IAW-Diskussionspapiere

Discussion Paper

30

Stochastische Überlagerung mit Hilfe der Mischungsverteilung

Revidierte und ergänzte Fassung

Gerd Ronning

Oktober 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
Internet: 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

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

---

Stochastische Überlagerung mit Hilfe der Mischungsverteilung[1]

GERD RONNING

(Stand: 11. Oktober 2007- Version 28)

Inhaltsverzeichnis

[1]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 und Hans Schneeweiß für hilfreiche Kommentare und Hinweise.

1 Einleitung

Diese Arbeit beschäftigt sich mit speziellen Aspekten der stochastischen Überlagerung, die als Anonymisierungsmethode zur Sicherstellung der faktischen Anonymität von Mikrodaten eingesetzt wird.[2] 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.

In dieser Fassung wurden gegenüber der Fassung, die als IAW-Diskussionspapier erschien (Version 20 - April 2007) **folgende Änderungen** vorgenommen:

- Ein neuer Abschnitt 4 über Lognormalverteilungen als Mischungsverteilungen wurde aufgenommen, die Nummerierung der folgenden Abschnitte ändert sich entsprechend.
- In Abschnitt 7 wurden die Ergebnisse für die multiplikative Überlagerung (allgemeiner Fall im Querschnitt) überarbeitet und erweitert.
- In Abschnitt 9 wurden die Ergebnisse für die Wahrscheinlichkeitsgrenzwerte des "naiven" Schätzers, die (wegen Nichtberücksichtigung des Absolutglieds) bisher fehlerhaft waren, korrigiert. Dies betrifft auch die Ableitung entsprechender Korrekturschätzer.
- Außerdem wurde der Abschnitt 10 (neu) über Paneldaten um einen Unterabschnitt ergänzt, in dem die Auswirkungen der Korrelation der individuellen Effekte mit den Regressoren (Nichtexogenität der Regresoren) betrachtet werden (Abschnitt 10.6 neu).
- In Appendix C wird jetzt ein alternativer Beweis für den naiven Schätzer im Fall der multiplikativen Überlagerung präsentiert. Die entsprechenden Ergebnisse im Abschnitt 9.4 wurden korrigiert.
- In Appendix D.1 wird die Darstellung des Within-Schätzers um die Darstellung der Kovarianzmatrix ergänzt. igiert.
- Ein Beweis für die alternative Ableitung des 'Within'-Schätzers, der im STATA-Handbuch vorgeschlagen wird, findet sich jetzt in Abschnitt D.2.
- Außerdem wird in Abschnitt D.5 eine rechentechnisch günstigere Darstellung des Schätzers beschrieben.
- In Appendix D.6 findet sich eine ausführliche Formulierung für den Korrekturschätzer im **multiplen** Regressionsmodell bei multiplikativer Überlagerung mit tels der "speziellen" Höhne-Spezifikation.
- Ein neuer Appendix E betrachtet nun die multiplikative Überlagerung allgemein und geht auf die Unterschiede zur Höhne-Überlagerung ein bei Schätzung mittels des Within-Schätzers ein.

[2] Siehe Ronning et al (2005).

Ferner wurden einige Fehler korrigiert.

2 Eindimensionale Mischungsverteilungen

2.1 Stetige Verteilung

Es seien V_1 und V_2 zwei stetige Zufallsvarible mit Dichtefunktionen $f_1(v)$ und $f_2(v)$ sowie Erwartungswerten μ_i und Varianzen σ_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

$$g(u) \quad = \quad \alpha\, f_1(u) \;+\; (1 \,-\, \alpha)\, f_2(u) \qquad , \quad 0 < \alpha < 1 \quad , \tag{2-1}$$

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 α_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.[3] Diese Modellvorstellung nutzt man aus, um Zufallszahlen aus Mischungsverteilungen zu erzeugen.

Symbolisch schreiben wir die "Mischung" auch wie folgt:

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

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

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

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

$$E[U] \quad = \quad \alpha\, \mu_1 \;+\; (1 \,-\, \alpha)\, \mu_2 \quad . \tag{2-3}$$

Ferner ergibt sich für den Erwartungswert der quadrierten Zufallsvariablen

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

Daraus folgt für die Varianz von U[4]

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

[3] Siehe z.B. McLachlan und Peel (2000 section 1.4 "Interpretation of Mixture Models").

[4] Im folgenden wird $V[U]$ und σ_u^2 gleichgesetzt. Davon zu unterscheiden ist der später eingeführte Mischungsverteilungs-Parameter σ_m^2.

Die letzte Zeile zeigt, daß nicht die absolute Größe der beiden Parameter μ_1 und μ_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.[5] 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 Beliebig viele stetige (oder diskrete) Zufallsvariable

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

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

und für Erwartungswert und Varianz erhalten wir

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

sowie[6]

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

3 Mehrdimensionale Mischungsverteilungen

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

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

[5] $Var[V_1 + V_2] \;=\; Var[V_1] \, + \, Var[V_2] \, + \, 2\, cov[V_1, V_2]$.

[6] Eine Vereinfachung wie in Zeile 5 von Formel (2-4) für den Spezialfall von zwei Mischungskomponenten ist nicht möglich.

sowie Erwartungsvektor

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

und Kovarianzmatrix

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

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

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

gegeben ist.

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

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

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

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

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

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

Andererseits läßt sich unter Verwendung der Definition (3-1) für die gemeinsame Dichte-

[7] Siehe z.B. die Definition bei Graybill, Introduction to Matrices with Applications in Statistics. 1969. Kap. 10.4.

funktion schreiben:

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

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

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

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

$$
cov[\mathbf{U}] = E\left[\mathbf{U}\,\mathbf{U}'\right] - E\left[\mathbf{U}\right] E\left[\mathbf{U}\right]'
$$

außerdem das folgende Ergebnis herleiten:[8] [9]

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

[8] Man beachte, daß

$$
E\left[\mathbf{U}\,\mathbf{U}'\right] = \begin{pmatrix} E[U_1^2] & E[U_1U_2] & \ldots & E[U_1U_k] \\ E[U_1U_2] & E[U_2^2] & \ldots & E[U_2U_k] \\ \vdots & \vdots & \ddots & \vdots \\ E[U_kU_1] & \ldots & \ldots & E[U_k^2] \end{pmatrix}
$$

und beispielsweise

$$
E[U_1U_2] = \sigma_{12} + \mu_1\mu_2
$$

gilt.

[9] Aus (3-3) folgt, daß das (t,s)-Element der Kovarianz-Matrix (3-3) wie folgt aussieht:

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

Dabei ist $\sigma_i(t,s)$ das (t,s)-Element aus Matrix $\boldsymbol{\Sigma}_\mathbf{i}$ und $\mu_i(t)$ das t-Element aus dem Vektor $\boldsymbol{\mu}_i$. In der Überlagerung zwecks Anonymisierung duerften 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. Dann ergibt sich für die additive Überlagerung

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

und für die multiplikative Überlagerung

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

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

Man beachte, daß auch bei Nullkorrelation in den Ausgangsverteilungen eine von Null verschiedene Korrelation erzeugt wird, sofern nicht alle Elemente des Mittelwertvektors für alle ** gleich Null ist. Daß bei unterschiedlichen Mittelwerten eine von Null verschiedenen Korrelation erzeugt wird, zeigt auch die Graphik 3/1. Man könnte dieses Bild auch als Streudiagramm mit zwei "schwergewichtigen" Ausreißern interpretieren, die die Korrelation erzeugen.

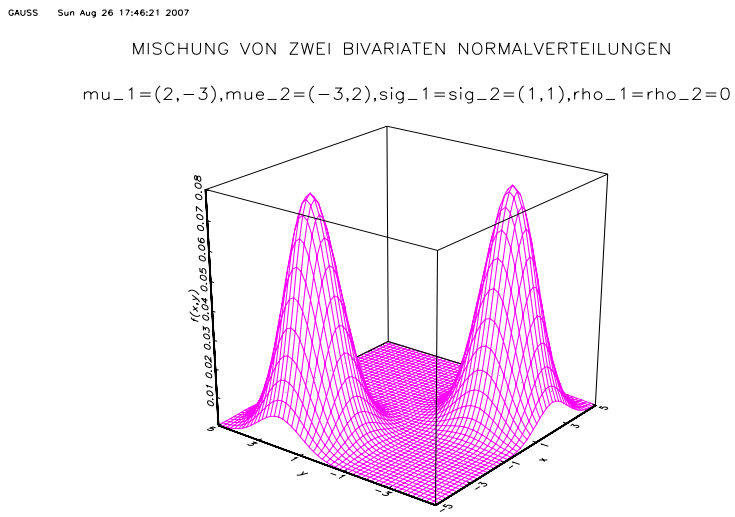

Abbildung 3/1: Mischung von zwei bivariaten Normalverteilungen

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

$$\mathbf{\Sigma}_j \;=\; \sigma_j^2 \begin{pmatrix} 1 & \varrho_j & \varrho_j & \varrho_j \\ \varrho_j & 1 & \varrho_j & \varrho_j \\ \varrho_j & \varrho_j & 1 & \varrho_j \\ \rho & \varrho_j & \varrho_j & 1 \end{pmatrix} \quad = \quad \sigma_j^2 \left[(1-\varrho_j)\,\mathbf{I} \;+\; \varrho_j\,\boldsymbol{\iota}\boldsymbol{\iota}' \right] \tag{3-4}$$

gegeben, wobei ϱ_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 $\mathbf{U}$

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

4 Die Logarithmische Normalverteilung als Mischungsverteilung

4.1 Eindimensionale Verteilung

Ich entnehme die folgenden Ergebnisse meinem Buch 'Statistische Methoden in der empirischen Wirtschaftsforschung'.

Es sei Y eine normalverteilte Zufallsvariable mit $\mathrm{E}(\mathrm{Y}) = \mu$ und $\mathrm{Var}(\mathrm{Y}) = \sigma^2$, d.h. $Y \sim N(\mu, \sigma^2)$. Dann sagen wir, $X := \exp(Y)$ sei lognormalverteilt und schreiben

$$X \sim L(\mu, \sigma^2) \quad . \tag{4-6}$$

Die Dichtefunktion ist gegeben durch

$$f(x) = \begin{cases} \frac{1}{x\,\sigma\,\sqrt{2\,\pi}} \cdot \exp\left\{-\frac{1}{2\,\sigma^2} \quad [\log_e(x) - \mu]^2\right\} & , x > 0 \\ 0 & , \text{sonst} \end{cases} \tag{4-7}$$

und für Erwartungswert und Varianz gilt:

$$\mathrm{E}(\mathrm{X}) = \exp(\mu + \frac{\sigma^2}{2}) \tag{4-8}$$

$$\mathrm{Var}(\mathrm{X}) = \exp(2\,\mu + \sigma^2)\,(\exp(\sigma^2) - 1) = \exp(2\,\mu + 2\sigma^2) - \exp(2\,\mu + \sigma^2) \tag{4-9}$$

Man beachte bei der Schreibweise in (4-6), daß μ und σ^2 *nicht* Erwartungswert und Varianz der Verteilung sind, sondern daß diese durch (4-8) und (4-9) gegeben sind. Ferner kann man zeigen , daß der Modalwert, also das Maximum der Dichtefunktion durch

$$h = \exp(\mu - \sigma^2) \tag{4-10}$$

gegeben ist, und daß für den Median bzw. Zentralwert

$$z = \exp(\mu) \tag{4-11}$$

gilt.

4.2 Multivariate Verteilung

Die multivariate Normalverteilung wird beispielsweise bei S.J. Press (Applied Multivariate Analysis, 1972 , Kap. 6.4) behandelt. Sie wird durch den Vektor $\boldsymbol{\mu}$ und die (positiv definite) Matrix $\boldsymbol{\Sigma}$ charakterisiert.

Für die gemeinsame Dichtefunktion des r-dimensionalen Vektors ergibt sich

$$f(v_1 \ldots, v_r) = \frac{1}{(2\pi)^{r/2}\,(\det(\boldsymbol{\Sigma})^{1/2} \prod_{s=1}^{r} v_s} \exp\left\{-\frac{1}{2}(\overline{\log \mathbf{v}} - \boldsymbol{\mu})'\,\boldsymbol{\Sigma}^{-1}\,(\overline{\log \mathbf{v}} - \boldsymbol{\mu})\right\}$$

Dabei ist

$$\overline{\log \mathbf{v}} \equiv \begin{pmatrix} \log(v_1) \\ \vdots \\ \log(v_r) \end{pmatrix}$$

Für die Erwartungswerte gilt

$$E[V_s] \quad = \quad \exp\left\{\mu_s + \frac{\sigma_s^2}{2}\right\}$$

und für die Kovarianzen ergibt sich

$$cov[V_s, V_t] \quad = \quad \exp\left\{\mu_s + \mu_t + \frac{\sigma_s^2 + \sigma_t^2}{2} + \rho_{st}\,\sigma_s\,\sigma_t\right\} - \exp\left\{\mu_s + \mu_t + \frac{\sigma_s^2 + \sigma_t^2}{2}\right\}$$

Daraus folgt speziell für die Varianzen ($t = s$)

$$var[V_s] \quad = \quad \exp\left\{2\,\mu_s + 2\,\sigma_s^2\right\} - \exp\left\{2\,\mu_s + \sigma_s^2\right\}$$

Demnach haben die Erwartungswerte und Varianzen im multivariaten Fall dieselbe Struktur wie im univariaten Fall.

Außerdem ist erwähnenswert, daß die Kovarianzen und damit die Korrelationskoeffizienten gleich Null sind, wenn die Parameter $\rho_{ts} = 0$ gesetzt werden. Andererseits impliziert die Parameterkonstellation $\rho_{ts} = 1$ nicht, daß die lognormalverteilten Zufallsvariablen 'perfekte' Korrelation aufweisen. Denn für $\rho_{ts} = 1$ ergibt sich

$$\begin{aligned} cov[V_s, V_t] &= \exp\left\{\mu_s + \mu_t + \tfrac{\sigma_s^2 + \sigma_t^2}{2} + \sigma_s\,\sigma_t\right\} - \exp\left\{\mu_s + \mu_t + \tfrac{\sigma_s^2 + \sigma_t^2}{2}\right\} \\ &= \exp\left\{\mu_s + \tfrac{\sigma_s^2}{2}\right\} \exp\left\{\mu_t + \tfrac{\sigma_t^2}{2}\right\} (\exp\{\sigma_s\,\sigma_t\} - 1) \end{aligned}$$

Die Korrelation wäre gleich 1, wenn dieser Ausdruck gleich dem Produkt der Strandardabweichungen wäre. Jedoch ist dieser durch einen zwar sehr ähnlich aussehenden, aber abweichenden Ausdruck gegeben:

$$\sqrt{var[V_s, V_t]} \quad = \quad \exp\left\{\mu_s + \frac{\sigma_s^2}{2}\right\} \exp\left\{\mu_t + \frac{\sigma_t^2}{2}\right\} \sqrt{(\exp\{\sigma_s\} - 1)\,(\exp\{\sigma_t\} - 1)}$$

Damit die Korrelation stets kleiner oder gleich 1 ist, muß demnach der zuerst gegebene Ausdruck größer sein als der zweite, bzw. es muß

$$(\exp\{\sigma_s\,\sigma_t\} - 1) \quad > \quad \sqrt{(\exp\{\sigma_s\} - 1)\,(\exp\{\sigma_t\} - 1)}$$

gelten. Für den Fall $\sigma_s = \sigma_t$ ist dies offensichtlich der Fall.[10] Anders gesagt bedeutet dies, daß die Korrelation **nie** gleich 1 sein kann!!

4.3 Eindimensionale Mischungsverteilung basierend auf der Lognormalverteilung

Eine eindimensionale Mischungsverteilung, deren **zwei** Komponenten lognormalverteilt sind, ergibt sich aus der allgemeinen Formel (2-1) unter Verwendung der speziellen Formel (4-7) für die Dichtefunktion der Lognormalverteilung. Ein Beispiel findet sich in Abbildung 4/2, rechte Seite. Dabei sind die Parameter wie folgt gewählt:[11]

$$\mu_1 = -0.2957, \sigma_1^2 = 0,1452\,,\ \mu_2 = 0,6688, \sigma_2^2 = 0,0488,$$

[10] Beweis für den allgemeinen Fall ???

[11] Dieses Beispiel verwendet auch Ronning (2005). Siehe Fußnote Seite 77.

d.h. die beiden Komponenten haben die folgenden Erwartungswerte:

$$E[V_1] = \exp\{-0,2231\} \;=\; 0,80 \quad , \quad E[V_2] = \exp\{0,6932\} \;=\; 2,00$$

Dies läßt sich optisch aus der Abbildung verifizieren. Man beachte, daß diese Mischungsverteilung stets nur positive Werte annimmt!

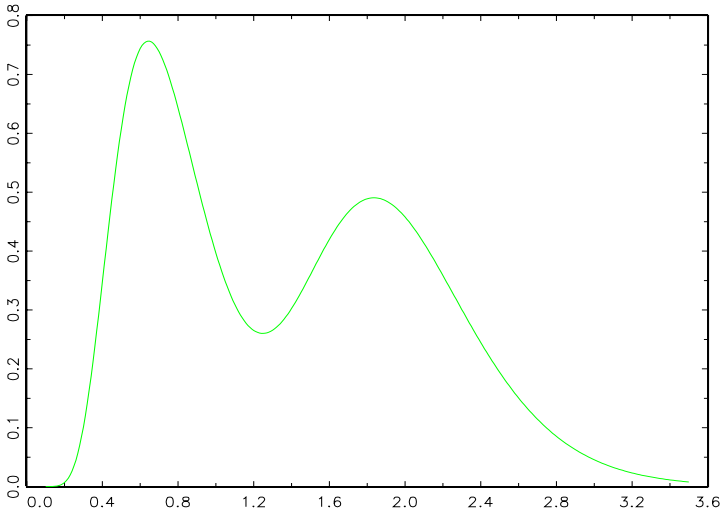

Abbildung 4/2: Mischung von zwei Logarithmischen Normalverteilungen

Für die multiplikative Überlagerung, bei der diese Mischungsverteilung eingesetzt werden soll, wird verlangt, daß der Erwartungswert dieser Verteilung gleich 1 ist. Formaler soll gelten:

$$0.5 \exp\left\{\mu_1 + \frac{\sigma_1^2}{2}\right\} + 0.5 \exp\left\{\mu_2 + \frac{\sigma_2^2}{2}\right\} = 1 \quad .$$

Dies ließe sich am einfachsten dadurch erreichen, daß man

$$\mu_1 \;=\; -\sigma_1^2/2 \;= \quad \text{und} \quad \mu_2 \;=\; -\sigma_2^2/2$$

setzt. Allerdings ist die resultierende Mischungsverteilung **eingipflig** und damit für unsere Zwecke wertlos.

Eine unter vielen Varianten, die zu einer bimodalen Verteilung führen, ist wie folgt: Wir setzen

$$\sigma_1^2 \;=\; \sigma_2^2 \;=\; \sigma^2 \;,\; \sigma^2 \text{ ein beliebiger positiver Wert}$$

und schreiben die obige Bedingung als

$$0.5 \cdot \exp\left\{\sigma^2/2\right\} \; (\exp\{\mu_1\} \;+\; \exp\{\mu_2\}) \quad = \quad 1$$

Sodann wählen wir zwei positive Zahlen c und d mit der Restriktion $c + d < 2$ und setzen

$$\mu_1 \;=\; \log(c) \quad \text{und} \quad \mu_2 \;=\; \log(d).$$

Dann läßt sich die Bedingung wie folgt schreiben:

$$c \;+\; d \quad = \quad 2 \cdot \exp\left\{-\sigma^2/2\right\}$$

Daraus bestimmen wir den Wert für σ^2 wie folgt:

$$\sigma^2 = 2\ (\log(2) - \log(c+d))$$

Allerdings ist die Restriktion, daß c und d positiv und in der Summe kleiner als 2 sein sollen, vermutlich eine starke Restriktion. Numerische Beispiele dazu sind geplant.

4.4 Bezug zur Mischungsverteilung basierend auf Normalverteilungen

Für die Mischungsverteilung mit lognomalverteilten Komponenten gilt gemäß (2-1) und (4-7)

$$g(x) \ = \ \alpha \frac{1}{x\,\sigma_1\,\sqrt{2\,\pi}}\ \exp\left\{-\frac{(\log(x)-\mu_1)^2}{2\,\sigma_1^2}\right\} + (1-\alpha)\,\frac{1}{x\,\sigma_2\,\sqrt{2\,\pi}}\ \exp\left\{-\frac{(\log(x)-\mu_2)^2}{2\,\sigma_2^2}\right\} \tag{4-12}$$

Es soll die daraus resultierende Verteilung für die Zufallsvariable $Y = \log(X)$ bestimmt werden. Wir wenden den Transformationssatz für stetige Verteilungen an, der in allgemeiner Form wie folgt lautet:

> Für die Dichte der Zufallsvariablen $Y = h(X)$ gilt
>
> $$\phi_y(y) = \psi_x(v(y))\ |v'(y)| \quad ,$$
>
> wobei $v(y) = h^{-1}(y)$ die Umkehrfunktion von h und $\psi_x(x)$ die Dichtefunktion von X ist.

In unserem Fall gilt $Y = \log(X)$, d.h. $h(x) = \log(x)$ und damit $v(y) = \exp(y)$ bzw. $v'(y) = \exp(y) = \exp(\log(x)) = x$. Wir erhalten also

$$\phi(y) \ = \ \psi(exp(x))\ x$$

Anders gesagt: Die Dichtefunktion von X ist mit x zu multiplizieren, außerdem tritt an die Stelle von x jetzt $\exp(y)$. Damit erhalten wir für die Zufallsvariable $Y = \log(x)$ den folgenden Ausdruck:

$$\begin{aligned}\phi(y) \ &= \ x\ \left(\alpha \frac{1}{x\,\sigma_1\,\sqrt{2\,\pi}}\ \exp\left\{-\frac{(\log(\exp(y))-\mu_1)^2}{2\,\sigma_1^2}\right\} \ + \ (1-\alpha)\,\frac{1}{x\,\sigma_2\,\sqrt{2\,\pi}}\ \exp\left\{-\frac{(\log(\exp(y))-\mu_2)^2}{2\,\sigma_2^2}\right\}\right) \\ &= \ \alpha \frac{1}{\sigma_1\,\sqrt{2\,\pi}}\ \exp\left\{-\frac{(y-\mu_1)^2}{2\,\sigma_1^2}\right\} \ + \ (1-\alpha)\,\frac{1}{\sigma_2\,\sqrt{2\,\pi}}\ \exp\left\{-\frac{(y-\mu_2)^2}{2\,\sigma_2^2}\right\}\end{aligned} \tag{4-13}$$

Dies ist jedoch exakt die Formel für die Mischung von zwei Normalverteilungen mit Erwartungswert

$$E[Y] \quad = \quad \alpha\,\mu_1 \ + \ (1-\alpha)\,\mu_2 \qquad . \tag{4-14}$$

Demnach läßt sich die Mischung von logarithmischen Normalverteilungen durch Logarithmierung in eine Mischung von Normalverteilungen überführen, allerdings mit einem

Erwartungswert, der von der jeweiligen Parameterkonstellation abhängt (und **nicht automatisch den Wert Null annimmt**[12], wenn die Mischung der logarithmischen Normalverteilungen den Erwartungswert Eins hat!!) Beispielsweise ergibt sich bei der oben gewählten Methode zur Festlegung der Parameter als Erwartungswert

$$E[Y] \quad = \quad \alpha \log(c) \; + \; (1-\alpha) \log(d) \qquad .$$

Der Erwartungswert ist nur dann Null, wenn ($\alpha = 0.5$ und)

$$d = 1/c$$

gilt, also z.B. $c = 3/2$ und $d = 2/3$.

