.6 sowie zusätzlich aus C.1 verwendet werden. Für jeweils ein Element aus dem zuvor bestimmten Vektor ergibt sich

$$\begin{aligned}
&\sum_{i=1}^n \sum_{t=1}^T (x_{itk}^a - \overline{x}_{i\bullet k}^a) (y_{it} - \overline{y}_{i\bullet}) \\
&= \sum_{i=1}^n \sum_{t=1}^T \left\{ ((1 + \delta D_i) (x_{itk} - \overline{x}_{i\bullet k}) + x_{itk} \varepsilon_{itk} - \overline{x}_{i\bullet k} \overline{\varepsilon}_{i\bullet k}) \left( \sum_{h=1}^K (x_{ith} - \overline{x}_{i\bullet h}) \beta_h + \eta_{it} - \overline{\eta}_{i\bullet} \right) \right\} \\
&= \sum_{t=1}^T \sum_{i=1}^n (1 + \delta D_i) \sum_{h=1}^K (x_{ith} - \overline{x}_{i\bullet h}) (x_{itk} - \overline{x}_{i\bullet k}) \beta_h \\
&\quad + \sum_{t=1}^T \sum_{i=1}^n (1 + \delta D_i) (x_{itk} - \overline{x}_{i\bullet k}) (\eta_{it} - \overline{\eta}_{i\bullet}) \\
&\quad + \sum_{t=1}^T \sum_{i=1}^n \left\{ \sum_{h=1}^K \{ (x_{itk} \varepsilon_{itk} - \overline{x}_{i\bullet k} \overline{\varepsilon}_{i\bullet k}) (x_{ith} - \overline{x}_{i\bullet h}) \beta_h \} \right\} \\
&\quad + \sum_{t=1}^T \sum_{i=1}^n \{ (x_{itk} \varepsilon_{itk} - \overline{x}_{i\bullet k} \overline{\varepsilon}_{i\bullet k}) (\eta_{it} - \overline{\eta}_{i\bullet}) \}
\end{aligned}$$

Eine Analyse der einzelnen Summanden entsprechend Abschnitt 9.4.6 ergibt, daß nur der erste Summand zu einem von Null verschiedenen Grenzwert führt, d.h. für bestimmtes  $k$  ergibt sich

$$\text{plim}_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n (1 + \delta D_i) \sum_{h=1}^K (x_{ith} - \overline{x}_{i\bullet h}) (x_{itk} - \overline{x}_{i\bullet k}) \beta_h = \sum_{h=1}^K \sigma_{kh} \beta_h$$

Demnach gilt

$$\text{plim}_{n \rightarrow \infty} \frac{1}{n} \mathbf{X}^{\mathbf{a}'}\mathbf{M}_W\mathbf{y} = \begin{pmatrix} \sum_{h=1}^K \sigma_{1h} \beta_h \\ \sum_{h=1}^K \sigma_{2h} \beta_h \\ \vdots \\ \vdots \\ \sum_{h=1}^K \sigma_{Kh} \beta_h \end{pmatrix} \quad (\text{C-9})$$

Für den Wahrscheinlichkeitsgrenzwert des naiven Panelschätzers (C-3) ergibt sich dann aus (C-8) und (C-9)

$$\begin{aligned}
\text{plim}_{n \rightarrow \infty} \hat{\beta}^a &= \left( \text{plim}_{n \rightarrow \infty} \frac{1}{n} \mathbf{X}^a{}' \mathbf{M}_W \mathbf{X}^a \right)^{-1} \text{plim}_{n \rightarrow \infty} \frac{1}{n} \mathbf{X}^a{}' \mathbf{M}_W \mathbf{y} \\
&= \left( \begin{array}{cccc}
(1 + \delta^2) \sigma_1^2 + \sigma_\varepsilon^2 (\sigma_1^2 + \mu_1^2) & (1 + \delta^2) \sigma_{12} & \dots & (1 + \delta^2) \sigma_{1k} \\
(1 + \delta^2) \sigma_{21} & (1 + \delta^2) \sigma_2^2 + \sigma_\varepsilon^2 (\sigma_2^2 + \mu_2^2) & \dots & (1 + \delta^2) \sigma_{2K} \\
\vdots & \vdots & \ddots & \vdots \\
(1 + \delta^2) \sigma_{K1} & (1 + \delta^2) \sigma_{K2} & \dots & (1 + \delta^2) \sigma_K^2 + \sigma_\varepsilon^2 (\sigma_K^2 + \mu_K^2)
\end{array} \right)^{-1} \begin{pmatrix} \sum_{h=1}^K \sigma_{1h} \beta_h \\ \sum_{h=1}^K \sigma_{2h} \beta_h \\ \vdots \\ \sum_{h=1}^K \sigma_{Kh} \beta_h \end{pmatrix}
\end{aligned} \tag{C-10}$$

Wie auch im Fall eines einzigen Regressors ist dieser Schätzer nur dann konsistent, wenn sowohl  $\delta = 0$  als auch  $\sigma_\varepsilon^2 = 0$  gilt, d.h. wenn sowohl der Zuschlag  $\delta$  als auch die Restüberlagerung durch  $\varepsilon$  entfällt.

### C.3 Gemeinsame Überlagerung aller Variablen

Falls alle Variablen überlagert wird (und das wird der Normalfall sein), ist auch die abhängige Variable  $y$  in die Anonymisierung einbezogen. Wie in Abschnitt 9.4.7 bereits für einen einzigen Regressor ausgeführt, gilt dann für die anonymisierte abhängige Variable

$$\begin{aligned}
y_{it}^a &= y_{it} (1 + \delta D_i + \varepsilon_{ity}) \\
&= (1 + \delta D_i + \varepsilon_{ity}) (\sum_{h=1}^K \beta_h x_{ith} + \tau_i + \eta_{it}) \\
&= (1 + \delta D_i) \tau_i + (1 + \delta D_i) (\sum_{h=1}^K \beta_h x_{ith} + \eta_{it}) + \tau_i \varepsilon_{ity} + \varepsilon_{ity} (\sum_{h=1}^K \beta_h x_{ith} + \eta_{it})
\end{aligned}$$

wobei jetzt  $\varepsilon_{ity}$  anzeigt, daß es sich um die Störvariable für die Überlagerung von  $y$  handelt. Für den Mittelwert erhalten wir

$$\bar{y}_{i\bullet}^a = (1 + \delta D_i) \tau_i + (1 + \delta D_i) \left( \sum_{h=1}^K \beta_h \bar{x}_{i\bullet h} + \bar{\eta}_{i\bullet} \right) + \tau_i \bar{\varepsilon}_{i\bullet y} + \sum_{h=1}^K \beta_h \bar{x}_{i\bullet h} \bar{\varepsilon}_{i\bullet y} + \bar{\varepsilon}_{i\bullet y} \bar{\eta}_{i\bullet}$$

wobei

$$\bar{x}_{i\bullet ky} = \frac{1}{T} \sum_t x_{itk} \varepsilon_{ity}$$

und

$$\bar{\varepsilon}_{i\bullet y} = \frac{1}{T} \sum_t \varepsilon_{ity} \eta_{it}$$

Dann erhalten wir für die Abweichung vom Mittelwert

$$\begin{aligned}
y_{it}^a - \bar{y}_{i\bullet}^a &= (1 + \delta D_i) \sum_{h=1}^K (x_{ith} - \bar{x}_{i\bullet h}) \beta_h + (1 + \delta D_i) (\eta_{it} - \bar{\eta}_{i\bullet}) + \tau_i (\varepsilon_{ity} - \bar{\varepsilon}_{i\bullet y}) \\
&\quad + \sum_{h=1}^K \beta_h (x_{ith} \varepsilon_{ity} - \bar{x}_{i\bullet h} \bar{\varepsilon}_{i\bullet y}) + (\varepsilon_{ity} \eta_{it} - \bar{\varepsilon}_{i\bullet y} \bar{\eta}_{i\bullet})
\end{aligned} \tag{C-11}$$