Wählt man andererseits - ausgehend von der Mischung von zwei Normalverteilungen - die (multiplikative) Spezifikation a la Höhne, d.h. ($\alpha = 0.5$ und)

$$\mu_1 = 1 + \delta \quad , \quad \mu_2 = 1 - \delta \quad , \; \sigma^2 \; = \delta^2 + \sigma_\varepsilon^2 \quad ,$$

dann ergibt sich für die Mischung der logarithmischen Normalverteilungen

$$E[X] = 0.5 \, \exp\left\{1 + \delta \; + \; \frac{\delta^2 + \sigma_\varepsilon^2}{2}\right\} \; + \; 0.5 \, \exp\left\{1 - \delta \; + \; \frac{\delta^2 + \sigma_\varepsilon^2}{2}\right\}$$

Dieser Ausdruck kann, weil die Varianz σ_ε^2 beliebig gewählt werden kann, auch beliebige Werte annehmen!

Entsprechend erhalten wir für die additive Spezifikation, d.h. ($\alpha = 0.5$ und)

$$\mu_1 = \mu \quad , \quad \mu_2 = -\mu \quad , \; \sigma^2 \; = \mu^2 + \sigma_\varepsilon^2 \quad ,$$

als Erwartungswert

$$E[X] = 0.5 \, \exp\left\{\mu \; + \; \frac{\mu^2 + \sigma_\varepsilon^2}{2}\right\} \; + \; 0.5 \, \exp\left\{-\mu \; + \; \frac{\mu^2 + \sigma_\varepsilon^2}{2}\right\}.$$

5 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.[13] 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

[12] Das gilt übrigens auch bereits für die Beziehung zwischen Normalverteilung und logarithmischer Normalverteilung selbst (ohne Mischung)!

[13] 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.

man zum Zwecke der Anonymisierung beide Mischungskomponenten symmetrisch wählen, insbesondere die Gewichte und die Varianzen gleich groß wählen.

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.

5.1 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.

5.1.1 Additive Überlagerung

Im Fall der additiven Überlagerung wählen wir[14]

$$\mu_1 = \mu_m \ , \ \mu_2 = -\mu_m \tag{5-1}$$

und

$$\sigma_1^2 \ = \ \sigma_2^2 \ = \ \sigma_m^2 \tag{5-2}$$

sowie

$$\alpha_1 \ = \ \alpha_2 \ = \ \frac{1}{2} \tag{5-3}$$

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

$$E[U] \quad = \quad 0$$

und

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

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

5.1.2 Multiplikative Überlagerung

Im Fall der multiplikativen Überlagerung wählen wir die alternative Spezifikation

$$\mu_1 = 1 + \delta_m \ , \ \mu_2 = 1 - \delta_m \tag{5-4}$$

für die Erwartungswerte. Im Übrigen werden die Spezifikationen des additiven Falls beibehalten: Für die Varianzen gilt (5-2) und für die Gewichte gilt (5-3).

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

$$E[U] \quad = \quad 1$$

und

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

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

[14] Der Index "m" steht für Mischung und bezeichnet Parameter, die die einzelnen Komponenten der Mischungsverteilung charakterisieren, wenn sie für alle Mischungskomponenten identisch gewählt werden.

5.2 Mehrdimensionaler Fall

Entsprechend dem Vorgehen im Abschnitt 5.1 beschreiben wir auch für mehdimensionale 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.

5.2.1 Additive Überlagerung

Im Fall der additiven Überlagerung wählen wir

$$\boldsymbol{\mu}_1 = \boldsymbol{\mu}_m \ , \ \boldsymbol{\mu}_2 = -\boldsymbol{\mu}_m \tag{5-5}$$

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

$$\boldsymbol{\Sigma}_1 \ = \ \boldsymbol{\Sigma}_2 \ = \ \boldsymbol{\Sigma}_m \tag{5-6}$$

sowie

$$\alpha_1 \ = \ \alpha_2 \ = \ \frac{1}{2} \tag{5-7}$$

gelten.

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

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

und

$$cov[\mathbf{U}] \quad = \quad \boldsymbol{\Sigma}_m \ + \ \boldsymbol{\mu}_m \boldsymbol{\mu}_m' \quad . \tag{5-8}$$

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

5.2.2 Multiplikative Überlagerung

Im Fall der multiplikativen Überlagerung wählen wir die alternative Spezifikation

$$\boldsymbol{\mu}_1 = \boldsymbol{\iota} + \boldsymbol{\delta}_m \ , \ \boldsymbol{\mu}_2 = \boldsymbol{\iota} - \boldsymbol{\delta}_m \tag{5-9}$$

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 (5-2) und für die Gewichte gilt (5-3).

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

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

und

$$cov[\mathbf{U}] \quad = \quad \boldsymbol{\Sigma}_m \ + \ \boldsymbol{\delta}_m \boldsymbol{\delta}_m' \quad . \tag{5-10}$$

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

$$\boldsymbol{\delta}_m \ = \ \delta_m \, \boldsymbol{\iota} \tag{5-11}$$

gilt, ergibt sich für die Kovarianzmatrix die spezielle Form

$$cov[\mathbf{U}] \quad = \quad \mathbf{\Sigma}_m \; + \; \delta_m^2 \, \boldsymbol{\iota}\boldsymbol{\iota}' \quad .$$

Falls wir weiterhin noch

$$\mathbf{\Sigma}_m \quad = \quad \sigma_m^2 \, \mathbf{I} \tag{5-12}$$

verlangen, erhalten wir

$$cov[\mathbf{U}] \quad = \quad \sigma_m^2 \, \mathbf{I} \; + \; \delta_m^2 \, \boldsymbol{\iota}\boldsymbol{\iota}' \quad .$$

Dieses spezielle Ergebnis wird uns später bei der Betrachtung des Höhne-Verfahrens wieder begegnen. Siehe insbesondere Abschnitt 6.2. Aber auch hier sollte bereits die Bemerkung gemacht werden, daß die Mischungsverteilung mit zwei Komponenten, die mittels des Parameters δ_m symmetrische Zu- und Abschläge erzeugt, dazu führt, daß die r Komponenten miteinander korreliert sind. Denn man kann die Kovarianzmatrix auch wie folgt schreiben:

$$cov[\mathbf{U}] \; = \; (\sigma_m^2 + \delta_m^2) \, \left((1-\rho) \, \mathbf{I} \; + \; \rho \, \boldsymbol{\iota}\boldsymbol{\iota}' \right)$$

Dabei steht ρ für den zuvor abgeleiteten Korrelationskoeffizienten

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

Wohlgemerkt: Obwohl die Verteilungen der einzelnen Mischungskomponenten skalare Kovarianzmatrix besitzen , weist die Mischungsverteilung selbst positive Korrelation aus. Siehe (5-12) in Verbindung mit (5-2).

6 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.

6.1 Ein einziges Merkmal

6.1.1 Additiver Fall

Wir betrachten das Modell

$$U \quad = \quad \mu \, D \; + \; \varepsilon \quad , \tag{6-1}$$

wobei μ ein Parameter mit beliebigem Wert ist. Für die diskrete Zufallsvariable D gilt

$$D \; = \; \begin{cases} +1 & \text{mit Wahrscheinlichkeit } \alpha \\ -1 & \text{mit Wahrscheinlichkeit } 1-\alpha \end{cases} \tag{6-2}$$

und die stetige Zufallsvariable ε ist wie folgt spezifiziert:

$$E[\varepsilon] \;=\; 0 \quad , \quad V[\varepsilon] \;=\; \sigma_\varepsilon^2 \tag{6-3}$$

wobei die Annahme der Normalverteilung üblicherweise hinzutritt. Ferner wird angenommen, daß D und ε 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

$$h(d) \quad = \quad \alpha^{\frac{1+d}{2}}\;(1\;-\;\alpha)^{\frac{1-d}{2}} \tag{6-4}$$

für $d\,\varepsilon\,\{+1\,,\,-1\}$ gegeben.

Für Erwartungswert und Varianz von U gemäß (6-1) ergibt sich dann

$$E[U] \quad = \quad \mu\,(2\alpha - 1) \tag{6-5}$$

sowie

$$V[U] \quad = \quad \mu^2\,V[D] \;+\; V[U] \;=\; 4\,\alpha\,(1-\alpha)\,\mu^2 \;+\; \sigma_\varepsilon^2 \quad . \tag{6-6}$$

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

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

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

6.1.2 Multiplikativer Fall

Im Fall der multiplikativen Überlagerung verwenden wir statt (6-1) den Ansatz

$$U \quad = \quad (1 \;+\; \delta\,D) \;+\; \varepsilon \tag{6-7}$$

und erhalten in diesem Fall für Erwartungswert und Varianz

$$E[U] \quad = \quad 1 \;+\; \delta\,(2\alpha - 1) \tag{6-8}$$

sowie

$$V[U] \quad = \quad \delta^2\,V[D] \;+\; V[U] \;=\; 4\,\alpha\,(1-\alpha)\,\delta^2 \;+\; \sigma_\varepsilon^2 \quad . \tag{6-9}$$

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

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

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

6.1.3 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

$$\varepsilon \sim N(0\,,\,\sigma_\varepsilon^2) \qquad . \tag{6-10}$$

Additiver Fall: Für gegebenes D ist U in (6-1) 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

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

mit $h(d)$ aus (6-4). Dann erhalten wir die Randdichte von U durch folgende Formel:

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

oder

$$f(u) \quad = \quad \alpha\, N(\mu_1\,,\,\sigma_\varepsilon^2) \,+\, (1-\alpha)\, N(\mu_2\,,\,\sigma_\varepsilon^2) \quad . \tag{6-11}$$

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

Multiplikativer Fall: Für gegebenes D ist U in (6-7) normalverteilt, d.h.

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

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

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

mit $h(d)$ aus (6-4)und für die Randdichte von U ergibt sich

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

oder

$$f(u) \quad = \quad \alpha\, N(1+\delta\,,\,\sigma_\varepsilon^2) \,+\, (1-\alpha)\, N(1-\delta\,,\,\sigma_\varepsilon^2) \quad . \tag{6-12}$$

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

6.2 Mehrere Merkmale gemeinsam

Höhne hat vorgeschlagen, alle Merkmale mit demselben (multiplikativen) Zu- bzw. Abschlag δ 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. [15] 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 9 ebenfalls benötigt wird.

6.2.1 Formale Darstellung des Verfahrens

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

$$U_j \quad = \quad (1 + \delta\, D) \; + \; \varepsilon_j \quad , j = 1, \ldots, r\,, \tag{6-13}$$

Für alle r Komponenten gemeinsam kann man das auch schreiben als

$$\mathbf{U} \quad = \quad (1 + \delta\, D)\, \boldsymbol{\iota} \; + \; \boldsymbol{\varepsilon} \quad , \tag{6-14}$$

wobei δ ein Parameter mit beliebigem Wert ist. Genau wie in Abschnitt 6.1.1 ist die diskrete Zufallsvariable D durch

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

gegeben (siehe (6-2)) und der r-dimensionale stetige Zufallsvektor $\boldsymbol{\varepsilon}$ ist entsprechend (6-3) spezifiziert:

$$E[\boldsymbol{\varepsilon}] \; = \; \mathbf{0} \quad , \quad cov[\boldsymbol{\varepsilon}] \; = \; \sigma_\varepsilon^2\, \mathbf{I} \tag{6-15}$$

wobei die Annahme der Normalverteilung üblicherweise hinzutritt. Genau wie in Abschnitt 6.1.1 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)

$$E[U_j] \quad = \quad 1 \; + \; \delta\,(2\alpha - 1) \qquad , \quad j = 1, \ldots, r \tag{6-16}$$

oder auch kompakter

$$E[\mathbf{U}] \quad = \quad (1 \; + \; \delta\,(2\alpha - 1))\; \boldsymbol{\iota} \qquad . \tag{6-17}$$

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

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

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

[15] Darauf werden wir in Abschnitt 8 näher eingehen.

korreliert, selbst wenn die Komponenten ε_j des Zufallsvektors $\boldsymbol{\varepsilon}$ unkorreliert sind!! Um die Kovarianz zu bestimmen, betrachten wir zunächst den Erwartungswert des Produktes $U_i U_j$. Dafür erhalten wir[16]

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

und für die Kovarianz selbst erhalten wir

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

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 δ^2.

Für die Varianz von U_j gilt ferner (siehe die Ergebnisse in Abschnitt 6.1.2)

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

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

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

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

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

Die durch das Verfahren erzeugte Korrelation ist also umso größer, je größer der gemeinsame Zuschlag δ und je kleiner die Überlagerungsvarianz σ_ε^2 ist!

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

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

die sich im 'symmetrischen' Fall auf die einfache Form

$$cov[\mathbf{U}] \;=\; \sigma_\varepsilon^2\mathbf{I}\;+\;\delta^2\,\boldsymbol{\iota}\boldsymbol{\iota}' \;=\; \frac{1}{\sigma_\varepsilon^2+\delta^2}\left((1-\rho)\,\mathbf{I}\;+\;\rho\,\boldsymbol{\iota}\boldsymbol{\iota}'\right) \tag{6-18}$$

reduziert, wobei ρ für den zuvor abgeleiteten Korrelationskoeffizienten

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

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

[16] Wir verwenden

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

6.2.2 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

$$E[\mathbf{U}] \quad = \quad \boldsymbol{\iota} \quad \text{und} \quad cov[\mathbf{U}] \; = \; \sigma_\varepsilon^2 \mathbf{I} \; + \; \delta^2 \, \boldsymbol{\iota}\boldsymbol{\iota}' \tag{6-19}$$

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

$$\boldsymbol{\mu}_j = \delta \boldsymbol{\iota} \quad \text{und} \quad \boldsymbol{\Sigma}_j = \sigma_\varepsilon^2 \mathbf{I} \tag{6-20}$$

und der 'symmetrische Fall'

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

sowohl bei der Mischungsverteilung als auch im Höhneverfahren unterstellt wird.

Es bleibt zu prüfen, ob auch im mehrdimensionalen Fall die Dichtefunktion des Höhneverfahren mit der Definition der Mischungsverteilung gemäß (3-1) übereinstimmt. Bei dieser Analyse gehen wir wie im eindimensionalen Fall vor. Siehe dazu Abschnitt 6.1.3. 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 U zu erhalten. Dabei unterstellen wir im folgenden wieder zusätzlich Normalverteilung:

$$\boldsymbol{\varepsilon} \; \sim \; N(\mathbf{0} \, , \, \sigma_\varepsilon^2 \mathbf{I}) \qquad . \tag{6-21}$$

Für gegebenes D ist $\mathbf{U}$ in (6-14) normalverteilt, d.h.

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

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

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

mit $h(d)$ aus (6-4)und für die Randdichte von $\mathbf{U}$ ergibt sich

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

oder

$$f(\mathbf{u}) \quad = \quad \alpha \, N((1 \, + \, \delta)\boldsymbol{\iota} \, , \, \sigma_\varepsilon^2 \mathbf{I}) \; + \; (1 - \alpha) \, N((1 \, - \, \delta)\boldsymbol{\iota} \, , \, \sigma_\varepsilon^2 \mathbf{I}) \quad . \tag{6-22}$$

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

7 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

$$E[Y] \quad = \quad \mu_y \quad \text{sowie} \quad V[Y] \quad = \quad \sigma_y^2 \tag{7-1}$$

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

$$E[\mathbf{Y}] \quad = \quad \boldsymbol{\mu}_y \quad \text{sowie} \quad cov[\mathbf{Y}] \quad = \quad \boldsymbol{\Sigma}_{yy} \tag{7-2}$$

7.1 Additive Überlagerung

Im eindimensionalen Fall gilt

$$Y^a \quad = \quad Y \quad + \quad U \tag{7-3}$$

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

$$\mathbf{Y}^a \quad = \quad \mathbf{Y} \quad + \quad \mathbf{U} \tag{7-4}$$

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

Für Erwartungswert und Varianz bzw. Kovarianzmatrix ergibt sich

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

und

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

Falls die Überlagerung mittels der Mischungsverteilung aus Abschnitt 5.1.1 (bzw. mit dem Höhne-Verfahren) erfolgt, ergibt sich - im eindimensionalen Fall - für die Varianz der überlagerten Variablen

$$V[Y^a] \quad = \quad \sigma_y^2 \; + \; \sigma_m^2 \; + \; \mu_m^2 \tag{7-5}$$

wobei μ_m und σ_m^2 die Parameter aus der Mischungsverteilung sind, siehe (5-1).

Im multivariaten Fall ergibt sich aus (5-8) als Kovarianzmatrix der überlagerten Variablen

$$cov[\mathbf{Y}^a] \quad = \quad cov[\mathbf{Y}] \; + \; \boldsymbol{\Sigma}_m \; + \; \boldsymbol{\mu}_m \boldsymbol{\mu}_m' \quad , \tag{7-6}$$

wobei $\boldsymbol{\mu}_m$ und $\boldsymbol{\Sigma}_m$ die Parameter aus der Mischungsverteilung bezeichnen. Siehe Abschnitt 5.2.1.

7.2 Multiplikative Überlagerung in univariaten Verteilungen

Üblicherweise wird man bei der multiplikativen Überlagerung unterstellen, daß sowohl die zu überlagernde Variable Y als auch die Überlagerun gsvariable 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

$$Y^a \quad = \quad Y\,U \tag{7-7}$$

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:

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

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

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

und damit für den unbedingten Erwartungswert

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

Ferner erhalten wir für die bedingte Varianz

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

und damit für die unbedingte Varianz[17]

$$V[Y^a] \quad = \quad \sigma_y^2 \,+\, \sigma_u^2\,(\mu_y^2 \,+\, \sigma_y^2)\,. \tag{7-8}$$

Falls die Überlagerung mittels der Mischungsverteilung aus Abschnitt 5.1.2 (bzw. mit dem Höhne-Verfahren) erfolgt, ergibt sich - im eindimensionalen Fall - für die Varianz der überlagerten Variablen

$$V[Y^a] \quad = \quad \sigma_y^2 \,+\, (\sigma_m^2 \,+\, \delta_m^2)\,(\mu_y^2 \,+\, \sigma_y^2) \tag{7-9}$$

wobei δ_m der Mittelwertparameter und σ_m^2 der Varianzparameter aus der Mischungsverteilung sind, siehe (5-4) und (5-2).

7.3 Multiplikative Überlagerung in multivariaten Verteilungen

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

[17] Es wird die Formel für bedingte und unbedingte Varianzen verwendet:

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

Komponenten des Merkmals-Vektors mit **demselben Faktor** multiplikativ überlagert. Dies hat den Vorteil, daß bei der Bildung von Quotienten dieser Faktor tendenziell (siehe jedoch Genaueres in Abschnitt 8) eliminiert wird, d.h. daß die Quotienten durch diese Art der multiplikativen Überlagerung nicht verändert werden!! Formal ergibt sich in diesem Fall

$$\mathbf{Y}^a \quad = \quad U\,\mathbf{Y} \qquad \text{(konstante multiplikative Überlagerung)} \tag{7-10}$$

Dabei soll für die (nichtnegative) Störvariable U gelten:

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

Im Fall der eigentlichen multiplikativen Überlagerung erhalten wir dagegen

$$\mathbf{Y}^a \quad = \quad \mathbf{U} \odot \mathbf{Y} \qquad \text{(eigentliche multiplikative Überlagerung)} \tag{7-11}$$

wobei $\mathbf{U}$ ein **positiver** Zufalls-Vektor mit

$$E[\mathbf{U}] = \boldsymbol{\iota} \quad \text{und} \quad cov[\mathbf{U}] \text{ eine beliebige positiv definite Matrix}$$

ist und $\odot$ das Hadamard-Produkt bezeichnet.[18]

Es sind aber auch allgemeine Spezifikationen denkbar.

- Beispielsweise lassen sich anonymisierte Variablen als Linearkombinationen von überlagerten Originalvariablen bilden:

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

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

- Schließlich kann man auch noch

$$\mathcal{Y}^a \;=\; \mathcal{U} \;\odot\; \mathcal{Y} \tag{7-12}$$

  betrachten, wobei die Zufalls-Matrizen $\mathcal{Y}$ und $\mathcal{U}$ identische Dimensionen besitzen. Wir bezeichnen dies als variablenspezifische multiplikative Überlagerung bei Matrizen.

7.3.1 Überlagerung mit Hilfe eines konstanten Faktors

Im "konstanten" Fall ergibt sich aus (7-10) für den bedingten Erwartungswert

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

und damit für den unbedingten Erwartungswert

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

[18] Siehe Marshall und Olkin (1979 Seite 257-8) für einige Eigenschaften des Hadamard-Produktes.

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

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

sowie

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

Insgesamt gilt also für die bedingte Kovarianzmatrix

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

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

$$\begin{aligned} cov[\mathbf{Y}^a] \quad &= \quad E[cov[\mathbf{Y}^a|\mathbf{Y}]] \ + \ cov[E[\mathbf{Y}^a|\mathbf{Y}]] \ = \ \sigma_u^2\, E[\mathbf{Y}\, \mathbf{Y}'] \ + \ cov[\mathbf{Y}] \\ &= \quad \sigma_u^2 \left\{ cov[\mathbf{Y}] \ + \ \boldsymbol{\mu}_y\, \boldsymbol{\mu}_y' \right\} \ + \ cov[\mathbf{Y}] \end{aligned} \tag{7-13}$$

mit

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

Man beachte die Entsprechung zu (7-8) für den eindimensionalen Fall.

7.3.2 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 von $\mathbf{U}$ eine Rolle. Wir untersuchen zunächst den einfachsten Fall, in dem alle U_k identische Varianz σ_u^2 besitzen und miteinander unkorreliert sind.

Aus (7-11) ergibt sich für den bedingten Erwartungswert

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

und damit für den unbedingten Erwartungswert

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

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

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

sowie

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

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

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

Zur Berechnung der unbedingten Kovarianzmatrix verwenden wir wieder

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

und erhalten in diesem Fall

$$\begin{aligned} cov[\mathbf{Y}^a] &= \sigma_u^2\, E\left[\begin{pmatrix} Y_1^2 & & & & \\ & Y_2^2 & & & \\ & & \ddots & & \\ & & & Y_{k-1}^2 & \\ & & & & Y_k^2 \end{pmatrix}\right] + cov[\mathbf{Y}] \\ &= \sigma_u^2 \begin{pmatrix} \sigma_{y(1)}^2 + \mu_{y(1)}^2 & & & & \\ & \sigma_{y(2)}^2 + \mu_{y(2)}^2 & & & \\ & & \ddots & & \\ & & & \sigma_{y(k-1)}^2 + \mu_{y(k-1)}^2 & \\ & & & & \sigma_{y(k)}^2 + \mu_{y(k)}^2 \end{pmatrix} \\ &\quad + \quad cov[\mathbf{Y}] \end{aligned} \tag{7-14}$$

mit

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

Man beachte auch hier die Entsprechung zu (7-8) im eindimensionalen Fall.

Demnach stimmen die Varianzen des "eigentlichen" 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 (7-13) mit (7-14). Wegen der identischen Varianzen bedeutet dies eine geringere Korrelation im Fall der "eigentlichen" Überlagerung.