Um den Wahrscheinlichkeitsgrenzwert für den "naiven" Panelschätzer zu bestimmen, benötigen wir bei "zusätzlicher" Überlagerung der abhängigen Variablen

$$\text{plim}_{n \rightarrow \infty} \frac{1}{n} \mathbf{X}^a{}' \mathbf{M}_W \mathbf{y}^a$$

Dazu betrachten wir

$$\begin{aligned}
\mathbf{X}^{\mathbf{a}'} \mathbf{M}_W \mathbf{y}^{\mathbf{a}} &= \sum_{i=1}^n \mathbf{X}^{\mathbf{a}'} \langle i \rangle \left( \mathbf{I}_T - \frac{1}{T} \boldsymbol{\iota}_T \boldsymbol{\iota}_T' \right) \mathbf{y}^{\mathbf{a}} \langle i \rangle \\
&= \sum_{i=1}^n \begin{pmatrix} (\mathbf{x}_1^{\mathbf{a}} \langle i \rangle - \overline{\mathbf{x}}_1^{\mathbf{a}} \langle i \bullet \rangle)' (\mathbf{y}^{\mathbf{a}} \langle i \rangle - \overline{\mathbf{y}} \langle i \bullet \rangle) \\ (\mathbf{x}_1^{\mathbf{a}} \langle i \rangle - \overline{\mathbf{x}}_1^{\mathbf{a}} \langle i \bullet \rangle)' (\mathbf{y}^{\mathbf{a}} \langle i \rangle - \overline{\mathbf{y}}^{\mathbf{a}} \langle i \bullet \rangle) \\ \vdots \\ (\mathbf{x}_1^{\mathbf{a}} \langle i \rangle - \overline{\mathbf{x}}_1^{\mathbf{a}} \langle i \bullet \rangle)' (\mathbf{y}^{\mathbf{a}} \langle i \rangle - \overline{\mathbf{y}}^{\mathbf{a}} \langle i \bullet \rangle) \end{pmatrix} \\
&= \sum_{i=1}^n \sum_{t=1}^T \begin{pmatrix} (x_{it1}^{\mathbf{a}} - \overline{x}_{i\bullet 1}^{\mathbf{a}}) (y_{it}^{\mathbf{a}} - \overline{y}_{i\bullet}^{\mathbf{a}}) \\ (x_{it2}^{\mathbf{a}} - \overline{x}_{i\bullet 2}^{\mathbf{a}}) (y_{it}^{\mathbf{a}} - \overline{y}_{i\bullet}^{\mathbf{a}}) \\ \vdots \\ (x_{itK}^{\mathbf{a}} - \overline{x}_{i\bullet K}^{\mathbf{a}}) (y_{it}^{\mathbf{a}} - \overline{y}_{i\bullet}^{\mathbf{a}}) \end{pmatrix}
\end{aligned}$$

Im folgenden werden die einzelnen Terme zur Berechnung des Grenzwertes von

$$\text{plim}_{n \rightarrow \infty} \frac{1}{nT} \sum_{t=1}^T \sum_{i=1}^n (x_{itk}^{\mathbf{a}} - \overline{x}_{i\bullet k}^{\mathbf{a}}) (y_{it}^{\mathbf{a}} - \overline{y}_{i\bullet}^{\mathbf{a}}) \quad ,$$

unter Verwendung von (C-5) und (C-11) systematisch dargestellt.