7.3.3 Überlagerung mit Hilfe von korrelierten Störvariablen

Wir untersuchen nun den später interessierenden Fall, daß die Störvariablen U_j des Zufallsvektors $\mathbf{U}$ miteinander korreliert sind. Die Kovarianzmatrix insgesamt soll jetzt allgemein durch $cov[\mathbf{U}]$ gegeben sein. Dies betrifft insbesondere die Überlagerung mit Hilfe einer Mischungsverteilung, auf die wir im folgenden Abschnitt 7.3.4 eingehen, was dann jedoch eine spezielle Struktur der Kovarianzmatrix zur Folge hat.

Wieder gilt für den unbedingten Erwartungswert

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

Ferner erhalten wir für die bedingte Varianz-Kovarianz-Matrix wegen

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

sowie

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

und somit für die bedingte Kovarianzmatrix insgesamt:

$$cov[\mathbf{Y}^a|\mathbf{Y}] \quad = \quad cov[\mathbf{U}] \odot \mathbf{Y}\,\mathbf{Y}' \quad .$$

Zur Berechnung der unbedingten Kovarianzmatrix verwenden wir wieder

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

und erhalten in diesem Fall

$$\begin{aligned} cov[\mathbf{Y}^a] &= cov[\mathbf{U}] \odot E[\mathbf{Y}\,\mathbf{Y}'] + cov[\mathbf{Y}] \\ &= cov[\mathbf{U}] \odot (cov[\mathbf{Y}] + \boldsymbol{\mu}_y\boldsymbol{\mu}_y') \quad + \quad cov[\mathbf{Y}] \quad . \end{aligned} \tag{7-15}$$

7.3.4 Überlagerung mit Hilfe der Mischungsverteilung

Wir haben in Abschnitt 5.2 und speziell für multiplikative Überlagerung im Unterabschnitt 5.2.2 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 Mischungsverteilug 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 Höhneverfahren bezeichnen: Alle Merkmale werden mit demselben Zuschlag versehen, d.h. gemäß (6-13) soll gelten:

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

Ferner sollen die Störterme ε_j den "klassischen" Annahmen genügen:

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

Siehe dazu (6-3). Dies führt zu folgender Kovarianzmatrix:

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

wobei ρ durch

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

gegeben ist. Siehe (6-18). Demnach gilt für Varianzen und Kovarianzen:

$$V[U_j] = \sigma_\varepsilon^2 + \delta^2 \quad \text{und} \quad cov[U_j, U_k] = \delta^2 \quad . \tag{7-16}$$

Dies entspricht der speziellen multiplikativen Überlagerung, wie sie in Abschnitt 5.2.2 beschrieben wurde.[19] Für diesen speziellen Fall wollen wir deshalb im folgenden Erwartungsvektor und Kovarianzmatrix des Vektors $\mathbf{Y}^a$ bestimmen.

Es gilt wieder

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

[19] Dort wurden die etwas abweichenden Bezeichnungen σ_m^2 und δ_m gewählt.

Deshalb ergibt sich aus (7-11)für den bedingten Erwartungswert

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

und damit für den unbedingten Erwartungswert

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

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

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

und für die bedingten Kovarianzen erhalten wir

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

Insgesamt ergibt sich also für die bedingte Kovarianzmatrix :

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

Zur Berechnung der unbedingten Kovarianzmatrix verwenden wir wieder

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

und erhalten in diesem Fall

$$cov[\mathbf{Y}^a] \quad = \quad \left(\sigma_\varepsilon^2\mathbf{I} + \delta^2\,\boldsymbol{\iota}\boldsymbol{\iota}'\right) \odot \left(cov[\mathbf{Y}] + \boldsymbol{\mu}_y\boldsymbol{\mu}_y'\right) + cov[\mathbf{Y}] \qquad (7\text{-}17)$$

Alternativ könnte man in dieser Formel die Parameter σ_m^2 und δ_m verwenden. Die Entsprechung zu (7-8) im eindimensionalen Fall ist hier nur noch bedingt gegeben.

7.3.5 Zusammenfassung

Die Ergebnisse der Unterabschnitte von Abschnitt 7.3 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 Fällen eine Kovarianzmatrix mit additiver Struktur: Zu der Kovarianzmatrix $cov[\mathbf{Y}]$ wird eine zweite Matrix addiert, in der die ersten und zweiten Momente von $\mathbf{Y}$ sowie die Parameter der Überlagerungsvariablen eine Rolle spielen, d.h. es tritt eine zweite positiv semi-definite Matrix zu der Matrix $cov[\mathbf{Y}]$ hinzu, was den (unterschiedlich großen) Effizienzverlust charakterisiert.

7.4 Korrelation bei multiplikativer Überlagerung

Es soll nun die Auswirkung der Überlagerung auf den Korrelationskoeffeizienten eingehender untersucht werden. Dabei konzentrieren wir uns auf die Resultate aus Abschnitt 7.3.4, in dem Mischungsverteilungen vom Höhne-Typ verwendet werden. Insbesondere ist hier

die daraus resultierende Formel (7-17)für die Kovarianzmatrix relevant. Aus dieser Formel ergibt sich (unter der Annahme konstanter Korrelation ρ für alle Paare)

$$\begin{aligned} V[Y_j^a] &= \left(\delta^2 + \sigma_\varepsilon^2\right)\left(\sigma_j^2 + \mu_j^2\right) + \sigma_j^2 \\ cov[Y_j^a, Y_k^a] &= \delta^2\left(\rho\,\sigma_j\,\sigma_k + \mu_j\,\mu_k\right) + \rho\,\sigma_j\,\sigma_k \end{aligned} \tag{7-18}$$

sowie

$$corr[Y_j^a, Y_k^a] = \frac{\rho\,(1+\delta^2)\,\sigma_j\,\sigma_k + \delta^2\,\mu_j\,\mu_k}{\sqrt{\left\{(\delta^2+\sigma_\varepsilon^2)\left(\sigma_j^2+\mu_j^2\right)+\sigma_j^2\right\}\left\{(\delta^2+\sigma_\varepsilon^2)\left(\sigma_k^2+\mu_k^2\right)+\sigma_k^2\right\}}} \tag{7-19}$$

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 ρ 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 μ_j und μ_k ab. Setzt man andererseits den Zählerausdruck von (7-19) gleich Null, so ergibt sich daraus

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

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

Aus der Formel (7-19) 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:

$$corr[Y_j^a, Y_k^a] = \frac{\rho\,(1+\delta^2)\,\gamma_j\,\gamma_k + \delta^2}{\sqrt{\left\{(\delta^2+\sigma_\varepsilon^2)\left(\gamma_j^2+1\right)+\gamma_j^2\right\}\left\{(\delta^2+\sigma_\varepsilon^2)\left(\gamma_k^2+1\right)+\gamma_j^2\right\}}} \tag{7-20}$$

mit

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

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

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

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

d.h. es hängt von nur von δ sowie γ_j und γ_k ab, ob beide Korrelationskoeffizienten das gleiche oder unterschiedliche Vorzeichen besitzen. Da δ in der Anonymisierungspraxis stets kleiner als 1 und üblicherweise einen Wert von etwa 0,10 aufweist (und damit $\delta^2 = 0,01$ 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 Orginalvariablen exakt positiv bzw. exakt negativ korreliert sind. Im Fall $\rho = 1$ erhalten wir aus (7-19)

$$corr[Y_j^a\,,\,Y_k^a]_{\rho=1} \;=\; \frac{(1\,+\,\delta^2)\,\sigma_j\,\sigma_k \;+\; \delta^2\,\mu_j\,\mu_k}{\sqrt{\left\{(\delta^2\,+\,\sigma_\varepsilon^2)\,\left(\sigma_j^2\,+\,\mu_j^2\right)\,+\,\sigma_j^2\right\}\,\left\{(\delta^2\,+\,\sigma_\varepsilon^2)\,\left(\sigma_k^2\,+\,\mu_k^2\right)\,+\,\sigma_k^2\right\}}} \tag{7-21}$$

und im Fall $\rho = -1$

$$corr[Y_j^a\,,\,Y_k^a]_{\rho=-1} \;=\; \frac{\delta^2\,\mu_j\,\mu_k \;-\; (1\,+\,\delta^2)\,\sigma_j\,\sigma_k}{\sqrt{\left\{(\delta^2\,+\,\sigma_\varepsilon^2)\,\left(\sigma_j^2\,+\,\mu_j^2\right)\,+\,\sigma_j^2\right\}\,\left\{(\delta^2\,+\,\sigma_\varepsilon^2)\,\left(\sigma_k^2\,+\,\mu_k^2\right)\,+\,\sigma_k^2\right\}}} \tag{7-22}$$

Vor allem die Formel (7-22) besagt, daß die Korrelation der überlagerten Variablen durchaus positiv sein kann, obwohl die Orginalvariablen 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 Orginalvariablen die Korrelation der überlagerten Variablen stets kleiner als 1 sein. Dies sieht man deutlicher, wenn man

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

setzt und dann die rechte Seite von (7-21) 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 unmitelbar zu sehen, daß für $\sigma_\varepsilon^2 > 0$ der Ausdruck kleiner als 1 ist.[20]

8 Einfluß der Überlagerung auf Quotienten

8.1 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

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

nicht zu sehr vom Erwartungswert der Zufallsvariablen

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

abweicht.[21]

[20] Beweis für den allgemeinen Fall steht noch aus!

[21] Man könnte auch die stärkere Forderung aufstellen, daß die Verteilung von Z nicht zu sehr von der Verteilung von Z^a abweichen soll.

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

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

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

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

was allerdings meist auch nicht sehr gut analysierbar ist.[22]

Allerdings läßt sich - für den allgemeinen Fall - folgende Approximation angeben:[23]

$$E[\frac{Y_1}{Y_2}] \;\approx\; \frac{E[Y_1]}{E[Y_2]} \;-\; \frac{1}{(E[Y_2])^2}\, cov[Y_1, Y_2] \;+\; \frac{E[Y_1]}{(E[Y_2])^3}\, V[Y_2] \tag{8-1}$$

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 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 6 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**.

8.2 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

$$U_j \quad = \quad (1 \,+\, \delta_j D_j) \,+\, \varepsilon_j \quad , \quad j = 1, 2 \quad , \tag{8-2}$$

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

$$Y_j^a \quad = \quad Y_j \cdot U_j \quad = \quad Y_j \cdot (1 \,+\, \delta_j D_j \,+\, \varepsilon_j) \quad , \quad j = 1, 2 \quad , \tag{8-3}$$

[22] Von Nutzen ist oft die Jensensche Ungleichung, nach der ganz allgemein für eine nichtnegative Zufallsvariable W gilt:

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

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

[23] 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.

gegeben.

Der Vorschlag von Höhne, der bereits im ersten Anonymisierungsprojekt realisiert wurde[24], läuft darauf hinaus, daß die folgenden Restriktionen eingeführt werden:

$$\delta_1 \;=\; \delta_2 \;=\; \delta \tag{8-4}$$

und

$$D_1 \;=\; D_2 \quad \text{mit Wahrscheinlichkeit 1} \quad . \tag{8-5}$$

Dann ergibt sich für das Verhältnis der anonymisierten Variablen

$$Z^a \quad = \quad \frac{Y_1 \cdot ((1 \,+\, \delta D) \,+\, \varepsilon_1)}{Y_2 \cdot ((1 \,+\, \delta D) \,+\, \varepsilon_2)} \tag{8-6}$$

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

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

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 (8-6) steht, sind auch unter dieser Unabhängigkeitsannahme die beiden Terme nicht stochastisch unabhängig. Jedoch können wir den **bedingten Erwartungswert** für gegebenes D wie folgt bestimmen:

$$\begin{aligned} E[Z^a|D] \quad &= \quad E[Y_1] \cdot ((1 \,+\, \delta D))\, E\left[\frac{1}{Y_2 \cdot ((1 \,+\, \delta D) \;+\, \varepsilon_2)}\right] \\ &= \quad E[Y_1] \cdot ((1 \,+\, \delta D))\, E\left[\frac{1}{Y_2}\right] \cdot E\left[\frac{1}{((1 \,+\, \delta D) \;+\, \varepsilon_2)}\right] \quad . \end{aligned} \tag{8-7}$$

Für den zweiten und dritten Erwartungswert auf der rechten Seite kann man folgende Approximation angeben[25]:

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

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 ε wird im allgemeinen bei der Anonymisierung nur eine kleine Varianz aufweisen, was eine nur unbedeutende Verzerrung impliziert.

Für den Erwartungswert in (8-7) ergibt sich damit

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

[24] Siehe Ronning et al (2005).

[25] Siehe Mood Graybill Boes 1974 S. 181 sowie Formel (8-1) 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!

Dafür können wir in grober Approximation auch schreiben:

$$\begin{aligned} E[Z^a|D] &\approx (1 + \delta D)E[Y_1] \left(\frac{1}{E[Y_2]}\right) \left(\frac{1}{(1 + \delta D)}\right) + \mathbf{R} \\ &= \frac{E[Y_1]}{E[Y_2]} + \mathbf{R}\,, \end{aligned} \tag{8-9}$$

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 σ_ε^2 gewählt wird. Siehe (8-7).

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.

8.3 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 8.1.

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

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

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 μ_x und μ_y relativ groß gewählt habe.[26]

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: $\rho\ \varepsilon\ \{-0.999\,,\,0\,,\,+0.999\}$ AlleBeispiele wurden mit $n = 100$ Beobachtungen und 50 Wiederholungen simuliert. Die Ergebnisse sind in der folgenden Tabelle zusammengefaßt.[27]

μ_x	μ_y	σ_x	σ_y	ρ	ρ^a	r^a	$\frac{\mu_x}{\mu_y}$	$\overline{z}$	$\sigma_{\overline{z}}$	$\overline{z}^a$	$\sigma_{\overline{z}^a}$	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
ρ^a gibt den theoretischen Korrelationskeoffizienten gemäß (7-19) an.
r^a gibt den empirischen Korrelationskoeffizienten über alle Wiederholungen gemittelt an.
$\overline{z}$ gibt den geschätzten Wert von Z über alle Wiederholungen gemittelt an.
$\sigma_{\overline{z}}$ gibt die entsprechende Standardabweichung an.
$\overline{z}^a$ gibt den geschätzten Wert von Z^a über alle Wiederholungen gemittelt an.
$\sigma_{\overline{z}^a}$ gibt die entsprechende Standardabweichung an.
"bias" gibt die Summe aus dem zweiten und dritten Summanden auf der rechten Seite von (8-1) an.

Als erstes fällt auf, daß bei allen drei Szenarios die Korrelation ρ^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 ρ^a, während für die Originalvariablen $\rho = -0.999$ gilt! Wenn ρ positiv ist, nimmt die Diskrepanz ab, ist aber immer noch deutlich sichtbar. Im übrigen werden die Ergebnisse für ρ^a durch das empirische Maß r^a bestätigt. Bemerkenswert ist die Gleichheit von ρ^a in den Szenarios A und C.

[26] Eine formal bessere Lösung wäre, sowohl für die Ausgangsvariablen als auch für die Überlagerungs-Reste Lognormalverteilungen zu verwenden. Dabei sollten die Ausgangsvariablen im Intervall $[0, \infty)$ und die Überlagerungen im Intervall $[(-(1 + \delta D), \infty)$ variieren.

[27] Alle Ergebnisse wurden mit den GAUSS-Programmen QUOT01.prg, QUOT02.prg und QUOT03.prg (Laptop Ronning, Februar 2007) erzeugt.

Hier wirkt sich aus, daß die Variationskoeffizienten für X und Y in beiden Fällen identisch sind. Aus (7-19) ergibt sich, daß dann das Maß ρ^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 (8-1) 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 ρ und ρ^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.

9 Lineare Modelle mit Fehler in den Variablen aus Mischungsverteilungen

9.1 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:

$$\mathbf{y} \quad = \quad \beta_0 \, \boldsymbol{\iota} \; + \; \mathbf{X} \, \boldsymbol{\beta} \; + \; \boldsymbol{\eta} \tag{9-1}$$

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:

$$E[\boldsymbol{\eta}] \; = \; \mathbf{0} \quad \text{und} \quad cov[\boldsymbol{\eta}] \; = \; \sigma_\eta^2 \, \mathbf{I}. \tag{9-2}$$

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

$$\hat{\boldsymbol{\beta}} \quad = \quad \left(\mathbf{X}' \mathbf{M}_\iota \mathbf{X} \right)^{-1} \mathbf{X}' \mathbf{M}_\iota \mathbf{y} \; = \; \boldsymbol{\beta} \; + \; \left(\mathbf{X}' \mathbf{M}_\iota \mathbf{X} \right)^{-1} \mathbf{X}' \mathbf{M}_\iota \boldsymbol{\eta} \tag{9-3}$$

und seine Kovarianzmatrix ist durch

$$cov[\hat{\boldsymbol{\beta}}] \quad = \quad \sigma_\eta^2 \, \left(\mathbf{X}' \mathbf{M}_\iota \mathbf{X} \right)^{-1} \tag{9-4}$$

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

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

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

$$\hat{\boldsymbol{\beta}}^a \quad = \quad \left(\mathbf{X^{a\prime}M}_\iota\mathbf{X^a}\right)^{-1}\mathbf{X^{a\prime}M}_\iota\mathbf{y^a} \tag{9-5}$$

Für den Fall **unkorrelierter** Fehlervariablen können wir auf Ergebnisse aus Rosemann (2006) 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 6.2) 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

$$\mathbf{y}^a \;=\; \mathbf{y} + \mathbf{u} \quad \text{und} \quad \mathbf{X}^a \;=\; \mathbf{X} + \mathbf{U} \tag{9-6}$$

im Fall der multiplikativen Überlagerung gilt

$$\mathbf{y}^a \;=\; \mathbf{y} \odot \mathbf{u} \quad \text{und} \quad \mathbf{X}^a \;=\; \mathbf{X} \odot \mathbf{U} \tag{9-7}$$

Dabei ist $\mathbf{u}$ ein Zufallsvektor und $\mathbf{U}$ eine Zufallsmatrix. Man beachte, daß im Gegensatz zu den Abschnitten 1 bis 7 hier eine andere Symbolik verwendet wird.[28] Für (9-7) können wir auch schreiben:

$$\mathbf{y}^a \;=\; \mathbf{y} \odot \mathbf{u}_y \quad \text{und} \quad (\mathbf{x}_1^a, \ldots, \mathbf{x}_K^a) \;=\; (\mathbf{x}_1 \odot \mathbf{u}_1, \ldots, \mathbf{x}_K \odot \mathbf{u}_K)\,, \tag{9-8}$$

womit die separate Überlagerung jedes Regressor $\mathbf{x}_k$ besonders gut analysiert werden kann. Es wird sich allerdings herausstellen, daß eine zeilenweise d.h. beobachtungspunktspezifische Darstellung mehr weiterhilft, indem wir (9-6) wie folgt schreiben:

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

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:

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

[28] In Abschnitt 3 haben wir $\mathbf{U}$ als Zufalls**vektor** eingeführt. Jetzt müssen wir die Überlagerung für alle n Beobachtungen modellieren, deshalb die geänderte Symbolik.

9.2 Spezielle und flexible Höhne-Spezifikation

Im folgenden werden wir die spezielle Höhne-Spezifikation (6-13) bzw. (6-14), die einen identischen Zuschlag δ 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:

$$\mathbf{u}\langle i\rangle \quad = \quad (1 + \delta\, D\langle i\rangle)\, \boldsymbol{\iota} \;+\; \boldsymbol{\varepsilon}\langle i\rangle \quad , \quad i = 1, \ \ldots\ , n\ . \tag{9-10}$$

wobei die stochastischen Spezifikationen der Zufallsvariablen D und des Zufallsvektors $\boldsymbol{\varepsilon}$ für jedes i durch (6-2) und (6-3) 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

$$\mathbf{u}\langle i\rangle \quad = \quad (\boldsymbol{\iota} + \delta\, \mathbf{D}\langle i\rangle) \;+\; \boldsymbol{\varepsilon}\langle i\rangle \quad , \quad i = 1, \ \ldots\ , n\ . \tag{9-11}$$

wobei $\mathbf{D}$ jetzt ein r-dimensionaler Zufallsvektor ist, dessen Komponenten jeweils die Spezifikation (6-2) erfüllen und die stochastisch voneinander unabhängig sind. Man kann zeigen, daß Erwartungswerte und Varianzen dieser Spezifikation denen von (9-10) bzw. (6-14) entsprechen. Dagegen sind alle Kovarianzen gleich Null.[29]: Insbesondere gilt für den "symmetrischen" Fall

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

9.3 Additive Überlagerung im linearen Regressionsmodell

9.3.1 Allgemeine Bemerkungen

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 (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

$$\text{plim}\ \frac{1}{n}\, \mathbf{X}'\mathbf{M}_\iota\mathbf{X} \quad \equiv \quad \mathbf{Q} \tag{9-12}$$

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, wie dies auch Rosemann (2006) tut. Der Unterschied besteht allerdings nur darin, daß ich statt der Originalwerte die Abweichungen vom Mittel betrachte, also zentrierte Momente statt der unzentrierten Momente.

[29] Die stochastischen Eigenschaften dieser Spezifikation werden in Appendix B beschrieben! Insbesondeer läßt sich zeigen, daß in diesem Fall keine Äquivalenz zu einer Mischungsverteilung besteht.

9.3.2 Ausschließliche Überlagerung der Regressoren

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

$$\text{plim}\hat{\boldsymbol{\beta}}^a \quad = \quad (cov[\mathbf{u}] \,+\, \mathbf{Q})^{-1}\, \mathbf{Q}\, \boldsymbol{\beta} \tag{9-13}$$

wobei $cov[\mathbf{u}]$ durch (7-6) 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 $\boldsymbol{\beta}$.[30] 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:

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

Dabei ist $\hat{\boldsymbol{\beta}}^a$ durch (9-5) gegeben. Im Fall der Anonymisierung ist allerdings nur die Regressormatrix $\mathbf{X}^a$ bekannt, die Matrix $\mathbf{Q}$ also unbekannt. Da jedoch

$$\text{plim}\,\frac{1}{n}\mathbf{X^{a\prime}M}_\iota\mathbf{X^a} \quad = \quad cov[\mathbf{x}] \,+\, cov[\mathbf{u}] \quad = \quad \mathbf{Q} \,+\, cov[\mathbf{u}]$$

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

$$\hat{\boldsymbol{\beta}}^{a\,,\,korr} \quad = \quad \left(\frac{1}{n}\mathbf{X^{a\prime}M}_\iota\mathbf{X^a} \,-\, cov[\mathbf{u}]\right)^{-1} \left(\frac{1}{n}\mathbf{X^{a\prime}M}_\iota\mathbf{X^a}\right) \hat{\boldsymbol{\beta}}^a \quad . \tag{9-14}$$

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. Für den speziellen Fall erhalten wir aus Appendix A

$$cov[\mathbf{u}] \quad = \quad \sigma_\varepsilon^2\mathbf{I} \,+\, \mu^2\, \boldsymbol{\iota\iota}' \;=\; \tfrac{1}{\sigma_\varepsilon^2+\mu^2}\,((1-\rho)\,\mathbf{I} \,+\, \rho\,\boldsymbol{\iota\iota}')$$

wobei ρ durch

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

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

Dagegen gilt im Fall der "flexiblen" Spezifikation

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

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

[30] 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.).