Summanden von $y_{it}^{\mathbf{a}} - \overline{y}_{i\bullet}^{\mathbf{a}}$	Summanden von $x_{it}^{\mathbf{a}} - \overline{x}_{i\bullet}^{\mathbf{a}}$	
	$(1 + \delta D_i) (x_{itk} - \overline{x}_{i\bullet k})$	$x_{itk} \varepsilon_{itk} - \overline{x} \overline{\varepsilon}_{i\bullet k}$
$(1 + \delta D_i) \sum_{h=1}^K (x_{ith} - \overline{x}_{i\bullet h}) \beta_h$	$(1 + \delta^2) \sum_{h=1}^K \sigma_{kh} \beta_h$	0
$(1 + \delta D_i) (\eta_{it} - \overline{\eta}_{i\bullet})$	0	0
$\tau_i (\varepsilon_{ity} - \overline{\varepsilon}_{i\bullet y})$	0	0
$\sum_{h=1}^K \beta_h (x_{ith} \varepsilon_{ity} - \overline{x} \overline{\varepsilon}_{i\bullet hy})$	0	0
$(\varepsilon_{ity} \eta_{it} - \overline{\varepsilon} \overline{\eta}_{i\bullet y})$	0	0

**Hinweis:** In den einzelnen Zellen wird der jeweilige Grenzwert angegeben.

Eine Analyse analog zu der in Abschnitt 9.4.7 zeigt, daß ausschließlich das linke oberste Element einen von Null verschiedenen Grenzwert aufweist. Wir erhalten demnach

$$\text{plim}_{n \rightarrow \infty} \frac{1}{n} \mathbf{X}^{\mathbf{a}'} \mathbf{M}_W \mathbf{y}^{\mathbf{a}} = (1 + \delta^2) \begin{pmatrix} \sum_{h=1}^K \sigma_{1h} \beta_h \\ \sum_{h=1}^K \sigma_{2h} \beta_h \\ \vdots \\ \vdots \\ \sum_{h=1}^K \sigma_{Kh} \beta_h \end{pmatrix} \quad (\text{C-12})$$

und für den naiven Panelschätzer (C-3) ergibt sich in diesem Fall

$$\begin{aligned}
\text{plim}_{n \rightarrow \infty} \beta^{\mathbf{a}} &= \left( \text{plim}_{n \rightarrow \infty} \frac{1}{n} \mathbf{X}^{\mathbf{a}'} \mathbf{M}_W \mathbf{X}^{\mathbf{a}} \right)^{-1} \text{plim}_{n \rightarrow \infty} \frac{1}{n} \mathbf{X}^{\mathbf{a}'} \mathbf{M}_W \mathbf{y}^{\mathbf{a}} \\
&= (1 + \delta^2) \begin{pmatrix} (1 + \delta^2) \sigma_1^2 + \sigma_\varepsilon^2 (\sigma_1^2 + \mu_1^2) & (1 + \delta^2) \sigma_{12} & \dots & (1 + \delta^2) \sigma_{1k} \\ (1 + \delta^2) \sigma_{21} & (1 + \delta^2) \sigma_2^2 + \sigma_\varepsilon^2 (\sigma_2^2 + \mu_2^2) & \dots & (1 + \delta^2) \sigma_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ (1 + \delta^2) \sigma_{K1} & (1 + \delta^2) \sigma_{K2} & \dots & (1 + \delta^2) \sigma_K^2 + \sigma_\varepsilon^2 (\sigma_K^2 + \mu_K^2) \end{pmatrix}^{-1} \begin{pmatrix} \sum_{h=1}^K \sigma_{1h} \beta_h \\ \sum_{h=1}^K \sigma_{2h} \beta_h \\ \vdots \\ \vdots \\ \sum_{h=1}^K \sigma_{Kh} \beta_h \end{pmatrix} \\
&\hspace{25em} (\text{C-13})
\end{aligned}$$

d.h. genau wie im Fall nur eines Regressors verändert sich bei zusätzlicher Überlagerung der abhängigen Variablen der Wahrscheinlichkeitsgrenzwert um den Faktor  $(1 + \delta^2)$ . Siehe Abschnitt 9.4.7.