9.3.3 Gemeinsame Überlagerung aller Variablen

Derselbe Wahrscheinlichkeitsgrenzwert ergibt sich, wenn auch die abhängige Variable (additiv) überlagert wird und zwischen den Überlagerungsvariablen keine Korrelation besteht.[31] Rosemann (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" Höhneverfahren diese Korrelation besteht, wie bereits oben demonstriert. In diesem Fall ergibt sich folgender Wahrscheinlichkeitsgrenzwert:

$$\text{plim}\hat{\boldsymbol{\beta}}^a \quad = \quad (cov[\mathbf{u}] + \mathbf{Q})^{-1} \, (\mathbf{Q}\,\boldsymbol{\beta} + cov[\mathbf{u}_x, u_y]) \tag{9-15}$$

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:

$$cov[\mathbf{u}_x, u_y] \quad = \quad \mu^2 \, \boldsymbol{\iota} \qquad . \tag{9-16}$$

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 (9-15) folgender Korrekturschätzer:

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

Dabei ist $\hat{\boldsymbol{\beta}}^a$ durch (9-5) gegeben. Entsprechend (9-14) erhalten wir in diesem Fall als operationalen Korrekturschätzer

$$\hat{\boldsymbol{\beta}}^{a\,,korr} \quad = \quad \left(\frac{1}{n} \mathbf{X^{a\prime}} \mathbf{M}_\iota \mathbf{X^a} - cov[\mathbf{u}] \right)^{-1} \left(\frac{1}{n} \mathbf{X^{a\prime}} \mathbf{M}_\iota \mathbf{X^a} \, \hat{\boldsymbol{\beta}}^a + cov[\mathbf{u}_x, u_y] \right) \quad , \tag{9-17}$$

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

$$\frac{1}{n} \mathbf{X^{a\prime}} \mathbf{M}_\iota \mathbf{X^a} \, \hat{\boldsymbol{\beta}}^a \quad = \quad \frac{1}{n} \mathbf{X^{a\prime}} \mathbf{M}_\iota \mathbf{y^a}$$

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 (2007 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.

9.4 Multiplikative Überlagerung im linearen Regressionsmodell

9.4.1 Allgemeine Bemerkungen

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

[31] Allerdings steigt die Varianz des Schätzers.

hinweg.[32] In diesem Fall ist die spezielle Formulierung der multiplikativen Überlagerung der Regressormatrix aus (9-8) besonders geeignet:

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

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

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

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:

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

Bei Rosemann (2006) wird im Einzelnen ausgeführt, wie die Wahrscheinlichkeitsgrenzwerte im Fall der multiplikativen Überlagerung abgeleitet werden. Allerdings ist zu beachten, daß im Gegensatz zu Rosemann (2006) **in dieser Arbeit ein Absolutglied (siehe (9-1)) explizit berücksichtigt** wird, wodurch sich die Ergebnisse deutlich verändern. Deshalb sollen die - modifizierten - Beweise im Folgende ebenfalls geliefert werden.[33]

9.4.2 Ausschließliche Überlagerung der Regressoren

Im Fall der ausschließlichen Überlagerung der Regressoren betrachten wir

$$\hat{\boldsymbol{\beta}}^a \quad = \quad \left(\mathbf{X}^{a\prime}\mathbf{M}_\iota\mathbf{X}^a\right)^{-1}\, \mathbf{X}^{a\prime}\mathbf{M}_\iota\mathbf{y}\;.$$

Die Matrix $\mathbf{X}^{a\prime}\mathbf{M}_\iota\mathbf{X}^a\,/\,n$ enthält die die empirischen **zentrierten zweiten** Momente der multiplikativ überlagerten Regressoren und tendiert deshalb gegen die Kovarianzmatrix einer bestimmten Zeile von $\mathbf{X}^a$ tendiert, die wir oben mit

$$\mathbf{x}^a\langle i\rangle \quad = \quad \mathbf{x}\langle i\rangle \odot \mathbf{u}\langle i\rangle \qquad ,\; i = 1, \ldots, n\,,$$

bezeichnet haben.

Wir verwenden nun die Formel (7-15), die in der hier verwendeten Symbolik (mit $\mathbf{x}\langle i\rangle \odot \mathbf{u}\langle i\rangle$ statt $\mathbf{Y}^a = \mathbf{Y} \odot \mathbf{U}$) zu folgendem Ergebnis führt:

$$cov[\mathbf{x}^a\langle i\rangle] \quad = \quad cov[\mathbf{u}\langle i\rangle] \;\odot\; (cov[\mathbf{x}\langle i\rangle] \,+\, E[\mathbf{x}\langle i\rangle]\, E[\mathbf{x}\langle i\rangle]') \qquad + \qquad cov[\mathbf{x}\langle i\rangle \qquad .$$

Wegen der identischen Verteilung über alle n Stichprobenpunkte kann man dann auch schreiben:

$$cov[\mathbf{x}^a] \quad = \quad cov[\mathbf{u}] \;\odot\; (cov[\mathbf{x}] \,+\, \boldsymbol{\mu}_x\boldsymbol{\mu}_x') \qquad + \qquad cov[\mathbf{x}] \qquad . \tag{9-18}$$

[32] Diese Annahme wird natürlich auch bei der additiven Überlagerung gemacht, wurde dort aber nicht speziell problematisiwert.

[33] Siehe Lin (1986) und Hwang(1989) für den Beweis im Fall nicht zentrierter Momente. Die vor allem bei Hwang (1989) verwendete Beweistechnik ist hier nicht anwendbar, weil der Operator $\mathbf{M}_\iota$ das arithmtische Mittel über alle n Stichprobenpunkte fordert. Siehe jedoch Appendix C, der unter expliziter Beachtung des Absolutglieds den Wahrscheinlichkeitsgrenzwert für den Gesamtvektor $(\alpha, \boldsymbol{\beta}')'$ bestimmt.

Unter Beachtung von (9-12) ergibt sich daraus

$$\operatorname{plim} \frac{1}{n} \mathbf{X^{a\prime} M_\iota X^a} \quad = \quad cov[\mathbf{u}] \, \odot \, (\mathbf{Q} + \boldsymbol{\mu}_x \, \boldsymbol{\mu}_x') \, + \, \mathbf{Q}$$

Wir benötigen ferner die Matrix

$$\frac{1}{n} \mathbf{X^{a}}' \mathbf{M}_\iota \mathbf{y} \quad = \quad \frac{1}{n} \mathbf{X^{a\prime} M}_\iota \, (\alpha \, \boldsymbol{\iota} \, + \, \mathbf{X} \, \boldsymbol{\beta} \, + \, \boldsymbol{\varepsilon}) \, = \, \frac{1}{n} \mathbf{X^{a\prime} M}_\iota \, (\mathbf{X} \, \boldsymbol{\beta} \, + \, \boldsymbol{\varepsilon})$$

Wegen

$$\operatorname{plim} \frac{1}{n} \mathbf{X^{a\prime} M}_\iota \mathbf{X} \, = \, \mathbf{Q} \quad \text{und} \quad \operatorname{plim} \frac{1}{n} \mathbf{X^{a\prime} M}_\iota \boldsymbol{\varepsilon} \, = \, \mathbf{0}$$

erhalten wir

$$\operatorname{plim} \frac{1}{n} \mathbf{X^{a\prime} M}_\iota \mathbf{y} \quad = \quad \operatorname{plim} \frac{1}{n} \mathbf{X^{a\prime} M}_\iota \, (\alpha \, \boldsymbol{\iota} \, + \, \mathbf{X} \, \boldsymbol{\beta} \, + \, \boldsymbol{\varepsilon}) \quad = \quad \mathbf{Q} \, \boldsymbol{\beta}$$

und damit

$$\operatorname{plim} \hat{\boldsymbol{\beta}}^a \quad = \quad \left(cov[\mathbf{u}] \, \odot \, (\mathbf{Q} + \boldsymbol{\mu}_x \, \boldsymbol{\mu}_x') \, + \, \mathbf{Q} \right)^{-1} \, \mathbf{Q} \, \boldsymbol{\beta} \, , \tag{9-19}$$

wobei $\mathbf{Q}$ durch (9-12) gegeben ist.

9.4.3 Gemeinsame Überlagerung aller Variablen

Im Fall der zusätzlichen Überlagerung auch der abhängigen Variablen lautet der Schätzer

$$\hat{\boldsymbol{\beta}}^a \quad = \quad \left(\mathbf{X}^{a\prime} \mathbf{M}_\iota \mathbf{X}^a \right)^{-1} \, \mathbf{X}^{a\prime} \mathbf{M}_\iota \mathbf{y}^a \, .$$

Somit ist der Wahrscheinlichkeits-Grenzwert des Vektors

$$\frac{1}{n} \mathbf{X^{a\prime} M}_\iota \mathbf{y^a}$$

zu bestimmen. Da dieser Vektor die empirischen Kovarianzen zwischen $x_{ik}^a \, = \, x_{ik} \cdot u_{ik}$ und $y_i^a \, = \, y_i \cdot u_{iy}$ enthält, tendieren die einzelnen Elemente dieses Vektors gegen die entsprechenden theoretischen Kovarianzen, die im folgenden zu bestimmen sind. Dabei verwenden wir die Modellannahme

$$y_i \quad = \quad \alpha \, + \, \sum_{\ell=1}^{K} \beta_\ell \, x_{i\ell} \, + \, \varepsilon_i \quad , \, i = 1, \ldots, n \, ,$$

die äquivalent zu (9-1) ist.

Wir verwenden wieder die Formel (7-15) für die Kovarianz, die wir hier - unter Vernachlässigung des Beobachtungsindex i - wie folgt schreiben:

$$cov \left[\begin{pmatrix} x_k^a \\ y^a \end{pmatrix} \right] \quad = \quad E_{xy} \left[cov \left[\begin{pmatrix} x_k^a \\ y^a \end{pmatrix} \quad \middle| \quad \begin{pmatrix} x_k \\ y \end{pmatrix} \right] + cov_{xy} \left[E \left[\begin{pmatrix} x_k^a \\ y^a \end{pmatrix} \quad \middle| \quad \begin{pmatrix} x_k \\ y \end{pmatrix} \right] \right].$$

Für die bedingte Kovarianz ergibt sich

$$cov \left[\begin{pmatrix} x_k^a \\ y^a \end{pmatrix} \quad \middle| \quad \begin{pmatrix} x_k \\ y \end{pmatrix} \right] \quad = \quad x_k \cdot y \cdot cov[u_k, u_y] \quad = \quad x_k \, (\alpha + \sum_{\ell=1}^{K} \beta_\ell \, x_\ell + \varepsilon) \, cov[u_k, u_y]$$

und für den Vektor der bedingten Erwartungswerte erhalten wir

$$E\left[\begin{pmatrix} x_k^a \\ y^a \end{pmatrix} \;\middle|\; \begin{pmatrix} x_k \\ y \end{pmatrix}\right] \;=\; \begin{pmatrix} x_k \\ y \end{pmatrix}$$

Der Erwartungswert der bedingten Kovarianzmatrix ergibt

$$E[x_k\,(\alpha \,+\, \sum_{\ell=1}^{K} \beta_\ell\, x_\ell \,+\, \varepsilon)\, cov[u_k, u_y]] \;=\; (\alpha\,\mu_k \,+\, \sum_{\ell=1}^{K} \beta_\ell\,(\sigma_{k\ell} \,+\, \mu_k\,\mu_\ell))\; cov[u_k, u_y]$$

und für die Kovarianz des Vektors der bedingten Erwartungswerte erhalten wir

$$cov\,[x_k\,,\, y] \;=\; cov[x_k, \alpha] \,+\, cov[x_k, \sum_{\ell=1}^{K} \beta_\ell\, x_\ell] \,+\, cov[x_k, \varepsilon] \;=\; \sum_{\ell=1}^{K} \beta_\ell\, \sigma_{k\ell}$$

Wenn wir die ergebnisse für alle K Elemente gemeinsam schreiben, so erhalten wir für den Vektor der Ergebnisse aus dem ersten Ausdruck

$$\left\{\alpha\,\boldsymbol{\mu}_x \;+\; (\mathbf{Q} + \boldsymbol{\mu}_x\boldsymbol{\mu}_x')\boldsymbol{\beta}\right\} \odot cov[\mathbf{u}_x, u_y]$$

und der Vektor, der aus dem zweiten Ergebnis folgt, lautet

$$\mathbf{Q}\,\boldsymbol{\beta}$$

Damit erhalten wir folgenden Wahrscheinlichkeitsgrenzwert für den "naiven" Schätzer im Fall der gemeinsamen Überlagerung aller Variablen:[34]

$$\text{plim}\,\hat{\boldsymbol{\beta}}^a \;=\; \left(cov[\mathbf{u_x}] \odot \left\{\mathbf{Q} \,+\, \boldsymbol{\mu}_x\,\boldsymbol{\mu}_x'\right\} \,+\, \mathbf{Q}\right)^{-1} \left(\mathbf{Q}\,\boldsymbol{\beta} \,+\, \left\{\alpha\,\boldsymbol{\mu}_x \,+\, (\mathbf{Q} + \boldsymbol{\mu}_x\boldsymbol{\mu}_x')\boldsymbol{\beta}\right\} \odot cov[\mathbf{u}_x, u_y]\right) \tag{9-20}$$

wobei $cov[\mathbf{u_x}]$ die $(n\times n)$-Kovarianzmatrix der Überlagerungsvariablen aus $\mathbf{u_x}$ und $cov[\mathbf{u}_x, u_y]$ den n-dimensionalen Vektor der Kovarianzen zwischen den Elementen aus $\mathbf{u_x}$ und u_y bezeichnet. Konsistenz ist demnach nur dann gegeben, wenn beide gerade erwähnten Ausdrücke gleich der Nullmatrix bzw. gleich dem Nullvektor sind. Insbesondere muß die Kovarianzmatrix $cov[\mathbf{u_x}]$ eine Nullmatrix sein, d.h. auch die Diagonalelemente (= Varianzen) müssen gleich Null sein!

Im Fall der Höhneüberlagerung gilt

$$cov[\mathbf{u}_x] \;=\; \sigma_\varepsilon^2\mathbf{I} \,+\, \delta^2\,\boldsymbol{\iota}\boldsymbol{\iota}' \tag{9-21}$$

und

$$cov[\mathbf{u}_x, u_y] \;=\; \begin{cases} \delta^2\,\boldsymbol{\iota} & \text{für die spezielle Spezifikation} \\ \mathbf{0} & \text{für die flexible Spezifikation} \end{cases} \tag{9-22}$$

9.4.4 Korrekturschätzer

Vorausgesetzt daß $\mathbf{Q}$ und $cov[\mathbf{u}]$ bekannt sind oder konsistent geschätzt werden können, läßt sich aus dem Ergebnis (9-19) für den Fall der ausschließlichen (multiplikativer) Überlagerung der Regressoren ein konsistenter "Korrektur-Schätzer" wie folgt definieren:

$$\hat{\boldsymbol{\beta}}^{a\,,\,korr} \;=\; \mathbf{Q}^{-1}\,\left(cov[\mathbf{u}] \,\odot\, (\mathbf{Q} \,+\, \boldsymbol{\mu}_x\,\boldsymbol{\mu}_x') \,+\, \mathbf{Q}\right)\,\hat{\boldsymbol{\beta}}^a \quad .$$

[34] Siehe auch den alternativen Beweis in Appendix C.

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

$$\text{plim}\,\frac{1}{n}\mathbf{X^{a\prime}M}_\iota\mathbf{X^a} \quad = \quad cov[\mathbf{u}] \,\odot\, \mathbf{Q}$$

(siehe Hwang 1986) und somit[35]

$$\text{plim}\,\left(\frac{1}{n}\mathbf{X^{a\prime}M}_\iota\mathbf{X^a} \,\div\, cov[\mathbf{u}]\right) \quad = \quad \mathbf{Q}$$

und außerdem

$$\text{plim}\,\frac{1}{n}\mathbf{X^{a\prime}M}_\iota\mathbf{y} \quad = \quad \mathbf{Q}\,\boldsymbol{\beta}$$

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

$$\hat{\boldsymbol{\beta}}^{a\,,\,korr} \quad = \quad \left(\mathbf{X^{a\prime}M}_\iota\mathbf{X^a} \,\div\, cov[\mathbf{u}]\right)^{-1}\,\mathbf{X^{a\prime}M}_\iota\mathbf{y} \quad , \tag{9-23}$$

was der Formel (16.105) bei Rosemann (2006) entspricht.

Dagegen ist die Ableitung eines Korrekturschätzers **im Fall der gemeinsamen (multiplikativen) Überlagerung** komplizierter, weil eine direkte Auflösung nach $\boldsymbol{\beta}$ nicht möglich ist.

10 Ergänzende Überlegungen für Paneldaten

10.1 Allgemeines

10.1.1 Überlagerungsstrategien

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

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

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

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

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

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

interpretieren.

[35] Das Symbol $\div$ bezeichnet die "Hadamard-Division".

Es stellt sich die Frage, in welcher Weise die Überlagerung nun vorgenommen werden soll. Wir beschränken uns hier auf die Spezifikation[36]

$$U_{jt} \quad = \quad 1 \; + \; \delta \, D \; + \; \varepsilon_{jt} \quad \text{(”alle Merkmale und Zeitpunkte gemeinsam”)} \quad , \qquad \text{(10-1)}$$

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

10.1.2 Das einfache lineare Panelmodell

Entsprechend dem Vorgehen von Biørn (1996) (und abweichend von der Darstellung im Abschnitt 9.4 für Querschnittsdaten betrachten wir hier das **einfache** Regressionsmodell

$$y_{it} \quad = \quad \alpha \; + \; \beta \, x_{it} \; + \; \eta_{it} \quad i = 1, \ldots, n \; t = 1, \ldots T \qquad \text{(10-2)}$$

mit einem einzigen Regressor.

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 10.4.5, 10.4.6, 10.4.7 und 10.4.8. 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 10.2) darstellen. Dabei werden wir uns wieder auf die multiplikative Überlagerung und dort auf die ”spezielle” Höhne-Variante konzentrieren. Ferner werden wir in Abschnitt 10.3 das einfache lineare Panelmodell mit Individualeffekten einführen.

10.2 Überlagerung von Paneldaten

10.2.1 Symbolik

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

$$y_{itj} \quad i = 1, \ldots, n, \; t = 1, \ldots, T, \; j = 1, \ldots, r, \qquad \text{(10-3)}$$

bezeichnet das Merkmal[37] y_j für Beobachtungseinheit i im Zeitpunkt t, wobei insgesamt r Merkmale betrachtet werden.

[36] Denkbar sind folgende Varianten:

U_{jt}	$=$	$1 + \delta_{jt} \, D_{jt} \; + \; \varepsilon_{jt}$	alle Merkmale und Zeitpunkte getrennt
U_{jt}	$=$	$1 + \delta_j \, D_j \; + \; \varepsilon_{jt}$	alle Merkmale getrennt
U_{jt}	$=$	$1 + \delta_t \, D_t \; + \; \varepsilon_{jt}$	alle Zeitpunkte getrennt
U_{jt}	$=$	$1 + \delta \, D \; + \; \varepsilon_{jt}$	alle Merkmale und Zeitpunkte gemeinsam

Weitere Varianten, in denen δ 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.

[37] Das Symbol y steht hier für jedes beliebige Merkmal und nicht für die abhängige Variable. Dies entspricht dem Vorgehen im Abschnitt 7.

10.2.2 Additive Überlagerung

Wenn wir die additive Entsprechung zu (10-1) für die Überlagerung verwenden, d.h.

$$u_{itj} \quad = \quad y_{itj} \; + \; \mu D_i \; + \; \varepsilon_{itj} \quad , \tag{10-4}$$

so ergibt sich für die anonymisierten Merkmale

$$y^a_{itj} \quad = \quad y_{itj} \; + \; u_{itj} \quad = \quad y_{itj} \; + \; \mu D_i \; + \; \varepsilon_{itj} \tag{10-5}$$

Dabei soll ε_{itj} konstante Varianz besitzten, ferner sollen alle $(itj)-$Kombinationen unkorreliert sein:

$$\begin{array}{rcll} E[\varepsilon_{itj}] & = & 0 & \left\{ \begin{array}{l} i = 1, \ldots, n, \\ t = 1, \ldots, T, \\ j = 1, \ldots, r. \end{array} \right. \\ var[\varepsilon_{itj}] & = & \sigma^2_\varepsilon & \\ cov[\varepsilon_{itj}, \varepsilon_{hsk}] & = & 0 \; , \; i \neq h, t \neq s, j \neq k & \end{array} \tag{10-6}$$

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 7 (Anonymisierung mittels stochastischer Überlagerung) für Querschnittsdaten läßt sich zeigen, daß

$$var[y^a_{itj}] \quad = \quad \sigma^2_{tj} \; + \; \mu^2 \; + \; \sigma^2_\varepsilon \tag{10-7}$$

und

$$cov[y^a_{itj}, y^a_{hsk}] \quad = \quad \left\{ \begin{array}{ccc} \sigma_{jk} \; + \; \mu^2 & \text{falls} & i = h, \, t = s, \, j \neq k \\ \sigma_{ts} \; + \; \mu^2 & \text{falls} & i = h, \, t \neq s, \, j = k \\ 0 & \text{falls} & i \neq h \end{array} \right. \tag{10-8}$$

für alle $(j, k, t, s)-$Kombinationen gilt. Dies entspricht (7-6) im Abschnitt 7 (Anonymisierung mittels stochastischer Überlagerung) für Querschnittsdaten. In (10-7) wird jetzt zugelassen, daß die Varianz für Merkmal j über die Zeit hin variiert und in (10-8) bezeichnet σ_{jk} die (kontemporäre) Kovarianz zwischen den Merkmalen j und k, während σ_{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

$$corr[y^a_{itj}, y^a_{hsk}] \quad = \quad \left\{ \begin{array}{ccc} \frac{\sigma_{jk} + \mu^2}{\sqrt{(\sigma^2_{tj} + \mu^2 + \sigma^2_\varepsilon)(\sigma^2_{tk} + \mu^2 + \sigma^2_\varepsilon)}} & \text{falls} & i = h, \, t = s, \, j \neq k \\ \frac{\sigma_{ts} + \mu^2}{\sqrt{(\sigma^2_{tj} + \mu^2 + \sigma^2_\varepsilon)(\sigma^2_{sj} + \mu^2 + \sigma^2_\varepsilon)}} & \text{falls} & i = h, \, t \neq s, \, j = k \\ 0 & \text{falls} & i \neq h \end{array} \right. \tag{10-9}$$

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.

10.2.3 Multiplikative Überlagerung

Im multiplikativen Fall verwenden wir statt (10-5) die Spezifikation

$$y^a_{itj} \quad = \quad y_{itj}\,(1 \,+\, \delta D_i \,+\, \varepsilon_{itj}) \tag{10-10}$$

Dabei soll ε_{itj} wieder (10-6) 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 7 (Anonymisierung mittels stochastischer Überlagerung) für Querschnittsdaten ableiten. Dabei beschränken wir uns auf die Überlagerung mithilfe der Mischungsverteilung, wie sie Abschnitt 7.3.4 (spezielle Höhne-Spezifikation) betrachtet wurde.

Es läßt sich zeigen, daß

$$var[y^a_{itj}] \quad = \quad \sigma^2_{tj} \,+\, (\delta^2 \,+\, \sigma^2_\varepsilon)\,(\sigma^2_{tj} \,+\, \mu^2_{tj}) \tag{10-11}$$

und

$$cov[y^a_{itj}, y^a_{hsk}] \quad = \quad \begin{cases} \sigma_{jk} \,+\, \delta^2\,(\sigma_{jk} + \mu_{tj}\mu_{tk}) & \text{falls} \quad i = h,\, t = s,\, j \neq k \\ \sigma_{ts} \,+\, \delta^2\,(\sigma_{ts} + \mu_{tj}\mu_{sj}) & \text{falls} \quad i = h,\, t \neq s,\, j = k \\ 0 & \text{falls} \quad i \neq h \end{cases} \tag{10-12}$$

gilt. Dabei bezeichnet μ_{tj} den Erwartungswert für Merkmal j im Zeitpunkt t. Varianzen und Kovarianzen entsprechen exakt der Struktur, wie sie in (7-17) 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:

$$corr[y^a_{itj}, y^a_{hsk}] \quad = \quad \begin{cases} \dfrac{\sigma_{jk} + \delta^2\,(\sigma_{jk}+\mu_{tj}\mu_{tk})}{\sqrt{\left\{\sigma^2_{tj} + (\delta^2+\sigma^2_\varepsilon)\,(\sigma^2_{tj}+\mu^2_{tj})\right\}\left\{\sigma^2_{tj} + (\delta^2+\sigma^2_\varepsilon)\,(\sigma^2_{tj}+\mu^2_{tj})\right\}}} & \text{falls} \quad i = h,\, t = s,\, j \neq k \\ \dfrac{\sigma_{ts} + \delta^2\,(\sigma_{ts}+\mu_{tj}\mu_{sj})}{\sqrt{\left\{\sigma^2_{tj} + (\delta^2+\sigma^2_\varepsilon)\,(\sigma^2_{tj}+\mu^2_{tj})\right\}\left\{\sigma^2_{tj} + (\delta^2+\sigma^2_\varepsilon)\,(\sigma^2_{tj}+\mu^2_{tj})\right\}}} & \text{falls} \quad i = h,\, t \neq s,\, j = k \\ 0 & \text{falls} \quad i \neq h \end{cases} \tag{10-13}$$

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.

10.3 Das einfache lineare Panelmodell mit Individualeffekten

10.3.1 Feste Individualeffekten

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

$$y_{it} \quad = \quad \alpha \,+\, \beta\, x_{it} \,+\, \sum_{h=1}^{n} \tau_h\, W_{ht} \,+\, \eta_{it} \quad , \; i = 1, \ldots, n \;,\; t = 1, \ldots T \tag{10-14}$$

wobei

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

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

$$\alpha \quad = \quad 0 \qquad .$$

In Matrixschreibweise erhalten wir dann

$$\mathbf{y} \quad = \quad \mathbf{x}\beta \; + \; \mathbf{W}\,\boldsymbol{\tau} \; + \; \boldsymbol{\eta} \tag{10-15}$$

wobei $\mathbf{x}$ ein Vektor und β ein Skalar ist. Die Matrix $\mathbf{W}$ enthält die individuenspezifischen Dummyvariablen.[38] Als sogenannter "Within-Schätzer" (oder auch "fixed effects estimator") ergibt sich die Teilschätzung nach der Kleinstquadrate-Methode:

$$\hat{\beta}_W \quad = \quad \frac{\mathbf{x}'\left(\mathbf{I} - \mathbf{W}(\mathbf{W}'\mathbf{W})^{-1}\mathbf{W}\right)\mathbf{y}}{\mathbf{x}'\left(\mathbf{I} - \mathbf{W}(\mathbf{W}'\mathbf{W})^{-1}\mathbf{W}\right)\mathbf{x}} \quad = \quad \frac{\sum_{t=1}^{T}\sum_{i=1}^{n}(x_{it} - \overline{x}_{i\bullet})(y_{it} - \overline{y}_{i\bullet})}{\sum_{t=1}^{T}\sum_{i=1}^{n}(x_{it} - \overline{x}_{i\bullet})^2} \tag{10-16}$$

mit

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

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

10.3.2 Stochastische Individualeffekte

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

$$y_{it} \quad = \quad \alpha \; + \; \beta\, x_{it} \; + \; \tau_i \; + \; \eta_{it} \quad , \, i = 1, \ldots, n \; , \; t = 1, \ldots T \tag{10-17}$$

wobei die τ_i nun Zufallsvariable sind. Weil daraus

$$\overline{y}_{i\bullet} \; = \; \alpha \; + \; \overline{x}_{i\bullet}\,\beta \; + \; \tau_i \; + \; \overline{\eta}_{i\bullet}$$

und

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

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

$$\hat{\beta}_W \quad = \quad \frac{\sum_{t=1}^{T}\sum_{i=1}^{n}(x_{it} - \overline{x}_{i\bullet})\left\{(x_{it} - \overline{x}_{i\bullet})\beta \; + \; (\eta_{it} \; - \; \overline{\eta}_{i\bullet})\right\}}{\sum_{t=1}^{T}\sum_{i=1}^{n}(x_{it} - \overline{x}_{i\bullet})^2} \tag{10-18}$$

d.h. $\hat{\beta}_W$ ist ein konsistenter Schätzer.[39]

Im Appendix Dwird gezeigt, daß die Varianz dieses Schätzers durch

$$var[\hat{\boldsymbol{\beta}}_W] \quad = \quad \frac{\sigma_\eta^2}{(\mathbf{x}'\,\mathbf{M}_W\,\mathbf{x})} \quad = \quad \frac{\sigma_\eta^2}{\sum_{t=1}^{T}\sum_{i=1}^{n}(x_{it} - \overline{x}_{i\bullet})^2}. \tag{10-19}$$

gegeben ist. Man beachte, daß die Varianz der Individualeffekte in dieser Formel nicht auftaucht.

[38] Für eine exakte Beschreibung siehe die Ergebnisse für das multiple Regressionsmodell in Appendix D.

[39] 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 τ_i und dem Regressor" dann problematisch ist, wenn man den Effekt von zeitlich nur wenig variierenden Regressoren schätzen will.

10.4 Schätzung des linearen Panelmodells aus anonymisierten Daten

10.4.1 Der naive Panelschätzer

Im Fall von anonymisierten Daten verwenden wir statt (10-16) den folgenden "naiven" Schätzer (Within-Schätzer, FE-Schätzer):

$$\hat{\beta}^a_W \quad = \quad \frac{\mathbf{x}^{a\prime}\left(\mathbf{I}-\mathbf{W}(\mathbf{W}'\mathbf{W})^{-1}\mathbf{W}\right)\mathbf{y^a}}{\mathbf{x}^{a\prime}\left(\mathbf{I}-\mathbf{W}(\mathbf{W}'\mathbf{W})^{-1}\mathbf{W}\right)\mathbf{x}^a} \quad = \quad \frac{\sum_{t=1}^{T}\sum_{i=1}^{n}(x^a_{it}-\overline{x}^a{}_{i\bullet})(y^a_{it}-\overline{y^a}_{i\bullet})}{\sum_{t=1}^{T}\sum_{i=1}^{n}(x^a_{it}-\overline{x}^a{}_{i\bullet})^2} \tag{10-20}$$

mit

$$\overline{x}^a{}_{i\bullet} \;=\; \frac{1}{T}\sum_{t=1}^{T} x^a_{it} \quad , \quad \overline{y^a}_{i\bullet} \;=\; \frac{1}{T}\sum_{t=1}^{T} y^a_{it} \qquad .$$

10.4.2 Additive Meßfehler allgemein

Biørn (1996 Kapitel 10.2.2) betrachtet ein lineares Panelmodel, in dem die Regressorvariable x latent ist und nur die meßfehlerbehaftete Variable x^a betrachtet werden kann[40], und leitet den Wahrscheinlichkeitsgrenzwert des naiven Panelschätzers (10-20) bei additiver Überlagerung ab. Das Fehlermodell lautet[41]

$$x^a_{it} \quad = \quad x_{it} \;+\; u_{it} \qquad .$$

Dann ergibt sich für den Wahrscheinlichkeitsgrenzwert[42]

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

[40] 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.

[41] Aus Gründen der Einheitlichkeit bleibe ich auch hier bei der Symbolik des Anonymisierungsfalls, verwende also x statt x^* und x^a statt x.

[42] Ich betrachte ausschließlich den Grenzwert für $n \to \infty$. Deshalb schreibe ich im folgenden oftmals nur plim statt $\text{plim}_{n\to\infty}$.

Da x, u und η gemäß Annahme miteinander unkorreliert sind, ergibt sich[43]

$$\begin{aligned} \text{plim}_{n\to\infty}\,\hat{\beta}^a &= \frac{\text{plim}\,\frac{1}{nT}\,\sum_{t=1}^{T}\sum_{i=1}^{n}(x_{it}-\overline{x}_{i\bullet})^2\,\beta}{\text{plim}\,\frac{1}{nT}\,\sum_{t=1}^{T}\sum_{i=1}^{n}(x_{it}-\overline{x}_{i\bullet})^2\;+\;\text{plim}\,\frac{1}{nT}\,\sum_{t=1}^{T}\sum_{i=1}^{n}(u_{it}-\overline{u}_{i\bullet})^2} \\ &\quad \frac{\sigma_x^2}{\sigma_x^2\;+\;\left(1\;-\;\frac{1}{T}\right)\,\sigma_u^2}\,\beta \end{aligned} \tag{10-21}$$

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!

10.4.3 Einschub: Korrekter Wahrscheinlichkeitsgrenzwert im Panelfall ?

Das Ergebnis (10-21) bedarf allerdings eines Kommentars: Bei Biørn (1996 Formel (10-59), Symbolik leicht verändert) wird der Wahrscheinlichkeitsgrenzwert von $\hat{\beta}^a$ wie folgt geschrieben:

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

mit

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

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 43 würde sich aber

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

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.

10.4.4 Meßfehler mit Faktorstruktur

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

[43] Dabei wird ausgenutzt, daß

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

und

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

gilt. Damit erhalten wir

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

unterstellt:

$$x^a_{it} \quad = \quad x_{it} \; + \; \tau_i \; + \; \phi_t \; + \; \varepsilon_{it} \tag{10-22}$$

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

$$\text{plim}_{n\to\infty}\, \hat{\beta}^a \quad = \quad \frac{\sigma_x^2}{\sigma_x^2 + \sigma_\phi^2 + (1 - \frac{1}{T})\sigma_\varepsilon^2} \; \beta \tag{10-23}$$

Siehe Biørn (1996), Formel (10-115), der allerdings auch hier wieder keinen expliziten Grenzwert bezüglich x angibt. Siehe meine Bemerkungen in Unterabschnitt 10.4.3.

Bemerkenswert ist, daß durch die Bildung von Abweichungen vom individuenspezifischen Mittelwert der Effekt τ_i bei der Schätzung keine Rolle spielt, im Gegensatz zum zeitspezifischen Effekt ϕ_t. Man beachte auch den engen Zusammenhang mit der additiven Überlagerung a la Höhne in (10-5), wo der zeitspezifische Effekt entfällt. Wir werden später sehen, daß deshalb bei der additiven Überlagerung der Bias des naiven Panelschätzers nur von σ_ε^2, nicht aber vom Parameter μ abhängt. Siehe dazu Abschnitt 10.4.5. Entsprechende Ergebnisse für den multiplikativen Fall werden in den Abschnitten 10.4.6 sowie 10.4.7 besprochen. Dort beeinflußt allerdings auch der Zuschlagsparameter (δ) den Bias der Schätzung.

10.4.5 Additive Überlagerung a la Höhne

Bei der additiven Überlagerung der Regressoren a la Höhne verwenden wir im Panelfall die Spezifikation (10-5), d.h.

$$x^a_{it} \quad = \quad x_{it} \; + \; \mu D_i \; + \; \varepsilon_{it}$$

und erhalten wegen

$$\overline{x}^a_{i\bullet} \quad = \quad \overline{x}_{i\bullet} \; + \; \mu D_i \; + \; \overline{\varepsilon}_{i\bullet}$$

und

$$x^a_{it} \; - \; \overline{x}^a_{i\bullet} \quad = \quad x_{it} \; - \; \overline{x}_{i\bullet} \; + \; \varepsilon_{it} \; - \; \overline{\varepsilon}_{i\bullet}$$

als Wahrscheinlichkeitsgrenzwert des "naiven" Panelschätzers

$$\text{plim}_{n\to\infty}\, \hat{\beta}^a \quad = \quad \frac{\sigma_x^2}{\sigma_x^2 + \left(1 \; - \; \frac{1}{T}\right) \; \sigma_\varepsilon^2} \; \beta \qquad , \tag{10-24}$$

d.h. der Bias hängt nur von der Varianz des Restterms ε, nicht aber von der Größenordnung des Zuschlag-Parameters μ ab! Da der (quadrierte) Zuschlagsparameter deutlich größer ist als die Varianz σ_ε^2, verringert sich der Bias gegenüber der Schätzung im Querschnittsfall (siehe Abschnitt 9.3) 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 (7-5) gilt

$$var[x^a] \quad = \quad \sigma_x^2 \; + \; \mu^2 \; + \; \sigma_\varepsilon^2 \quad .$$

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

$$\hat{\beta}^{a,korr} \quad = \quad \frac{\widehat{var[x^a]} \; - \; \mu^2 \; - \; \frac{1}{T}\sigma_\varepsilon^2}{\widehat{var[x^a]} \; - \; \mu^2 \; - \; \sigma_\varepsilon^2} \; \hat{\beta}^a_W \tag{10-25}$$

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

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

$$x^a_{it} \; - \; \overline{x}^a_{i\bullet} \quad = \quad x_{it} \; - \; \overline{x}_{i\bullet} \; + \; \varepsilon_{itx} \; - \; \overline{\varepsilon}_{i\bullet x}$$

und

$$y^a_{it} \; - \; \overline{y^a}_{i\bullet} \quad = \quad y_{it} \; - \; \overline{y}_{i\bullet} \; + \; \varepsilon_{ity} \; - \; \overline{\varepsilon}_{i\bullet y}$$

wobei die beiden Überlagerungsvariablen nun angeben, ob sie zu x oder y gehören. Wie in Abschnitt 9.3 ausführlich dargestellt, gilt gemäß Annahme (!!)

$$cov[\varepsilon_x, \varepsilon_y] \quad = \quad 0 \quad .$$

Demnach erhalten wir für den naiven Panelschätzer[44]

$$\text{plim}_{n\to\infty} \, \hat{\beta}^a \quad = \quad \frac{\sigma_x^2 \, \beta \; + \; \left(1 - \frac{1}{T}\right) \, cov[\varepsilon_x, \varepsilon_y]}{\sigma_x^2 + \left(1 - \frac{1}{T}\right) \, \sigma_\varepsilon^2} \quad = \quad \frac{\sigma_x^2 \, \beta}{\sigma_x^2 + \left(1 - \frac{1}{T}\right) \, \sigma_\varepsilon^2} \qquad , \quad \text{(10-26)}$$

Dies entspricht der Formel (10-24) 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, sofern die beiden Fehler - infolge Verwendung der Mischungsverteilung - miteinander korreliert sind! Siehe Abschnitt 9.3.

10.4.6 Multiplikative Überlagerung des Regressors

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

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

die direkt aus (10-1) folgt. Dann ergibt sich für die Berechnung der Abweichung vom Mittelwert

$$\overline{x}^a_{i\bullet} \quad = \quad (1 \; + \; \delta D_i) \, \overline{x}_{i\bullet} \; + \; \overline{x\varepsilon}_{i\bullet}$$

und

$$x^a_{it} \; - \; \overline{x}^a_{i\bullet} \quad = \quad (1 \; + \; \delta D_i) \, (x_{it} \; - \; \overline{x}_{i\bullet}) \; + \; x_{it} \, \varepsilon_{it} \; - \; \overline{x\varepsilon}_{i\bullet} \qquad \text{(10-27)}$$

[44] Hier wird die folgende Zerlegung verwendet: Für beliebige Zufallsvariablen λ_{itj} mit $cov[\lambda_{itj}, \lambda_{itk}] = \kappa$, $j \neq k$ gilt

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

und

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

gilt. Damit erhalten wir

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

Allerdings ist im obigen Fall die Kovarianz gleich Null!

wobei

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

Im Gegensatz zur additiven Überlagerung verschwindet also der Zuschlagsparameter (δ) 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}^a_W &= \frac{\sum_{t=1}^T\sum_{i=1}^n (x^a_{it} - \overline{x}^a{}_{i\bullet})(y_{it} - \overline{y}_{i\bullet})}{\sum_{t=1}^T\sum_{i=1}^n (x^a_{it} - \overline{x}^a{}_{i\bullet})^2} \\
&= \frac{\sum_{t=1}^T\sum_{i=1}^n ((1\ +\ \delta D_i)\,(x_{it}\ -\ \overline{x}_{i\bullet})\ +\ x_{it}\,\varepsilon_{it}\ -\ \overline{x\varepsilon}_{i\bullet})(y_{it} - \overline{y}_{i\bullet})}{\sum_{t=1}^T\sum_{i=1}^n \left((1\ +\ \delta D_i)\,(x_{it}\ -\ \overline{x}_{i\bullet})\ +\ x_{it}\,\varepsilon_{it}\ -\ \overline{x\varepsilon}_{i\bullet}\right)^2} \\
&= \frac{\sum_{t=1}^T\sum_{i=1}^n ((1\ +\ \delta D_i)\,(x_{it}\ -\ \overline{x}_{i\bullet})\ +\ x_{it}\,\varepsilon_{it}\ -\ \overline{x\varepsilon}_{i\bullet})((x_{it}\ -\ \overline{x}_{i\bullet})\,\beta\ +\eta_{it}\ -\ \overline{\eta}_{i\bullet})}{\sum_{t=1}^T\sum_{i=1}^n \left((1\ +\ \delta D_i)\,(x_{it}\ -\ \overline{x}_{i\bullet})\ +\ x_{it}\,\varepsilon_{it}\ -\ \overline{x\varepsilon}_{i\bullet}\right)^2} \\
&= \frac{\sum_{t=1}^T\sum_{i=1}^n \left\{(1\ +\ \delta D_i)\,(x_{it}\ -\ \overline{x}_{i\bullet})^2\ \beta\ +\ (1\ +\ \delta D_i)\,(x_{it}\ -\ \overline{x}_{i\bullet})\,(\eta_{it}\ -\ \overline{\eta}_{i\bullet})\right\}}{\sum_{t=1}^T\sum_{i=1}^n \left\{(1+\delta D_i)^2\,(x_{it}-\overline{x}_{i\bullet})^2\ +\ 2(1+\delta D_i)\,(x_{it}-\overline{x}_{i\bullet})\,(x_{it}\varepsilon_{it}-\overline{x\varepsilon}_{i\bullet})\ +\ (x_{it}\varepsilon_{it}\ -\ \overline{x\varepsilon}_{i\bullet})^2\right\}} \\
&\quad + \frac{\sum_{t=1}^T\sum_{i=1}^n \{(x_{it}\,\varepsilon_{it}\ -\ \overline{x\varepsilon}_{i\bullet})(x_{it}\ -\ \overline{x}_{i\bullet})\,\beta\ +(x_{it}\,\varepsilon_{it}\ -\ \overline{x\varepsilon}_{i\bullet})(\eta_{it}\ -\ \overline{\eta}_{i\bullet})\}}{\sum_{t=1}^T\sum_{i=1}^n \left\{(1+\delta D_i)^2\,(x_{it}-\overline{x}_{i\bullet})^2\ +\ 2(1+\delta D_i)\,(x_{it}-\overline{x}_{i\bullet})\,(x_{it}\varepsilon_{it}-\overline{x\varepsilon}_{i\bullet})\ +\ (x_{it}\varepsilon_{it}\ -\ \overline{x\varepsilon}_{i\bullet})^2\right\}}
\end{aligned} \tag{10-28}$$

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

$$\text{plim}(\hat{\beta}^a_W) \ = \ \frac{\frac{1}{T}\sum_{t=1}^T\{A_t\ +\ B_t\}}{\frac{1}{T}\sum_{t=1}^T\{\Phi_t\ +\ \Pi_t\ +\ \Psi_t\}} \ +\ \frac{\frac{1}{T}\sum_{t=1}^T\{\Gamma_t\ +\ \Delta_t\}}{\frac{1}{T}\sum_{t=1}^T\{\Phi_t\ +\ \Pi_t\ +\ \Psi_t\}} \tag{10-29}$$

wobei

$$\begin{aligned}
A_t &= E[(1+\delta D)(X_t - E[X_t])^2\beta] && \text{(10-30)}\\
B_t &= E[(1+\delta D)\,(X_t - E[X_t])\,(\eta_t - E[\eta_t)] && \text{(10-31)}\\
\Gamma_t &= E[(X_t\varepsilon_t - E[X_t\varepsilon_t])(X_t - E[X_t])\beta] && \text{(10-32)}\\
\Delta_t &= E[(X_t\varepsilon_t - E[X_t\varepsilon_t])\,(\eta_t - E[\eta_t)] && \text{(10-33)}\\
\Phi_t &= E[(1+\delta D)^2\,(X_t - E[X_t])^2] && \text{(10-34)}\\
\Pi_t &= 2\,E[(1+\delta D)\,(X_t - E[X_t])\,(X_t\varepsilon_t - E[X_t\varepsilon_t])] && \text{(10-35)}\\
\Psi_t &= E[(X_t\varepsilon_t - E[X_t\varepsilon_t])^2] && \text{(10-36)}
\end{aligned}$$

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 individuenspezifischen Mittelwerte enthalten. Siehe dazu meine Bemerkung in Unterabschnitt 10.4.3.

Für die einzelnen Ausdrücke ergibt sich folgendes: Wegen der Unabhängigkeit von D und X_t erhalten wir für (10-30)

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

wobei wir $E[D] = 0$ verwendet haben (symmetrischer Fall). Da X_t und η_t unabhängig sind, gilt für (10-31)

$$B_t \;=\; 0 \quad .$$

In (10-32) 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

$$\Gamma_t \;=\; 0 \qquad .$$

Für (10-33) ergibt sich

$$\Delta_t \;= 0 \quad ,$$

weil η_t stochastisch unabhängig von X_t und ε_t ist. In (10-34) 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 (10-35) erhalten wir

$$\Pi_t \;=\; 0 \quad .$$

Vergleiche dazu die Berechnung zu (10-31) oben. Außerdem kann wieder die Unabhängigkeit zwischen D und X_t ausgenutzt werden. In (10-36) 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" Höhne-Variante folgender Ausdruck:

$$\text{plim}(\hat{\beta}_W^a) \;=\; \frac{\sigma_x^2\,\beta}{(1+\delta^2)\,\sigma_x^2 \;+\; \sigma_\varepsilon^2\,(\sigma_x^2 \,+\, \mu_x^2)} \qquad . \tag{10-37}$$

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

$$\delta \;=\; 0 \qquad \textbf{und} \qquad \sigma_\varepsilon^2 \;=\; 0$$

gilt.

10.4.7 Multiplikative Überlagerung beider (aller) Variablen

Falls auch die abhängige Variable y nach dem Höhneverfahren multiplikativ überlagert wird, gilt für diese überlagerte Variable[45]

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

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

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

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

$$\begin{aligned} y^a_{it} - \overline{y^a}_{i\bullet} = {} & (1 + \delta D_i)\,(x_{it} - \overline{x}_{i\bullet})\,\beta + (1 + \delta D_i)\,(\eta_{it} - \overline{\eta}_{i\bullet}) + (\alpha + \tau_i)\,(\varepsilon_{ity} - \overline{\varepsilon}_{i\bullet y}) \\ & + (x_{it}\,\varepsilon_{ity} - \overline{x\varepsilon}_{i\bullet y})\,\beta + \left(\varepsilon_{ity}\,\eta_{it} - \overline{\varepsilon\eta}_{i\bullet y}\right) \end{aligned} \qquad (10\text{-}38)$$

Um den Wahrscheinlichkeitsgrenzwert von $\hat{\beta}^a_W$ zu bestimmen, ist der Grenzwert des Zählers von (10-20), d.h.