## C.4 Eine alternative Schreibweise

Zur Konstruktion eines Korrekturschätzers ist es nützlich, die zuvor abgeleiteten Grenzwerte anders darzustellen. Dazu definieren wir zunächst

$$\text{cov}[\mathbf{x}] = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \dots & \sigma_{1K} \\ \sigma_{21} & \sigma_2^2 & \dots & \sigma_{2K} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{K1} & \sigma_{K2} & \dots & \sigma_K^2 \end{pmatrix}$$

Dann können wir für (C-8) schreiben:

$$\text{plim}_{n \rightarrow \infty} \frac{1}{n} \mathbf{X}^{\mathbf{a}'} \mathbf{M}_W \mathbf{X}^{\mathbf{a}} = (1 + \delta^2) \text{cov}[\mathbf{x}] + \sigma_\varepsilon^2 \begin{pmatrix} \sigma_1^2 + \mu_1^2 & & & \\ & \sigma_2^2 + \mu_2^2 & & \\ & & \ddots & \\ & & & \sigma_K^2 + \mu_K^2 \end{pmatrix} \quad (\text{C-14})$$

Entsprechend ergeben sich aus (C-10) und (C-13)

$$\text{plim}_{n \rightarrow \infty} \frac{1}{n} \mathbf{X}^{\mathbf{a}'} \mathbf{M}_W \mathbf{y} = \text{cov}[\mathbf{x}] \boldsymbol{\beta} \quad (\text{C-15})$$

bzw.

$$\text{plim}_{n \rightarrow \infty} \frac{1}{n} \mathbf{X}^{\mathbf{a}'} \mathbf{M}_W \mathbf{y}^{\mathbf{a}} = (1 + \delta^2) \text{cov}[\mathbf{x}] \boldsymbol{\beta} \quad (\text{C-16})$$

und damit für den naiven Schätzer im Fall der ausschließlichen Überlagerung der Regressoren

$$\underset{n}{\text{plim}} \hat{\boldsymbol{\beta}}^{\mathbf{a}} = \left( (1 + \delta^2) \text{cov}[\mathbf{x}] + \sigma_\varepsilon^2 \begin{pmatrix} \sigma_1^2 + \mu_1^2 & & & \\ & \sigma_2^2 + \mu_2^2 & & \\ & & \ddots & \\ & & & \sigma_K^2 + \mu_K^2 \end{pmatrix} \right)^{-1} \text{cov}[\mathbf{x}] \boldsymbol{\beta} \quad (\text{C-17})$$

bzw. bei Überlagerung aller Variablen

$$\underset{n}{\text{plim}} \hat{\boldsymbol{\beta}}^{\mathbf{a}} = \left( \text{cov}[\mathbf{x}] + \frac{\sigma_\varepsilon^2}{(1 + \delta^2)} \begin{pmatrix} \sigma_1^2 + \mu_1^2 & & & \\ & \sigma_2^2 + \mu_2^2 & & \\ & & \ddots & \\ & & & \sigma_K^2 + \mu_K^2 \end{pmatrix} \right)^{-1} \text{cov}[\mathbf{x}] \boldsymbol{\beta} \quad (\text{C-18})$$

Aus diesen Formeln lassen sich Korrekturschätzer bestimmen. Dafür sind wieder geeignete Schätzer für die Mittelwerte und Varianzen sowie Kovarianzen der Regressorvariablen abzuleiten.

## D Einige nützliche Matrizen-Resultate

Bei der Analyse der multiplikativen Überlagerung spielt das Hadamard-Produkt eine wichtige Rolle. Ich liste deshalb im folgenden wichtige Eigenschaften auf, die auch Rosemann (2006) bereits verwendet hat.