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

unter Verwendung von (10-27) und (10-38) neu zu bestimmen, während für den Nenner die Ergebnisse aus dem Abschnitt 10.4.6 übernommen werden können. Dabei verwenden wir jetzt für die Überlagerung von x ebenfalls die Symbolik ε_{itx} für die hier relevante Störvariable und schreiben deshalb (10-27) und daraus folgende Ergebnis wie folgt:

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

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

und

$$x^a_{it} - \overline{x}^a_{i\bullet} = (1 + \delta D_i)\,(x_{it} - \overline{x}_{i\bullet}) + x_{it}\,\varepsilon_{itx} - \overline{x\varepsilon}_{i\bullet x} \qquad (10\text{-}39)$$

wobei

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

[45] Wieder unterstellen wir, daß das lineare Modell mit **stochastischen Effekten** gilt. Siehe Abschnitt 10.3.2. Dabei soll keine Korrelation zwischen diesem Effekt und dem Regressor x bestehen. Eine Verletzung dieser Annahme wird in Abschnitt 10.6 behandelt.

Tabelle 10.1: Erläuterungen zum Grenzwert (10-40)

Summanden von $y^a_{it} - \overline{y^a}_{i\bullet}$	Summanden von $x^a_{it} - \overline{x}^a_{i\bullet}$	
	$(1 + \delta D_i)\,(x_{it} - \overline{x}_{i\bullet})$	$x_{it}\,\varepsilon_{itx} - \overline{x\varepsilon}_{i\bullet x}$
$(1 + \delta D_i)\,(x_{it} - \overline{x}_{i\bullet})\,\beta$	$\Phi_t\beta$	0
$(1 + \delta D_i)\,\left(\eta_{ity} - \overline{\eta}_{i\bullet y}\right)$	0	0
$(\alpha + \tau_i)\,(\varepsilon_{ity} - \overline{\varepsilon}_{i\bullet y})$	0	0
$(x_{it}\,\varepsilon_{it} - \overline{x\varepsilon}_{i\bullet})\,\beta$	0	0
$\left(\varepsilon_{ity}\,\eta_{it} - \overline{\varepsilon\eta}_{i\bullet y}\right)$	0	0
Hinweis: In den einzelnen Zellen wird der jeweilige Grenzwert angegeben.		

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

Im folgenden werden die einzelnen Terme zur Berechnung des Grenzwertes des Zählers von (10-20), d.h. von

$$\text{plim}_{n\to\infty}\,\frac{1}{nT}\,\sum_{t=1}^{T}\sum_{i=1}^{n}(x^a_{it} - \overline{x}^a_{i\bullet})(y^a_{it} - \overline{y^a}_{i\bullet}) \qquad , \tag{10-40}$$

unter Verwendung von (10-39) und (10-38) systematisch dargestellt. Siehe dazu Tabelle 10.1.

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

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

führt, einen von Null verschiedenen Grenzwert ergibt. Wie bereits in Unterabschnitt 10.4.6 gezeigt, erhalten wir

$$\text{plim}_{n\to\infty}\,\frac{1}{n}\,\sum_{i}\,\left((1 + \delta D_i)^2\,(x_{it} - \overline{x}_{i\bullet})\right)^2\,\beta \quad = \quad \Phi_t\beta \quad = \quad (1 + \delta^2)\,\sigma_x^2\,\beta$$

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:

$$\text{plim}(\hat{\beta}^a_W) \;=\; \frac{(1 + \delta^2)\,\sigma_x^2\,\beta}{(1 + \delta^2)\,\sigma_x^2 \;+\; \sigma_\varepsilon^2\,(\sigma_x^2 \;+\; \mu_x^2)} \qquad . \tag{10-41}$$

Ein Vergleich mit dem in (10-37) 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

$$\delta \;=\; 0 \qquad \textbf{und} \qquad \sigma_\varepsilon^2 \;=\; 0$$

gilt.

10.4.8 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 10.4.7 wie folgt modifiziert: Es gibt nur eine Spalte, und zwar mit dem Summanden $x_{it} - \overline{x}_{i\bullet}$, der jeweils mit den Zeilenausdrücken bezüglich der abhängigen Variablen zu kombinieren ist. Der einzige von Null verschiedenen Grenzwert ergibt sich für die oberste Zeile mit $A_t = \sigma_x^2\beta$. Da auch der Nenner von $\hat{\beta}_W^a$ gegen σ_x^2 strebt, erhalten wir als Grenzwert insgesamt

$$\frac{\sigma_x^2}{\sigma_x^2}\,\beta \quad = \quad \beta \qquad .$$

10.5 Korrekturschätzer

Unter der Annahme, daß dem Datennutzer die Paranmeter der stochastischen Überlagerung, also δ und σ_ϵ^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 als auch y multiplikativ mit dem speziellen Höhne-Verfahren überlagert werden. Das entsprechende Ergebnis für den "naiven" Panelschätzer zeigt (10-41). Daraus erhalten wir bei Auflösung nach β folgenden Korrekturschätzer:

$$\hat{\beta}_W^{a,korr} \;=\; \frac{(1+\delta^2)\,\widehat{\sigma_x^2} \;+\; \sigma_\varepsilon^2\,(\widehat{\sigma_x^2} \;+\; \widehat{\mu_x}^2)}{(1\;+\;\delta^2)\,\widehat{\sigma_x^2}}\;\hat{\beta}_W^a \qquad . \tag{10-42}$$

Dabei ergibt sich unter Verwendung von (10-11) folgende Schätzung für σ_x^2:

$$\widehat{\sigma_x^2} \;=\; \frac{s_{x^a}^2 \;-\; (\delta^2 \;+\; \sigma_\varepsilon^2)\,\overline{x}^{a\,2}}{1 \;+\; \delta^2 \;+\; \sigma_\varepsilon^2} \qquad , \tag{10-43}$$

wobei

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

und

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

verwendet werden sollten. Ferner sollte

$$\widehat{\mu_x} \quad = \quad \overline{x}^a$$

benutzt werden.

Ein entsprechendes Ergebnis für den Korrekturschätzer im Fall mehrerer Regressoren findet sich in Appendix D.6.

10.6 Nichtexogenität des Regressors

Bekanntlich reagiert der in (10-16) gegebene Within-Schätzer nicht auf eine mögliche Korrelation zwischen Regressor x_{it} und Effekt τ_i, da der Effekt in der Schätzformel gar nicht

auftaucht. Siehe dazu die äquivalente Formulierung des Schätzers in (10-18). Dies gilt allerdings natürlich nur für den Fall nicht anonymisierter Variablen. Deshalb ist im Folgenden zu untersuchen, ob sich dieses Ergebnis ändert, wenn stochastisch überlagerte Variablen in der "naiven" Within-Schätzung (10-20) verwendet werden. Dabei werden wir - für additive und multiplikative Variante - zunächst den "allgemeinen" Fall betrachten und dann Folgerungen für den speziellen Fall der Überlagerung mittels Mischungsverteilung (spezielle Höhne-Spezifikation) anstellen.

10.6.1 Additiver Fall

In Abschnitt 10.4.2 habe ich den Bias für den Fall abgeleitet, daß nur der Regressor additiv überlagert wird, d.h. daß

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

und damit

$$\overline{x}_{i\bullet}^a \quad = \quad \overline{x}_{i\bullet} \; + \; \overline{u}_{i\bullet x}$$

und

$$x_{it}^a \; - \; \overline{x}_{i\bullet}^a \quad = \quad x_{it} \; - \; \overline{x}_{i\bullet} \; + \; u_{itx} \; - \; \overline{u}_{i\bullet x}$$

gilt. Zusätzlich soll nun y additiv überlagert sein, d.h. es soll ebenfalls

$$y_{it}^a \; = \; y_{it} \; + \; u_{ity} \; = \; \alpha \; + \; \beta \, x_{it} \; + \; \tau_i \; + \; \eta_{it} \; + \; u_{ity}$$

und damit

$$\overline{y}_{i\bullet}^a \quad = \quad \alpha \; + \; \beta \, \overline{x}_{i\bullet} \; + \; \tau_i \; + \; \overline{\eta}_{i\bullet} \; + \; \overline{u}_{i\bullet y}$$

sowie

$$y_{it}^a \; - \; \overline{y}_{i\bullet}^a \; = \; \beta \, (x_{it} \; - \; \overline{x}_{i\bullet}) \; + \; \eta_{it} \; - \; \overline{\eta}_{i\bullet} \; + \; u_{ity} \; - \; \overline{u}_{i\bullet y}$$

gelten. Damit ist klar, daß der Ausdruck

$$(x_{it}^a \; - \; \overline{x}_{i\bullet}^a) \, (y_{it}^a \; - \; \overline{y}_{i\bullet}^a)$$

den stochastischen Effekt τ_i nicht enthält und damit die möglicherweise vorhandene Korrelation zwischen x und τ die Bestimmung des Bias nicht berührt. Daß sich die Reststreuung durch die Hinzufügung von u_{ity} erhöht hat, wirkt sich bekanntlich nicht auf die Bestimmunng des Grenzwertes aus. Wir erhalten also denselben Wahrscheinlichkeitsgrenzwert wie in Abschnitt 10.4.2 (siehe (10-21)), den wir jetzt wie folgt schreiben:

$$\text{plim}_{n \to \infty} \, \hat{\beta}^a \quad = \quad \frac{\sigma_x^2}{\sigma_x^2 \; + \; \left(1 \; - \; \frac{1}{T}\right) \, \gamma_x^2} \, \beta \tag{10-1}$$

wobei wir jetzt für die Varianz von u_{itx} das Symbol γ_x^2 anstelle von σ_u^2 verwenden.

Falls die beiden Fehlervariablen korreliert sind, ergibt sich entsprechend Abschnitt 9.3 (siehe (9-15)) allerdings ein anderer Grenzwert:

$$\text{plim}_{n \to \infty} \, \hat{\beta}^a \quad = \quad \frac{\sigma_x^2}{\sigma_x^2 \; + \; \left(1 \; - \; \frac{1}{T}\right) \, \gamma_x^2} \, \beta \; + \; \frac{\gamma_{xy}}{\sigma_x^2 \; + \; \left(1 \; - \; \frac{1}{T}\right) \, \gamma_x^2}$$

Dabei bezeichnet γ_{xy} die Kovarianz zwischen u_x und u_y.

Allerdings ist diese Aussage nicht allgemein gültig: Wir wissen, daß die gemeinsame Überlagerung mittels der Mischungsverteilung gemäß der speziellen Höhne-Spezifikation Korrelation zwischen den Fehlern u_x und u_y bedingt, nicht aber zwischen den "Restfehlern" ε_x und ε_y und daß der gemeinsame Faktor D, der diese Korrelation bewirkt, durch die Differenzenbildung aus den Ausdrücken $u_{itx} - \overline{u}_{i\bullet x}$ und $u_{ity} - \overline{u}_{i\bullet y}$ herausfällt. Deshalb ergibt sich in diesem speziellen Fall **kein** zusätzlicher Effekt auf den Bias durch die Korrelation der Fehlervariablen.

Zusammenfassend kann festgestellt werden, daß im additiven Fall der naive Panelschätzer (10-20) durch die mögliche Korrelation zwischen Regressor x und stochastischem (individuenspezifischen) Effekt τ_i (d.h. $\sigma_{\tau x} \neq 0$) niemals beeinflußt wird. Dagegen ergibt sich bei Korrelation der Fehlervariablen (d.h. $\gamma_{xy} \neq 0$) ein Effekt, sofern die "allgemeine" Überlagerung verwendet wird, nicht aber im speziellen Höhnefall.

10.6.2 Multiplikativer Fall

Wir beginnen auch hier mit der "allgemeinen" Betrachtung , d.h. wir unterstellen

$$x^a_{it} \;=\; x_{it}\, u_{itx} \quad \text{und} \quad y^a_{it} \;=\; y_{it}\, u_{ity}$$

mit

$$V[u_{itx}] \;=\; \gamma_x^2 \quad \text{und} \quad V[u_{ity}] \;=\; \gamma_y^2 \quad \text{sowie} \quad COV[u_x\, u_y] \;=\; \gamma_{xy}$$

und erhalten

$$x^a_{it} \;-\; \overline{x}^a_{i\bullet} \quad = \quad x_{it}\, u_{itx} \;-\; \overline{xu}_{i\bullet x}$$

sowie

$$y^a_{it} \;-\; \overline{y}^a_{i\bullet} \quad = \quad (\alpha \;+\; \tau_i)\,(u_{ity} \;-\; \overline{u}_{i\bullet y}) \;+\; \beta\,(x_{it}\, u_{ity} \;-\; \overline{xu}_{i\bullet y}) \;+\; (\eta_{it}\, u_{ity} \;-\; \overline{\eta u}_{i\bullet y}) \quad .$$

Wir betrachten nun den Wahrscheinlichkeitsgrenzwert von

$$\frac{\frac{1}{n}\,\sum_{t=1}^{T}\sum_{i=1}^{n}(x^a_{it} - \overline{x}^a{}_{i\bullet})(y^a_{it} - \overline{y^a}_{i\bullet})}{\frac{1}{n}\,\sum_{t=1}^{T}\sum_{i=1}^{n}(x^a_{it} - \overline{x}^a{}_{i\bullet})^2}$$

getrennt für Nenner und Zähler, beginnend mit dem Nenner.

Wir erhalten direkt

$$\text{plim}_{n\to\infty}\; \frac{1}{n}\,\sum_{i=1}^{n}(x^a_{it} - \overline{x}^a{}_{i\bullet})^2 \quad = \quad \sigma_x^2 \;+\; \gamma_x^2(\sigma_x^2 \;+\; \mu_x^2) \qquad \text{(10-2)}$$

was aus der allgemeinen Formel der Varianz des Produktes von x und u folgt (siehe z.B. *****).

Ferner ergibt sich für die einzelnen Terme im Zähler Folgendes:[46]

$$\text{plim}_{n\to\infty}\; \frac{1}{n}\,\sum_{i=1}^{n}\,\alpha\,(x_{it}\, u_{itx} \;-\; \overline{xu}_{i\bullet x})\,(\,(u_{ity} \;-\; \overline{u}_{i\bullet y}) \quad = \quad \alpha\,\mu_x\,\gamma_{xy} \quad ,$$

[46] Die Beweise dazu finden sich in Appendix E.

$$\text{plim}_{n\to\infty}\ \frac{1}{n}\sum_{i=1}^{n}\tau_i\,(x_{it}\,u_{itx}\ -\ \overline{xu}_{i\bullet x})\,(u_{ity}\ -\ \overline{u}_{i\bullet y})\quad=\quad\sigma_{\tau x}\gamma_{xy}\quad,$$

$$\text{plim}_{n\to\infty}\ \frac{1}{n}\sum_{i=1}^{n}\beta\,(x_{it}\,u_{itx}\ -\ \overline{xu}_{i\bullet x})\,(x_{it}\,u_{ity}\ -\ \overline{xu}_{i\bullet y})\quad=\quad\beta(\sigma_x^2\ +\ \gamma_{xy}(\mu_x^2\ +\ \sigma_x^2))\quad,$$

$$\text{plim}_{n\to\infty}\ \frac{1}{n}\sum_{i=1}^{n}(x_{it}\,u_{itx}\ -\ \overline{xu}_{i\bullet x})\,(\eta_{it}\,u_{ity}\ -\ \overline{\eta u}_{i\bullet y})\quad=\quad 0\quad.$$

Demnach erhalten wir als Wahrscheinlichkeitsgrenzwert für den "naiven" Within-Schätzer:

$$\begin{aligned}\text{plim}_{n\to\infty}\ \hat{\beta}_W^a\ &=\ \frac{\alpha\,\mu_x\,\gamma_{xy}\ +\ \sigma_{\tau x}\gamma_{xy}\ +\ \beta(\sigma_x^2\ +\ \gamma_{xy}(\mu_x^2\ +\ \sigma_x^2))\ +\ 0}{\sigma_x^2\ +\ \gamma_x^2(\sigma_x^2\ +\ \mu_x^2)}\\ &=\ \frac{\sigma_x^2\,\beta\ +\ \left\{\alpha\,\mu_x\ +\ \sigma_{\tau x}\ +\ (\mu_x^2\ +\ \sigma_x^2)\,\beta)\right\}\,\gamma_{xy}}{\sigma_x^2\ +\ \gamma_x^2(\sigma_x^2\ +\ \mu_x^2)}\end{aligned}\qquad\text{(10-3)}$$

Formel (10-3) zeigt, daß im Fall einer multiplikativen Überlagerung die Schätzung von β durch die mögliche Korrelation zwischen Regressor x und stochastischem (individuenspezifischen) Effekt τ_i (d.h. $\sigma_{\tau x} \neq 0$) überhaupt nur dann beeinflußt werden kann, wenn die beiden Fehlervariablen miteinander korreliert sind, d.h. wenn $\gamma_{xy} \neq 0$ gilt. Dann hat übrigens auch das Absolutglied α einen Effekt auf die Schätzung von β.

Allerdings muß genau wie im additiven Fall (Abschnitt 10.6.1) auch hier eine Ergänzung bezüglich der (multiplikativen) Überlagerung mittels der speziellen Höhne-Spezifikation (Mischungsverteilung) erfolgen: Aus dem Studium der Tabelle 10.1 erkennt man, daß der Effekt τ_i in den Ausdrücken

$$(\alpha\ +\ \tau_i)\ (\varepsilon_{ity}\ -\ \overline{\varepsilon}_{i\bullet y})\ (1\ +\ \delta D_i)\,(x_{it}\ -\ \overline{x}_{i\bullet})$$

$$(\alpha\ +\ \tau_i)\ (\varepsilon_{ity}\ -\ \overline{\varepsilon}_{i\bullet y})\ (x_{it}\,\varepsilon_{itx}\ -\ \overline{x\varepsilon}_{i\bullet x})$$

auf den Regressor x "trifft". Da allerdings in beiden Fällen der Restfehler u_{ity} allein bzw. gemeinsam mit u_{ity} involviert ist und diese beiden Fehler gemäß Annahme unkorreliert sind, bleiben auch im Fall $\sigma_{\tau x} \neq 0$ die Ergebnisse aus Tabelle 10.1 gültig, d.h. es gilt weiterhin (10-41).

11 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 Statististical 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.(2005) *Statistische Methoden in der empirischen Wirtschaftsforschung*. Münster : Lit-Verlag.

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). Äuswirkungen datenverändernder Anonymisierungsverfahren auf die Analyse von Mikrodaten". IAW Tübingen.: 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.

Stata Corporation (2003). Stata Statistical Software. Release 8.0. "Cross-sectional Time-Series".

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 (Abschnitt 6.2)

In Abschnitt 6.2 haben wir nur die multiplikative Variante des "speziellen" Höhne-Verfahrens betrachtet. Hier ergänzen wir diese Ergebnisse für den "additiven" Fall:

$$\mathbf{U} \quad = \quad \mu\, D\, \boldsymbol{\iota} \; + \; \boldsymbol{\varepsilon} \quad , \tag{A-1}$$

wobei μ ein Parameter mit beliebigem Wert ist. Genau wie in Abschnitt 6.2 ist die diskrete Zufallsvariable D durch

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

gegeben (siehe (6-2)) und der r-dimensionale stetige Zufallsvektor $\boldsymbol{\varepsilon}$ ist entsprechend (6-3) spezifiziert:

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

wobei die Annahme der Normalverteilung üblicherweise hinzutritt. Genau wie in Abschnitt 6.2 wird die stochastische Unabhängigkeit von D und $\boldsymbol{\varepsilon}$ unterstellt.

Im symmetrischen Fall ergibt sich für den additiven Fall

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

und

$$cov[\mathbf{U}] \; = \; \sigma_{\varepsilon}^2 \mathbf{I} \; + \; \mu^2\, \boldsymbol{\iota}\boldsymbol{\iota}' \; = \; \frac{1}{\sigma_{\varepsilon}^2 + \mu^2} \left((1-\rho)\, \mathbf{I} \; + \; \rho\, \boldsymbol{\iota}\boldsymbol{\iota}' \right) \tag{A-3}$$

wobei

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

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

B Flexible Höhne-Spezifikation (Abschnitt 9)

B.1 Multiplikativer Fall

Im Abschnitt 9, in dem die Auswirkungen der Überlagerung auf die Schätzung eines linearen Modells untersuche, betrachte ich (9-11), die als eine "flexible" Variante von (9-10) interpretiert werden kann und die hier ohne den Beobachtungs-Index $\langle i \rangle$ geschrieben wird:

$$\mathbf{u} \quad = \quad (\boldsymbol{\iota} + \delta\, \mathbf{D}) \; + \; \boldsymbol{\varepsilon} \quad . \quad i = 1, \; \ldots \; , n \; . \tag{B-1}$$

Dabei ist $\mathbf{D}$ jetzt ein r-dimensionaler Zufallsvektor, dessen Komponenten jeweils die Spezifikation (6-2) 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 6.2.1 und 6.2.2 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

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

gilt, ergibt sich für Erwartungswert und Kovarianzmatrix

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

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

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

Diese Ergebnisse entsprechen für Erwartungswerte und Varianzen, aber **nicht für Kovarianzen** den Ergebnissen der spezielleren Höhne-Spezifikation (9-10). 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:

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

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 ε_j die einzelnen U_j (gegeben D_j) stochastisch voneinander unabhägig (bedingte Unabhängigkeit). Dann können wir für die multivariate bedingte Verteilung von $\mathbf{U}$ gegeben $\mathbf{D}$ schreiben:

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

Dabei bezeichnet N_j die eindimensionale Dichte der Normalverteilung.

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

$$h(\mathbf{d}) \quad = \quad h(d_1) \cdot \,\ldots\, \cdot h(d_r) \qquad .$$

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

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

Um die Randdichte des Vektors **U** zu bestimmen, müssen die einzelnen Komponenten von **D** 'heraussummiert' werden. Allerdings bedeutet dies, daß alle möglichen Ereignisse von **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:

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

Dann erhalten wir bei Summation über diese Ereignisse

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

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

Die allgemeine Formulierung der Dichtefunktion lautet demnach

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

B.2 Additiver Fall

Im additiven Fall schreiben wir die "flexible" Spezifikation als

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

Dabei ist **D** jetzt ein r-dimensionaler Zufallsvektor, dessen Komponenten jeweils die Spezifikation (6-2) erfüllen und die stochastisch voneinander unabhängig sind.

Im symmetrischen Fall ergibt sich für den additiven Fall

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

und

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

C Alternativer Beweis für (9-20) (Abschnitt 9.4)

Es soll gezeigt werden, daß auch mit der Beweis-Methode von Lin(1989) der Wahrscheinlichkeitsgrenzwert für den "naiven" Schätzer bei gemeinsamer multiplikativer Überlagerung aller Variablen abgeleitet werden kann. Siehe dazu Abschnitt 9.4.3.

Es wird unterstellt, daß die i-te Zeile der Regressormatrix $\mathbf{X}$ bzw $\mathbf{X^a}$ durch den Spaltenvektor

$$\mathbf{x}\langle i\rangle \quad = \quad \begin{pmatrix} 1 \\ x_{i2} \\ \vdots \\ x_{iK} \end{pmatrix} \quad = \quad \begin{pmatrix} 1 \\ \mathbf{x}_2 \end{pmatrix} \quad \text{und} \quad \mathbf{x^a}\langle i\rangle \quad = \quad \begin{pmatrix} 1 \\ x^a_{i2} \\ \vdots \\ x^a_{iK} \end{pmatrix} \quad = \quad \begin{pmatrix} 1 \\ \mathbf{x^a}_2 \end{pmatrix}$$

gegeben ist und der korrespondierende Überlagerungsvektor $\mathbf{u}_x$ folgende Gestalt hat:

$$\mathbf{u}_x\langle i\rangle \quad = \quad \begin{pmatrix} 1 \\ u_{2x} \\ \vdots \\ u_{Kx} \end{pmatrix} \quad = \quad \begin{pmatrix} 1 \\ \mathbf{u}_{2x} \end{pmatrix}$$

Damit wird dem Absolutglied bzw. dem Einsvektor in der Regressormatrix $\mathbf{X}$ Rechnung getragen.