Für "Hadamard-Produkt von zwei  $(m \times n)$ -Matrizen  $\mathbf{A}$  und  $\mathbf{B}$  gilt

$$\mathbf{A} \odot \mathbf{B} = \begin{pmatrix} a_{11}b_{11} & a_{12}b_{12} & \dots & a_{1,n-1}b_{1,n-1} & a_{1n}b_{1n} \\ a_{21}b_{21} & a_{22}b_{22} & \dots & a_{2,n-1}b_{2,n-1} & a_{2n}b_{2n} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ a_{m-1,1}b_{m-1,1} & a_{m-1,2}b_{m-1,2} & \dots & a_{m-1,n-1}b_{m-1,n-1} & a_{m-1,n}b_{m-1,n} \\ a_{m,1}b_{m,1} & a_{m,2}b_{m,2} & \dots & a_{m,n-1}b_{m,n-1} & a_{m,n}b_{m,n} \end{pmatrix}, \quad (\text{D-1})$$

d.h. die resultierende Matrix hat ebenfalls  $m$  Zeilen und  $n$  Spalten. Mit  $\div$  wird die entsprechende elementweise **Division** bezeichnet. Insbesondere gilt für beliebige Rechtecksmatrix

$$\mathbf{A} \div \mathbf{A} = \begin{pmatrix} 1 & 1 & \dots & 1 & 1 \\ 1 & 1 & \dots & 1 & 1 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 1 & \dots & 1 & 1 \\ 1 & 1 & \dots & 1 & 1 \end{pmatrix} = \boldsymbol{\iota} \boldsymbol{\iota}' \quad \text{und} \quad \mathbf{A} \odot \boldsymbol{\iota} \boldsymbol{\iota}' = \mathbf{A}, \quad (\text{D-2})$$

und für beliebigen Vektor  $\mathbf{a}$  gilt

$$\mathbf{a} \div \mathbf{a} = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \\ 1 \end{pmatrix} = \boldsymbol{\iota} \quad \text{und} \quad \mathbf{a} \odot \boldsymbol{\iota} = \mathbf{a}. \quad (\text{D-3})$$

Unmittelbar aus (D-1) folgt<sup>43</sup>

$$\text{vec}(\mathbf{A} \odot \mathbf{B}) = \text{vec}(\mathbf{A}) \odot \text{vec}(\mathbf{B}) \quad (\text{D-4})$$

Für  $m$ -dimensionale Vektoren  $\mathbf{a}$ ,  $\mathbf{c}$  und  $n$ -dimensionale Vektoren  $\mathbf{b}$ ,  $\mathbf{d}$  sind manchmal folgende Ergebnisse nützlich:

$$\mathbf{a} \odot \mathbf{b} = \mathbf{D}_a \mathbf{b} = \mathbf{D}_b \mathbf{a} \quad (\text{D-5})$$

Dabei sind  $\mathbf{D}_a$  und  $\mathbf{D}_b$  Diagonalmatrizen mit den Elementen des jeweiligen Vektors auf der Diagonalen. Ferner gilt

$$(\mathbf{a} \odot \mathbf{b})(\mathbf{c} \odot \mathbf{d})' = \mathbf{a} \mathbf{c}' \odot \mathbf{b} \mathbf{d}' \quad (\text{D-6})$$

In der Arbeit von Lin(1989) ist folgendes Ergebnis von besonderer Bedeutung, das nur indirekt mit dem Hadamard-Produkt zu tun hat: Für eine beliebige  $(m \times n)$ -Matrix  $\mathbf{A}$  gilt

$$\mathbf{A}' \mathbf{A} = \sum_{j=1}^n [(\mathbf{I}_n \otimes \mathbf{e}_j) \text{vec}(\mathbf{A}')] [(\mathbf{I}_n \otimes \mathbf{e}_j) \text{vec}(\mathbf{A}')]' \quad (\text{D-7})$$

Dabei ist  $\mathbf{e}_j$  ein  $n$ -dimensionaler Vektor und bezeichnet die  $j$ -te Spalte der Einheitsmatrix

<sup>43</sup>Der  $\text{vec}$ -Operator stapelt die Spalten einer Matrix übereinander.  $\text{vec}(\mathbf{A})$  ergibt einen  $mn$ -dimensionalen Vektor.