Es gilt

$$\begin{aligned} 1/n\,\mathbf{X^{a\prime}}\,\mathbf{X^a} &= 1/n\ \textstyle\sum_{i=1}^n \mathbf{x^a}\langle i\rangle\,\mathbf{x^a}\langle i\rangle' \\ &= 1/n\ \textstyle\sum_{i=1}^n (\mathbf{x}\langle i\rangle \odot \mathbf{u}_x\langle i\rangle)\,(\mathbf{x}\langle i\rangle \odot \mathbf{u}_x\langle i\rangle)' \\ &= 1/n\ \textstyle\sum_{i=1}^n \mathbf{x}\langle i\rangle\,(\mathbf{x}\langle i\rangle' \odot \mathbf{u}_x\langle i\rangle\mathbf{u}_x\langle i\rangle' \end{aligned}$$

und für den Erwartungswert ergibt sich

$$E[1/n\,\mathbf{X^{a\prime}}\,\mathbf{X^a}] \quad = \quad \begin{pmatrix} 1 & \boldsymbol{\mu}'_x \\ \boldsymbol{\mu}_x & \mathbf{P} \odot \mathbf{M} \end{pmatrix}$$

mit

$$\mathbf{P} \quad = \quad E[1/n\ \textstyle\sum_{i=1}^n \mathbf{x_2}\langle i\rangle\,(\mathbf{x_2}\langle i\rangle'] \quad = \quad cov[\mathbf{x}] \ + \ \boldsymbol{\mu}_x\,\boldsymbol{\mu}'_x$$

und

$$\mathbf{M} \quad = \quad E[\mathbf{u}_x\langle i\rangle\mathbf{u}_x\langle i\rangle'] \quad = \quad cov[\mathbf{u_{2x}}] \ + \ \boldsymbol{\iota}\,\boldsymbol{\iota}'$$

Unter Verwendung der Formel für die Invertierung einer zerlegten Matrix[47] erhalten wir ferner

$$\begin{pmatrix} 1 & \boldsymbol{\mu}'_x \\ \boldsymbol{\mu}_x & \mathbf{P} \odot \mathbf{M} \end{pmatrix}^{-1} = \begin{pmatrix} 1 + \boldsymbol{\mu}'_x(\mathbf{P} \odot \mathbf{M} - \boldsymbol{\mu}_x\boldsymbol{\mu}'_x)^{-1}\boldsymbol{\mu}_x & -\boldsymbol{\mu}'_x(\mathbf{P} \odot \mathbf{M} - \boldsymbol{\mu}_x\boldsymbol{\mu}'_x)^{-1} \\ -(\mathbf{P} \odot \mathbf{M} - \boldsymbol{\mu}_x\boldsymbol{\mu}'_x)^{-1}\boldsymbol{\mu}_x & (\mathbf{P} \odot \mathbf{M} - \boldsymbol{\mu}_x\boldsymbol{\mu}'_x)^{-1} \end{pmatrix}$$

[47] Allgemein gilt für die reguläre Matrix

$$\mathbf{A} = \begin{pmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{pmatrix} \quad ,$$

die Formel

$$\mathbf{A}^{-1} = \begin{pmatrix} \mathbf{A}_{11}^{-1}(\mathbf{I} + \mathbf{A}_{12}\mathbf{D}^{-1}\mathbf{A}_{21}\mathbf{A}_{11}^{-1}) & -\mathbf{A}_{11}^{-1}\mathbf{A}_{12}\mathbf{D}^{-1}) \\ -\mathbf{D}^{-1}\mathbf{A}_{21}\mathbf{A}_{11}^{-1} & \mathbf{D}^{-1} \end{pmatrix}$$

mit $\mathbf{D} = \mathbf{A}_{22} - \mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{A}_{12}$.

Ferner gilt

$$
\begin{aligned}
1/n\,\mathbf{X^{a\prime}\,y^{a}} &= 1/n\ \textstyle\sum_{i=1}^{n}\left(\mathbf{x}\langle i\rangle \odot \mathbf{u}_x\langle i\rangle\right)\left(y_i \cdot u_{yi}\right)\\
&= 1/n\ \textstyle\sum_{i=1}^{n}\mathbf{x}\langle i\rangle\, y_i \odot \mathbf{u}_x\langle i\rangle\, u_{yi}\\
&= 1/n\ \textstyle\sum_{i=1}^{n}\mathbf{x}\langle i\rangle\left\{\mathbf{x}\langle i\rangle'\begin{pmatrix}\alpha\\ \boldsymbol{\beta}\end{pmatrix}+\varepsilon_i\right\}\odot\ \mathbf{u}_x\langle i\rangle\, u_{yi}\\
&= 1/n\ \textstyle\sum_{i=1}^{n}\begin{pmatrix}1\\ \mathbf{x}_2\langle i\rangle\end{pmatrix}\left\{\begin{pmatrix}1\\ \mathbf{x}_2\langle i\rangle\end{pmatrix}'\begin{pmatrix}\alpha\\ \boldsymbol{\beta}\end{pmatrix}+\varepsilon_i\right\}\odot\ \mathbf{u}_x\langle i\rangle\, u_{yi}\\
&= 1/n\ \textstyle\sum_{i=1}^{n}\begin{pmatrix}\alpha\ +\ \mathbf{x}_2\langle i\rangle'\boldsymbol{\beta}\ +\ \varepsilon_i\\ \alpha\,\mathbf{x}_2\langle i\rangle\ +\ \mathbf{x}_2\langle i\rangle\,\mathbf{x}_2\langle i\rangle'\,\boldsymbol{\beta}\ +\ \mathbf{x}_2\langle i\rangle\,\varepsilon_i\end{pmatrix}\odot\begin{pmatrix}u_{yi}\\ \mathbf{u}_{2x}\langle i\rangle\, u_{yi}\end{pmatrix}
\end{aligned}
$$

und für den Erwartungswert ergibt sich in diesem Fall

$$
E\left[1/n\,\mathbf{X^{a\prime}\,y^{a}}\right] = \begin{matrix}\alpha\ +\ \boldsymbol{\mu}_x'\,\boldsymbol{\beta}\\ \left\{\alpha\,\boldsymbol{\mu}_x\ +\ \left(cov[\mathbf{x_2}]\ +\ \boldsymbol{\mu}_x\,\boldsymbol{\mu}_x'\right)\ \boldsymbol{\beta}\right\}\ \odot\ cov[\mathbf{u_{2x}}, u_y]\end{matrix}
$$

wobei

$$
E[\mathbf{u}_{2x}\langle i\rangle\, u_{yi}] \quad = \quad cov[\mathbf{u_{2x}}, u_y]\ +\ \boldsymbol{\iota}
$$

verwendet wurde.

Damit ergibt sich für den naiven Schätzer bei Beschränkung auf den $K-1$-dimensionalen Vektor $\boldsymbol{\beta}$:

$$
\begin{aligned}
\operatorname{plim}\hat{\boldsymbol{\beta}}^{a} &= -\left(\mathbf{P}\odot\mathbf{M}-\boldsymbol{\mu}_x\boldsymbol{\mu}_x'\right)^{-1}\boldsymbol{\mu}_x\cdot\left(\alpha\ +\ \boldsymbol{\mu}_x'\,\boldsymbol{\beta}\right)\\
&\qquad +\left(\mathbf{P}\odot\mathbf{M}-\boldsymbol{\mu}_x\boldsymbol{\mu}_x'\right)^{-1}\ \left\{\alpha\,\boldsymbol{\mu}_x\ +\ \left(cov[\mathbf{x_2}]\ +\ \boldsymbol{\mu}_x\,\boldsymbol{\mu}_x'\right)\ \boldsymbol{\beta}\right\}\ \odot\ \left\{cov[\mathbf{u_{2x}}, u_y]\ +\ \boldsymbol{\iota}\right\}\\
&= \left(\mathbf{P}\odot\mathbf{M}-\boldsymbol{\mu}_x\boldsymbol{\mu}_x'\right)^{-1}\ \left(cov[\mathbf{x_2}]\,\boldsymbol{\beta}+\ \left\{\alpha\,\boldsymbol{\mu}_x\ +\ \left(cov[\mathbf{x_2}]\ +\ \boldsymbol{\mu}_x\,\boldsymbol{\mu}_x'\right)\ \boldsymbol{\beta}\right\}\ \odot\ cov[\mathbf{u_{2x}}, u_y]\right)
\end{aligned}
$$

Wenn wir weiter beachten, daß

$$
\begin{aligned}
\mathbf{P}\odot\mathbf{M}-\boldsymbol{\mu}_x\boldsymbol{\mu}_x' &= \left(cov[\mathbf{x_2}\ +\ \boldsymbol{\mu}_x\,\boldsymbol{\mu}_x'\right)\odot\left(cov[\mathbf{u_{x2}}\ +\ \boldsymbol{\iota}\,\boldsymbol{\iota}'\right)\ -\ \boldsymbol{\mu}_x\boldsymbol{\mu}_x'\\
&= cov[\mathbf{u_{x2}}]\ \left(cov[\mathbf{x_2}]\ +\ \boldsymbol{\mu}_x\,\boldsymbol{\mu}_x'\right)\ +\ cov[\mathbf{x_2}] \qquad ,
\end{aligned}
$$

dann können wir schreiben:

$$
\operatorname{plim}\hat{\boldsymbol{\beta}}^{a} = \left(cov[\mathbf{u_{x2}}](cov[\mathbf{x_2}]+\boldsymbol{\mu}_x\,\boldsymbol{\mu}_x')+cov[\mathbf{x_2}]\right)^{-1}\left(cov[\mathbf{x_2}]\boldsymbol{\beta}+\left\{\alpha\boldsymbol{\mu}_x+\left(cov[\mathbf{x_2}]+\boldsymbol{\mu}_x\boldsymbol{\mu}_x'\right)\boldsymbol{\beta}\right\}\odot cov[\mathbf{u_{2x}}, u_y]\right).
$$

Dieser Ausdruck ist äquivalent mit der Formel (9-20) in Abschnitt 9.4. Dabei ist zu beachten, daß $cov[\mathbf{x_2}]\ =\ \mathbf{Q}$, $cov[\mathbf{u_{2x}}]\ =\ cov[\mathbf{u_x}]$ und $cov[\mathbf{u_{2x}}, u_y]\ =\ cov[\mathbf{u_x}, u_y]$ im Abschnitt 9.4.

D Höhne-Verfahren und Paneldaten für mehrere Regressoren (Abschnitt 10)

D.1 Allgemeines

Wir betrachten hier das **multiple** Panelmodell mit stochastischen Individiualeffekten und mit K Regressoren und einem Absolutglied[48]

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

[48] Dies entspricht der Formel (10-17) im Fall der Einfachregression.

das zum einfachen Modell aus Abschnitt 10.3.2 korrespondiert. Für Mittelwert und Abweichung vom Mittelwert ergibt sich in diesem Fall

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

und

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

Im folgenden verwenden wir auch die kompakte Schreibweise dieses Modells:

$$\mathbf{y} \quad = \quad \alpha\,\boldsymbol{\iota} \;+\; \mathbf{X}\boldsymbol{\beta} \;+\; \boldsymbol{\tau} \;+\; \boldsymbol{\eta} \tag{D-2}$$

Dabei hat die Regressormatrix folgende Struktur:

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

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

$$\mathbf{X}\langle i\rangle \quad = \quad \begin{pmatrix} \mathbf{x}_1\langle i\rangle & \mathbf{x}_2\langle i\rangle & \ldots & \mathbf{x}_{K-1}\langle i\rangle & \mathbf{x}_K\langle i\rangle \end{pmatrix} \quad , \quad i = 1, \ldots, n \quad ,$$

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

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

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

Entsprechend schreiben wir den nT-dimensionalen Vektor $\mathbf{y}$ wie folgt:

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

Dabei ist jeweils der n-dimensionale Vektor $\mathbf{y}\langle i\rangle$ durch

$$\mathbf{y}\langle i\rangle \quad = \quad \begin{pmatrix} y_{i1} \\ y_{i2} \\ \vdots \\ y_{i,T-1} \\ y_{iT} \end{pmatrix}$$

gegeben.

Da der individuenspezifische Effekt nur von i abhängt, ergibt sich für den Zufallsvektor $\boldsymbol{\tau}$ die folgende Form:

$$\boldsymbol{\tau} \quad = \quad \begin{pmatrix} \boldsymbol{\tau}\langle 1\rangle \\ \\ \boldsymbol{\tau}\langle 2\rangle \\ \\ \vdots \\ \boldsymbol{\tau}\langle n-1\rangle \\ \\ \boldsymbol{\tau}\langle n\rangle \\ \\ \end{pmatrix} \quad \text{mit} \quad \boldsymbol{\tau}\langle i\rangle \;=\; \tau_i\, \boldsymbol{\iota}_T$$

D.1.1 Der Schätzer

Die Matrix $\mathbf{W}$ im Panelmodell mit fixen Effekten (siehe Abschnitt 10.3.1) hat folgende Struktur:

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

Daraus ergibt sich die idempotente Matrix

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

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

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

Dieser Schätzer ist erwartungstreu!

D.1.2 Eine rechentechnisch attraktivere Darstellung des Schätzers

Da die Matrix $NT \times NT$-Matrix $\mathbf{M}_W$ relativ viel Speicherplatz benötigt, soll hier kurz eine alternative Darstellung präsentiert werden, die ohne diese Matrix auskommt. Zunächst ist zu beachten, daß für die Matrix $\mathbf{M}_W\,\mathbf{X}$ folgendes gilt:[49]

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

Ferner ergibt sich für die quadratische Form von $\mathbf{X}'\,\mathbf{M}_W\,\mathbf{X}$ der Ausdruck

$$\mathbf{X}'\,\mathbf{M}_W\,\mathbf{X} \quad = \quad \sum_{i=1}^{n} \mathbf{X}'\langle i\rangle \left(\mathbf{I}_T \,-\, \frac{1}{T}\,\boldsymbol{\iota}_T\,\boldsymbol{\iota}_T'\right)\,\mathbf{X}\langle i\rangle$$

[49] Dieser Zusamenhang wird in den Abschnitten D.3 und D.4 ausgenutzt.

Entsprechend gilt für $\mathbf{X}'\,\mathbf{M}_W\,\mathbf{y}$ der Ausdruck

$$\mathbf{X}'\,\mathbf{M}_W\,\mathbf{y} \quad = \quad \sum_{i=1}^{n} \mathbf{X}'\langle i\rangle \,\left(\mathbf{I}_T \,-\, \frac{1}{T}\,\boldsymbol{\iota}_T\,\boldsymbol{\iota}_T'\right)\,\mathbf{y}\langle i\rangle \;.$$

Somit ergibt sich für den Schätzer alternativ zu (D-3) der Ausdruck

$$\hat{\boldsymbol{\beta}}_W \;=\; \left(\sum_{i=1}^{n} \mathbf{X}'\langle i\rangle \,\left(\mathbf{I}_T \,-\, \frac{1}{T}\,\boldsymbol{\iota}_T\,\boldsymbol{\iota}_T'\right)\,\mathbf{X}\langle i\rangle\right)^{-1} \,\sum_{i=1}^{n} \mathbf{X}'\langle i\rangle \,\left(\mathbf{I}_T \,-\, \frac{1}{T}\,\boldsymbol{\iota}_T\,\boldsymbol{\iota}_T'\right)\,\mathbf{y}\langle i\rangle \;. \quad \text{(D-4)}$$

Eine entsprechende Darstellung des "naiven" Panelschätzers bei anonymisierten Daten wird in den Abschnitten D.3 und D.4 verwendet.

D.1.3 Kovarianzmatrix des Schätzers

Um die Kovarianzmatrix des Schätzers angeben zu können, müssen wir zunächst die Kovarianzmatrix des Störterms $\boldsymbol{\tau} + \boldsymbol{\eta}$ aus (D-2) angeben:

$$cov[\boldsymbol{\tau} + \boldsymbol{\eta}] \;=\; cov[\boldsymbol{\eta}] \,+\, cov[\boldsymbol{\tau}] \;=\; \mathbf{I}_n \,\otimes\, \left(\sigma_\eta^2\,\mathbf{I}_T \,+\, \sigma_\tau^2\,\boldsymbol{\iota}_T\,\boldsymbol{\iota}_T'\right)$$

Ferner können wir für den Schätzer schreiben:

$$\hat{\boldsymbol{\beta}}_W \quad = \quad \boldsymbol{\beta} \,+\, (\mathbf{X}'\,\mathbf{M}_W\,\mathbf{X})^{-1}\,\mathbf{X}'\,\mathbf{M}_W\,(\boldsymbol{\tau} + \boldsymbol{\eta}) \quad ,$$

woraus die zuvor behauptete Erwartungstreue/Unverzerrtheit unmittelbar folgt. (Die Regressoren werden dabei als fix betrachtet; bedingte Betrachtung!)

Daraus erhalten wir die Kovarianzmatrix des Schätzers wie folgt:

$$cov[\hat{\boldsymbol{\beta}}_W] \quad = \quad (\mathbf{X}'\,\mathbf{M}_W\,\mathbf{X})^{-1}\,\mathbf{X}'\,\mathbf{M}_W\,E[(\boldsymbol{\tau} + \boldsymbol{\eta})\,(\boldsymbol{\tau} + \boldsymbol{\eta})']\,\mathbf{M}_W\,\mathbf{X}\,(\mathbf{X}'\,\mathbf{M}_W\,\mathbf{X})^{-1} \quad . \quad \text{(D-5)}$$

wobei $cov[\boldsymbol{\tau} + \boldsymbol{\eta}] \;=\; E[(\boldsymbol{\tau} + \boldsymbol{\eta})\,(\boldsymbol{\tau} + \boldsymbol{\eta})']$ gilt. Demnach erhalten wir für den "mittleren" Ausdruck $\mathbf{M}_W\,E[(\boldsymbol{\tau} + \boldsymbol{\eta})\,(\boldsymbol{\tau} + \boldsymbol{\eta})']\,\mathbf{M}_W$:

$$\mathbf{M}_W\,\left(\mathbf{I}_n \,\otimes\, \left(\sigma_\eta^2\,\mathbf{I}_T \,+\, \sigma_\tau^2\,\boldsymbol{\iota}_T\,\boldsymbol{\iota}_T'\right)\right)\,\mathbf{M}_W \quad = \quad \sigma_\eta^2\,\mathbf{M}_W$$

und damit

$$cov[\hat{\boldsymbol{\beta}}_W] \quad = \quad \sigma_\eta^2\,(\mathbf{X}'\,\mathbf{M}_W\,\mathbf{X})^{-1} \quad . \quad \text{(D-6)}$$

Man beachte, daß die Varianz des individuenspezifischen Effektes in dieser Formel nicht mehr auftaucht.

Für den Spezialfall nur eines Regressors, den wir in Abschnitt 10.3 betrachten, erhalten wir daraus

$$var[\hat{\boldsymbol{\beta}}_W] \quad = \quad \frac{\sigma_\eta^2}{(\mathbf{x}'\,\mathbf{M}_W\,\mathbf{x})} \quad = \quad \frac{\sigma_\eta^2}{\sum_{t=1}^{T}\sum_{i=1}^{n}(x_{it} - \overline{x}_{i\bullet})^2}. \quad \text{(D-7)}$$

D.2 Eine alternative Ableitung des 'Within'-Schätzers

Das STATA-Handbuch zur statistischen Analyse von Paneldaten[50] schlägt folgendes Schätzverfahren für die 'Within'-Schätzung vor: Es soll eine Kleinstquadrateschätzung der Parameter β_k aus dem folgenden Ansatz bestimmt werden:

$$y_{it} - \overline{y}_{i\bullet} + \overline{\overline{y}}_{\bullet\bullet} \quad = \quad \sum_{k=1}^{K} \beta_k \left(x_{itk} - \overline{x}_{i\bullet k} + \overline{\overline{x}}_{\bullet\bullet k} \right) + \overline{\tau} + \eta_{it} - \overline{\eta}_{i\bullet k} + \overline{\overline{\eta}}_{\bullet\bullet k}) \quad i = 1, \ldots, n \ , \ t = 1, \ldots T \quad , \tag{D-8}$$

Dabei gilt

$$\overline{\overline{y}}_{\bullet\bullet} \quad = \quad \frac{1}{n} \frac{1}{T} \sum_i \sum_t \ y_{it} \quad ,$$

d.h. $\overline{\overline{y}}_{\bullet\bullet}$ gibt das 'Gesamtmittel' (grand mean) für y an. Entsprechend sind die Gesamtmittelwerte für den Störterm η und die Regressoren x_k definiert. Ferner gilt

$$\overline{\tau} \ = \ \frac{1}{n} \sum_i \ \tau_i \quad .$$

Wir unterstellen ein Absolutglied und damit

$$x_{it1} \ = \ 1 \quad \text{für alle } i \text{ und } t$$

und damit ergibt sich für (D-8)

$$y_{it} \ - \ \overline{y}_{i\bullet} \ + \ \overline{\overline{y}}_{\bullet\bullet} \quad = \quad \beta_1 \ + \ \sum_{k=2}^{K} \beta_k \left(x_{itk} \ - \ \overline{x}_{i\bullet k} \ + \ \overline{\overline{x}}_{\bullet\bullet k} \right) \ + \ \overline{\tau} \ + \ \left(\eta_{it} \ - \ \overline{\eta}_{i\bullet k} \ + \ \overline{\overline{\eta}}_{\bullet\bullet k} \right) \quad , \tag{D-9}$$

Im Folgenden wird gezeigt, daß dies äquivalent zur Schätzung der β-Kopeffizienten gemäß (D-3) ist. Wir geben ferner an, wie daraus eine Schätzung für das Absolutglied β_1 gewonnen werden kann.

Wir betrachten zunächst die Matrix

$$\mathbf{A} \ = \ (\mathbf{I}_n \otimes \mathbf{I}_T) \ - \ \left(\mathbf{I}_n \otimes \frac{1}{T} \boldsymbol{\iota}_T \boldsymbol{\iota}_T' \right) \ + \ \left(\frac{1}{n} \boldsymbol{\iota}_n \boldsymbol{\iota}_n' \otimes \frac{1}{T} \boldsymbol{\iota}_T \boldsymbol{\iota}_T' \right) \tag{D-10}$$

Wie man nachprüfen kann, ist diese Matrix symmetrisch idempotent und erfüllt

$$\mathbf{A} \left(\boldsymbol{\iota}_n \otimes \boldsymbol{\iota}_T \right) \quad = \quad \left(\boldsymbol{\iota}_n \otimes \boldsymbol{\iota}_T \right) \tag{D-11}$$

sowie[51]

$$\mathbf{M}_\iota \, \mathbf{A} \quad = \quad \mathbf{I}_n \otimes \left(\mathbf{I}_T \ - \ \frac{1}{T} \boldsymbol{\iota}_T \boldsymbol{\iota}_T' \right) \quad = \quad \mathbf{M}_W \tag{D-12}$$

[50] Siehe STATA-Handbuch (STATA release 8 Cross-Sectional Time Series. Routine XTREG, S. 207.