Beweis: Zunächst schreiben wir die Matrix  $\mathbf{A}$  wie folgt:

$$\mathbf{A} = \begin{pmatrix} \mathbf{a}'_1 \\ \mathbf{a}'_2 \\ \vdots \\ \mathbf{a}'_{m-1} \\ \mathbf{a}'_m \end{pmatrix} \quad \text{mit } \mathbf{a}'_j = (a_{j1}, \dots, a_{jn})$$

und damit

$$\mathbf{A}' \mathbf{A} = \sum_{j=1}^m \mathbf{a}_j \mathbf{a}'_j$$

Ferner gilt

$$\mathbf{A}' = (\mathbf{a}_1, \mathbf{a}_2 \dots \mathbf{a}_{m-1}, \mathbf{a}_m) \quad \text{und} \quad \text{vec}(\mathbf{A}') = \begin{pmatrix} \mathbf{a}_1 \\ \vdots \\ \vdots \\ \mathbf{a}_m \end{pmatrix}$$

sowie

$$(\mathbf{I}_n \otimes \mathbf{e}'_j) \text{vec}(\mathbf{A}') = (\mathbf{0}, \dots, \mathbf{I}_n, \dots, \mathbf{0}) \begin{pmatrix} \mathbf{a}_1 \\ \vdots \\ \mathbf{a}_j \\ \vdots \\ \mathbf{a}_m \end{pmatrix} = \mathbf{a}_j \quad .$$

Daraus folgt (D-7).

Für die lineare Form  $\mathbf{A} \mathbf{x}$  gilt folgendes Ergebnis: Es sei  $\mathbf{A}$  eine  $(m \times n)$ -Matrix und  $\mathbf{x}$  ein  $n$ -dimensionaler Vektor. Dann läßt sich die Linearform wie folgt darstellen:

$$\mathbf{A} \mathbf{x} = (\mathbf{I}_m \otimes \mathbf{x}') \text{vec}(\mathbf{A}') \quad (\text{D-8})$$

Beweis: Zunächst gilt

$$\mathbf{A} \mathbf{x} = \begin{pmatrix} \mathbf{a}'_1 \mathbf{x} \\ \vdots \\ \mathbf{a}'_m \mathbf{x} \end{pmatrix}$$

$$\mathbf{I}_m \otimes \mathbf{x}' = \begin{pmatrix} \mathbf{x}' & & & & \\ & \mathbf{x}' & & & \\ & & \ddots & & \\ & & & \mathbf{x}' & \\ & & & & \mathbf{x}' \end{pmatrix}$$

und damit

$$(\mathbf{I}_m \otimes \mathbf{x}') \text{vec}(\mathbf{A}') = \begin{pmatrix} \mathbf{x}' \mathbf{a}_1 \\ \vdots \\ \mathbf{x}' \mathbf{a}_m \end{pmatrix} \quad ,$$

womit (D-8) bewiesen ist.

Bisher erschienen:

Nr. 1

Das Einstiegsgeld – eine zielgruppenorientierte negative Einkommensteuer: Konzeption, Umsetzung und eine erste Zwischenbilanz nach 15 Monaten in Baden-Württemberg

*Sabine Dann / Andrea Kirchmann / Alexander Spermann / Jürgen Volkert*

Nr. 2

Die Einkommensteuerreform 1990 als natürliches Experiment. Methodische und konzeptionelle Aspekte zur Schätzung der Elastizität des zu versteuernden Einkommens

*Peter Gottfried / Hannes Schellhorn*

Nr. 3

Gut betreut in den Arbeitsmarkt? Eine mikroökonomische Evaluation der Mannheimer Arbeitsvermittlungsagentur

*Jürgen Jergler / Christian Pohnke / Alexander Spermann*

Nr. 4

Das IAW-Einkommenspanel und das Mikrosimulationsmodell SIMST

*Peter Gottfried / Hannes Schellhorn*

Nr. 5

A Microeconometric Characterisation of Household Consumption Using Quantile Regression

*Niels Schulze / Gerd Ronning*

Nr. 6

Determinanten des Überlebens von Neugründungen in der baden-württembergischen Industrie – eine empirische Survivalanalyse mit amtlichen Betriebsdaten

*Harald Strotmann*

Nr. 7

Die Baulandausweisungsumlage als ökonomisches Steuerungsinstrument einer nachhaltigkeitsorientierten Flächenpolitik

*Raimund Krumm*

Nr. 8

Making Work Pay: U.S. American Models for a German Context?

*Laura Chadwick, Jürgen Volkert*

# IAW-Diskussionspapiere

Nr. 9

Erste Ergebnisse von vergleichenden Untersuchungen mit anonymisierten und nicht anonymisierten Einzeldaten am Beispiel der Kostenstrukturerhebung und der Umsatzsteuerstatistik

*Martin Rosemann*

Nr. 10

Randomized Response and the Binary Probit Model

*Gerd Ronning*

Nr. 11

Creating Firms for a New Century: Determinants of Firm Creation around 1900

*Joerg Baten*

Nr. 12

Das fiskalische BLAU-Konzept zur Begrenzung des Siedlungsflächenwachstums

*Raimund Krumm*

Nr. 13

Generelle Nichtdiskontierung als Bedingung für eine nachhaltige Entwicklung?

*Stefan Bayer*

Nr. 14

Die Elastizität des zu versteuernden Einkommens. Messung und erste Ergebnisse zur empirischen Evidenz für die Bundesrepublik Deutschland.

*Peter Gottfried / Hannes Schellhorn*

Nr. 15

Empirical Evidence on the Effects of Marginal Tax Rates on Income – The German Case

*Peter Gottfried / Hannes Schellhorn*

Nr. 16

Shadow Economies around the World: What do we really know?

*Friedrich Schneider*

Nr. 17

Firm Foundations in the Knowledge Intensive Business Service Sector. Results from a Comparative Empirical Study in Three German Regions

*Andreas Koch / Thomas Stahlecker*

Nr. 18

The impact of functional integration and spatial proximity on the post-entry performance of knowledge intensive business service firms

*Andreas Koch / Harald Strotmann*

Nr. 19

Legislative Malapportionment and the Politicization of Germany's Intergovernmental Transfer System

*Hans Pitlik / Friedrich Schneider / Harald Strotmann*

Nr. 20

Implementation ökonomischer Steuerungsansätze in die Raumplanung

*Raimund Krumm*

Nr. 21

Determinants of Innovative Activity in Newly Founded Knowledge Intensive Business Service Firms

*Andreas Koch / Harald Strotmann*

Nr. 22

Impact of Opening Clauses on Bargained Wages

*Wolf Dieter Heinbach*

Nr. 23

Hat die Einführung von Gewinnbeteiligungsmodellen kurzfristige positive Produktivitätswirkungen? – Ergebnisse eines Propensity-Score-Matching-Ansatzes

*Harald Strotmann*

Nr. 24

Who Goes East? The Impact of Enlargement on the Pattern of German FDI

*Claudia M. Buch / Jörn Kleinert*

Nr. 25

Estimation of the Probit Model from Anonymized Micro Data

*Gerd Ronning / Martin Rosemann*

Nr. 26

Bargained Wages in Decentralized Wage-Setting Regimes

*Wolf Dieter Heinbach*

# IAW-Diskussionspapiere

Nr. 27

A Capability Approach for Official German Poverty and Wealth Reports:  
Conceptual Background and First Empirical Results

*Christian Arndt / Jürgen Volkert*

Nr. 28

Typisierung der Tarifvertragslandschaft – Eine Clusteranalyse der tarifvertraglichen  
Öffnungsklauseln

*Wolf Dieter Heinbach / Stefanie Schröpfer*

Nr. 29

International Bank Portfolios: Short- and Long-Run Responses to the Business Cycles

*Sven Blank / Claudia M. Buch*

Nr. 30

Stochastische Überlagerungen mit Hilfe der Mischungsverteilung

*Gerd Ronning*