[51] Der Tn-dimensionale Vektor $\boldsymbol{\iota}$ ist durch

$$\boldsymbol{\iota} \quad = \quad (\boldsymbol{\iota}_n \otimes \boldsymbol{\iota}_T)$$

definiert.

Bei Prämultiplikation des Vektors $\mathbf{y}$ mit $\mathbf{A}$ ergibt sich

$$\mathbf{A}\,\mathbf{y} \quad = \quad \mathbf{y} \; - \; \begin{pmatrix} \overline{y}_{1\bullet}\,\boldsymbol{\iota}_T \\ \overline{y}_{2\bullet}\,\boldsymbol{\iota}_T \\ \vdots \\ \overline{y}_{n\bullet}\,\boldsymbol{\iota}_T \end{pmatrix} + \begin{pmatrix} \overline{\overline{y}}_{\bullet\bullet}\,\boldsymbol{\iota}_T \\ \overline{\overline{y}}_{\bullet\bullet}\,\boldsymbol{\iota}_T \\ \vdots \\ \overline{\overline{y}}_{\bullet\bullet}\,\boldsymbol{\iota}_T \end{pmatrix} \quad ,$$

d.h. es ergibt sich die linke Seite von (D-8) als Vektor.

Ich benutze deshalb nun die Matrix $\mathbf{A}$ für eine zu (D-8) äquivalente kompakte Darstellung, die sich durch Prämultiplikation von (D-2) mit $\mathbf{A}$ ergibt:

$$\mathbf{A}\,\mathbf{y} \quad = \quad \mathbf{A}\,\mathbf{X}\boldsymbol{\beta} \; + \; \mathbf{A}\,\boldsymbol{\tau} \; + \; \mathbf{A}\,\boldsymbol{\eta} \tag{D-13}$$

Wegen der Struktur der Matrix $\mathbf{X}$ bei Vorhandensein eines Absolutglieds

$$\mathbf{X} \quad = \quad (\boldsymbol{\iota}_n \otimes \boldsymbol{\iota}_T \,,\, \mathbf{X}_2)$$

ergibt sich unter Beachtung der Ergebnisse für $\mathbf{A}$ oben

$$\mathbf{A}\,\mathbf{y} \quad = \quad \beta_1 \; (\boldsymbol{\iota}_n \otimes \boldsymbol{\iota}_T) \; + \; \mathbf{A}\,\mathbf{X}_2\boldsymbol{\beta}_2 \; + \; \mathbf{A}\,\boldsymbol{\tau} \; + \; \mathbf{A}\,\boldsymbol{\eta} \tag{D-14}$$

Der Kleinstquadrateschätzer für $\boldsymbol{\beta}_2$ läßt sich nun als Teilschätzung wie folgt schreiben:

$$\begin{array}{rcl} \hat{\boldsymbol{\beta}}_2 & = & \left((\mathbf{A}\,\mathbf{X}_2)'\,\mathbf{M}_{\boldsymbol{\iota}}\mathbf{A}\,\mathbf{X}_2\right)^{-1} \; (\mathbf{A}\,\mathbf{X}_2)'\,\mathbf{M}_{\boldsymbol{\iota}}\mathbf{A}\,\mathbf{y} \\ & = & \left(\mathbf{X}_2'\,\mathbf{M}_W\,\mathbf{X}_2\right)^{-1} \; \mathbf{X}_2'\,\mathbf{M}_W\,\mathbf{y} \end{array}$$

wobei $\mathbf{M}_{\boldsymbol{\iota}}\,\mathbf{A} \; = \; \mathbf{M}_W$ verwendet wurde. Dies ist aber genau der in (D-3) gegebene Schätzer.

Zur zusätzlichen Schätzung des Absolutglieds β_1 kann man die alternative Form der Teilschätzung verwenden.[52] Sie lautet

$$\begin{array}{rcl} \hat{\beta}_1 & = & \left((\mathbf{A}\,\boldsymbol{\iota})'\,\mathbf{A}\,\boldsymbol{\iota}\right)^{-1} \; (\mathbf{A}\,\boldsymbol{\iota})' \; \left(\mathbf{A}\,\mathbf{y} \; - \; \mathbf{A}\,\mathbf{X}_2\,\hat{\boldsymbol{\beta}}_2\right) \\ & = & \left(\mathbf{X}_2'\,\mathbf{A}\,\mathbf{X}_2\right)^{-1} \; \mathbf{X}_2' \; \left(\mathbf{A}\,\mathbf{y} \; - \; \mathbf{A}\,\mathbf{X}_2\,\hat{\boldsymbol{\beta}}_2\right) \\ & = & \dfrac{1}{n\,T}\,\boldsymbol{\iota}' \; \left(\mathbf{y} \; - \; \mathbf{X}_2\,\hat{\boldsymbol{\beta}}_2\right) \\ & = & \overline{\overline{y}} \quad - \quad \overline{\overline{\mathbf{x}}}'\,\hat{\boldsymbol{\beta}}_2 \end{array} \tag{D-15}$$

wobei $\overline{\overline{\mathbf{x}}}$ der $(K-1)$-dimensionale Vektor der Gesamtmittelwerte (grand means) für die einzelnen Regressoren ist.

D.3 Ausschließliche Überlagerung der Regressoren

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

$$\hat{\boldsymbol{\beta}}_W^a \quad = \quad (\mathbf{X}^{\mathbf{a}\prime}\,\mathbf{M}_W\,\mathbf{X}^{\mathbf{a}})^{-1}\,\mathbf{X}^{\mathbf{a}\prime}\,\mathbf{M}_W\,\mathbf{y} \quad . \tag{D-16}$$

[52] Siehe beispielsweise Greene (3. Auflage) Formel (6-24). Beachte: Der Tn-dimensionale Vektor $\boldsymbol{\iota}$ ist durch

$$\boldsymbol{\iota} \quad = \quad (\boldsymbol{\iota}_n \otimes \boldsymbol{\iota}_T)$$

definiert.

oder auch durch

$$\hat{\boldsymbol{\beta}}^a_W \;=\; \left(\sum_{i=1}^{n} \mathbf{X}^{\mathbf{a}\prime}\langle i\rangle \left(\mathbf{I}_T \;-\; \frac{1}{T}\,\boldsymbol{\iota}_T\,\boldsymbol{\iota}_T'\right) \mathbf{X}^{\mathbf{a}}\langle i\rangle\right)^{-1} \;\sum_{i=1}^{n} \mathbf{X}^{\mathbf{a}\prime}\langle i\rangle \left(\mathbf{I}_T \;-\; \frac{1}{T}\,\boldsymbol{\iota}_T\,\boldsymbol{\iota}_T'\right) \mathbf{y}^{\mathbf{a}}\langle i\rangle \;. \tag{D-17}$$

gegeben, wobei die zweite Form in Abschnitt D.1.2 erläutert wurde.

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

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

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

$$\begin{aligned}
&\mathbf{X}^{\mathbf{a}\prime}\,\mathbf{M}_W\,\mathbf{X}^{\mathbf{a}} \;= \\
&= \;\; \textstyle\sum_{i=1}^{n} \mathbf{X}^{\mathbf{a}\prime}\langle i\rangle \left(\mathbf{I}_T \;-\; \frac{1}{T}\,\boldsymbol{\iota}_T\,\boldsymbol{\iota}_T'\right) \mathbf{X}^{\mathbf{a}}\langle i\rangle \\
&= \sum_{i=1}^{n} \begin{pmatrix} \left(\mathbf{x}^a_1\langle i\rangle - \overline{\mathbf{x}^a_1}\langle i\bullet\rangle\right)^2 & \left(\mathbf{x}^a_1\langle i\rangle - \overline{\mathbf{x}^a_1}\langle i\bullet\rangle\right)'\left(\mathbf{x}^a_2\langle i\rangle - \overline{\mathbf{x}^a_2}\langle i\bullet\rangle\right) & \dots & \left(\mathbf{x}^a_1\langle i\rangle - \overline{\mathbf{x}^a_1}\langle i\bullet\rangle\right)'\left(\mathbf{x}^a_K\langle i\rangle - \overline{\mathbf{x}^a_K}\langle i\bullet\rangle\right) \\ \left(\mathbf{x}^a_2\langle i\rangle - \overline{\mathbf{x}^a_2}\langle i\bullet\rangle\right)'\left(\mathbf{x}^a_1\langle i\rangle - \overline{\mathbf{x}^a_1}\langle i\bullet\rangle\right) & \left(\mathbf{x}^a_2\langle i\rangle - \overline{\mathbf{x}^a_2}\langle i\bullet\rangle\right)^2 & & \left(\mathbf{x}^a_2\langle i\rangle - \overline{\mathbf{x}^a_2}\langle i\bullet\rangle\right)'\left(\mathbf{x}^a_K\langle i\rangle - \overline{\mathbf{x}^a_K}\langle i\bullet\rangle\right) \\ \vdots & \vdots & \ddots & \vdots \\ \left(\mathbf{x}^a_K\langle i\rangle - \overline{\mathbf{x}^a_K}\langle i\bullet\rangle\right)'\left(\mathbf{x}^a_1\langle i\rangle - \overline{\mathbf{x}^a_1}\langle i\bullet\rangle\right) & \left(\mathbf{x}^a_K\langle i\rangle - \overline{\mathbf{x}^a_K}\langle i\bullet\rangle\right)'\left(\mathbf{x}^a_2\langle i\rangle - \overline{\mathbf{x}^a_2}\langle i\bullet\rangle\right) & \dots & \left(\mathbf{x}^a_K\langle i\rangle - \overline{\mathbf{x}^a_K}\langle i\bullet\rangle\right)^2 \end{pmatrix} \\
&= \sum_{i=1}^{n}\sum_{t=1}^{T} \begin{pmatrix} \left(x^a_{it1} - \overline{x^a}_{i\bullet 1}\right)^2 & \left(x^a_{it1} - \overline{x^a}_{i\bullet 1}\right)\left(x^a_{it2} - \overline{x^a}_{i\bullet 2}\right) & \dots & \left(x^a_{it1} - \overline{x^a}_{i\bullet 1}\right)\left(x^a_{itK} - \overline{x^a}_{i\bullet K}\right) \\ \left(x^a_{it1} - \overline{x^a}_{i\bullet 2}\right)\left(x^a_{it1} - \overline{x^a}_{i\bullet 2}\right) & \left(x^a_{it2} - \overline{x^a}_{i\bullet 2}\right)^2 & \dots & \left(x^a_{it2} - \overline{x^a}_{i\bullet 2}\right)\left(x^a_{itK} - \overline{x^a}_{i\bullet K}\right) \\ \vdots & \vdots & \ddots & \vdots \\ \left(x^a_{itK} - \overline{x^a}_{i\bullet K}\right)\left(x^a_{it1} - \overline{x^a}_{i\bullet 1}\right) & \left(x^a_{itK} - \overline{x^a}_{i\bullet K}\right)\left(x^a_{it2} - \overline{x^a}_{i\bullet 2}\right) & \dots & \left(x^a_{itK} - \overline{x^a}_{i\bullet K}\right)^2 \end{pmatrix}
\end{aligned}$$

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

$$\overline{\mathbf{x}^a_k}\langle i\bullet\rangle \;=\; \frac{1}{T}\,\boldsymbol{\iota}_T\,\boldsymbol{\iota}_T'\,\mathbf{x}^a_k\langle i\rangle \;=\; \frac{1}{T}\sum_{t=1}^{T} \begin{pmatrix} x^a_{it1} \\ x^a_{it2} \\ \vdots \\ x^a_{itK} \end{pmatrix}$$

Ich greife nun auf die Ergebnisse aus Abschnitt 10.4.6 zurück, in dem die Überlagerung eines einzigen Regressors betrachtet wurde. Daraus ergibt sich bei Überlagerung mittels (10-1) für den k-ten Regressor

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

und

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

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

$$x^a_{itk} \;-\; \overline{x^a}_{i\bullet k} \;\;=\;\; (1 \;+\; \delta D_i)\,(x_{itk} \;-\; \overline{x}_{i\bullet k}) \;+\; x_{itk}\,\varepsilon_{itk} \;-\; \overline{x\varepsilon}_{i\bullet k} \tag{D-18}$$

Wie bereits in Abschnitt 10.4.6 bemerkt, verschwindet im Gegensatz zur additiven Überlagerung der Zuschlagsparameter (δ) in diesem Fall nicht! Ich habe im erwähnten Abschnitt auch bereits gezeigt, daß

$$\text{plim}_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \left(x^a_{itk} - \overline{x}^a_{i\bullet k}\right)^2 \quad = \quad (1 + \delta^2)\,\sigma_k^2 + \sigma_\varepsilon^2 \left(\sigma_k^2 + \mu_k^2\right) \tag{D-19}$$

gilt. Siehe die Ergebnisse zu (10-34) bis (10-36). Allerdings schreiben wir dieses Ergebnis jetzt für den k-ten Regressor mit Varianz σ_k^2 und Erwartungswert μ_k.

Zusätzlich benötigen wir jetzt ein ein Ergebnis für

$$\text{plim}_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \left(x^a_{itk} - \overline{x}^a_{i\bullet k}\right)\left(x^a_{ith} - \overline{x}^a_{i\bullet h}\right) \quad , k \neq h \quad .$$

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

$$\begin{aligned}
&\textstyle\sum_{i=1}^n \left(x^a_{itk} - \overline{x}^a_{i\bullet k}\right)\left(x^a_{ith} - \overline{x}^a_{i\bullet h}\right) \\
&= \textstyle\sum_{i=1}^n \left\{(1+\delta D_i)\left(x_{itk} - \overline{x}_{i\bullet k}\right) + \left(x_{itk}\varepsilon_{itk} - \overline{x\varepsilon}_{i\bullet k}\right)\right\}\left\{(1+\delta D_i)\left(x_{ith} - \overline{x}_{i\bullet h}\right) + \left(x_{ith}\varepsilon_{ith} - \overline{x\varepsilon}_{i\bullet h}\right)\right\} \\
&= \textstyle\sum_{i=1}^n \left\{(1+\delta D_i)^2\left(x_{itk} - \overline{x}_{i\bullet k}\right)\left(x_{ith} - \overline{x}_{i\bullet h}\right) + \left(x_{itk}\varepsilon_{itk} - \overline{x\varepsilon}_{i\bullet k}\right)\left(x_{ith}\varepsilon_{ith} - \overline{x\varepsilon}_{i\bullet h}\right)\right\} \\
&\quad + \textstyle\sum_{i=1}^n \left\{(1+\delta D_i)\left(x_{itk} - \overline{x}_{i\bullet k}\right)\left(x_{ith}\varepsilon_{ith} - \overline{x\varepsilon}_{i\bullet h}\right) + (1+\delta D_i)\left(x_{itk} - \overline{x}_{i\bullet k}\right)\left(x_{ith}\varepsilon_{ith} - \overline{x\varepsilon}_{i\bullet h}\right)\right\}
\end{aligned}$$

Aus dieser Zerlegung erkennt man, daß

$$\text{plim}_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \left(x^a_{itk} - \overline{x}^a_{i\bullet k}\right)\left(x^a_{ith} - \overline{x}^a_{i\bullet h}\right) \quad = \quad (1 + \delta^2)\,\sigma_{kh} \tag{D-20}$$

gilt, wobei σ_{kh} die Kovarianz der Regressoren k und h bezeichnet. Man beachte insbesondere, daß im Gegensatz zu (10-36) 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

$$\begin{aligned}
\underset{n\to\infty}{\text{plim}}\ \tfrac{1}{n}\,\mathbf{X^{a\prime}}\,\mathbf{M}_W\,\mathbf{X^a} \quad &= \quad \underset{n\to\infty}{\text{plim}}\ \tfrac{1}{n} \sum_{i=1}^{n} \mathbf{X^a}'_{\langle i\rangle} \left(\mathbf{I}_T - \tfrac{1}{T}\,\boldsymbol{\iota}_T\,\boldsymbol{\iota}_T'\right) \mathbf{X^a}_{\langle i\rangle} \\
&= \begin{pmatrix}
(1+\delta^2)\sigma_1^2 + \sigma_\varepsilon^2\left(\sigma_1^2+\mu_1^2\right) & (1+\delta^2)\,\sigma_{12} & \cdots & (1+\delta^2)\,\sigma_{1k} \\
(1+\delta^2)\,\sigma_{21} & (1+\delta^2)\sigma_2^2 + \sigma_\varepsilon^2\left(\sigma_2^2+\mu_2^2\right) & \cdots & (1+\delta^2)\,\sigma_{2K} \\
\vdots & \vdots & \ddots & \vdots \\
(1+\delta^2)\,\sigma_{K1} & (1+\delta^2)\,\sigma_{K2} & \cdots & (1+\delta^2)\sigma_K^2 + \sigma_\varepsilon^2\left(\sigma_K^2+\mu_K^2\right)
\end{pmatrix}
\end{aligned} \tag{D-21}$$

Inm 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

$$\text{plim}_{n\to\infty} \frac{1}{n}\,\mathbf{X^{a\prime}}\,\mathbf{M}_W\,\mathbf{y}$$

Dazu betrachten wir

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

Auch hier können die Ergebnisse aus Abschnitt 10.4.6 sowie zusätzlich aus D.1 verwendet werden. Für jeweils ein Element aus dem zuvor bestimmten Vektor ergibt sich

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

Eine Analyse der einzelnen Summanden entsprechend Abschnitt 10.4.6 ergibt, daß nur der erste Summand zu einem von Null verschiedenen Grenzwert führt, d.h. für bestimtes k ergibt sich

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

Demnach gilt

$$\text{plim}_{n\to\infty}\,\frac{1}{n}\,\mathbf{X^{a\prime}}\,\mathbf{M}_W\,\mathbf{y} \quad = \quad \begin{pmatrix} \sum_{h=1}^{K}\sigma_{1h}\,\beta_h \\ \sum_{h=1}^{K}\sigma_{2h}\,\beta_h \\ \vdots \\ \vdots \\ \sum_{h=1}^{K}\sigma_{Kh}\,\beta_h \end{pmatrix} \tag{D-22}$$

Für den Wahrscheinlichkeitsgrenzwert des naiven Panelschätzers (D-3) ergibt sich dann aus (D-21) und (D-22)

$$\text{plim}_{n\to\infty}\,\hat{\boldsymbol{\beta}}^a = \left(\text{plim}_{n\to\infty}\,\frac{1}{n}\,\mathbf{X^{a\prime}}\,\mathbf{M}_W\,\mathbf{X^a}\right)^{-1}\,\text{plim}_{n\to\infty}\,\frac{1}{n}\,\mathbf{X^{a\prime}}\,\mathbf{M}_W\,\mathbf{y}$$

$$= \begin{pmatrix} (1+\delta^2)\sigma_1^2+\sigma_\varepsilon^2\left(\sigma_1^2+\mu_1^2\right) & (1\,+\,\delta^2)\,\sigma_{12} & \dots & (1\,+\,\delta^2)\,\sigma_{1k} \\ (1\,+\,\delta^2)\,\sigma_{21} & (1+\delta^2)\sigma_2^2+\sigma_\varepsilon^2\left(\sigma_2^2+\mu_2^2\right) & \dots & (1\,+\,\delta^2)\,\sigma_{2K} \\ \vdots & \vdots & \ddots & \vdots \\ (1\,+\,\delta^2)\,\sigma_{K1} & (1\,+\,\delta^2)\,\sigma_{K2} & \dots & (1+\delta^2)\sigma_K^2+\sigma_\varepsilon^2\left(\sigma_K^2+\mu_K^2\right) \end{pmatrix}^{-1} \begin{pmatrix} \sum_{h=1}^K \sigma_{1h}\,\beta_h \\ \sum_{h=1}^K \sigma_{2h}\,\beta_h \\ \vdots \\ \vdots \\ \sum_{h=1}^K \sigma_{Kh}\,\beta_h \end{pmatrix}$$

(D-23)

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 δ als auch die Restüberlagerung durch ε entfällt.

D.4 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 10.4.7 bereits für einen einzigen Regressor ausgeführt, gilt dann für die anonymisierte abhängige Variable

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

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

$$\overline{y^a}_{i\bullet} \;=\; (1\,+\,\delta D_i)\,\tau_i + \,(1\,+\,\delta D_i)\,(\sum_{h=1}^K \beta_h\,\overline{x}_{i\bullet h}\,+\,\overline{\eta}_{i\bullet})\,+\,\tau_i\,\overline{\varepsilon}_{i\bullet y} + \sum_{h=1}^K \beta_h \overline{x\varepsilon}_{i\bullet hy} + \overline{\varepsilon\eta}_{i\bullet}$$

wobei

$$\overline{x\varepsilon}_{i\bullet ky} \;=\; \frac{1}{T}\sum_t x_{itk}\,\varepsilon_{ity}$$

und

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

Dann erhalten wir für die Abweichung vom Mittelwert

$$\begin{aligned} y_{it}^a\,-\,\overline{y^a}_{i\bullet} \;=\; (1\,+\,\delta D_i)\sum_{h=1}^K (x_{ith}\,-\,\overline{x}_{i\bullet h})\,\beta_h\,+\,(1\,+\,\delta D_i)\,(\eta_{it}\,-\,\overline{\eta}_{i\bullet})\,+\,\tau_i\,(\varepsilon_{ity}\,-\,\overline{\varepsilon}_{i\bullet y}) \\ +\sum_{h=1}^K \beta_h\,(x_{ith}\,\varepsilon_{ity}\,-\,\overline{x\varepsilon}_{i\bullet hy})\,+\,\left(\varepsilon_{ity}\,\eta_{it}\,-\,\overline{\varepsilon\eta}_{i\bullet y}\right) \end{aligned}$$

(D-24)

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

$$\text{plim}_{n\to\infty}\,\frac{1}{n}\,\mathbf{X^{a\prime}}\,\mathbf{M}_W\,\mathbf{y^a}$$

Dazu betrachten wir

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

Im folgenden werden die einzelnen Terme zur Berechnung des Grenzwertes von

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

unter Verwendung von (D-18) und (D-24) systematisch dargestellt.

	Summanden von $x_{it}^a\,-\,\overline{x}_{i\bullet}^a$	
Summanden von $y_{it}^a\,-\,\overline{y^a}_{i\bullet}$	$(1+\delta D_i)\,(x_{itk}\,-\,\overline{x}_{i\bullet k})$	$x_{itk}\,\varepsilon_{itk}\,-\,\overline{x\varepsilon}_{i\bullet k}$
$(1\,+\,\delta D_i)\,\sum_{h=1}^{K}\,(x_{ith}\,-\,\overline{x}_{i\bullet h})\,\beta_h$	$(1+\delta^2)\,\sum_{h=1}^{K}\,\sigma_{kh}\,\beta_h$	0
$(1\,+\,\delta D_i)\,(\eta_{it}\,-\,\overline{\eta}_{i\bullet})$	0	0
$\tau_i\,(\varepsilon_{ity}\,-\,\overline{\varepsilon}_{i\bullet y})$	0	0
$\sum_{h=1}^{K}\,\beta_h\,(x_{ith}\,\varepsilon_{ity}\,-\,\overline{x\varepsilon}_{i\bullet hy})$	0	0
$\left(\varepsilon_{ity}\,\eta_{it}\,-\,\overline{\varepsilon\eta}_{i\bullet y}\right)$	0	0
Hinweis: In den einzelnen Zellen wird der jeweilige Grenzwert angegeben.		

Eine Analyse analog zu der in Abschnitt 10.4.7 zeigt, daß ausschließlich das linke oberste Element einen von Null verschiedenen Grenzwert aufweist. Wir erhalten demnach

$$\text{plim}_{n\to\infty}\,\frac{1}{n}\,\mathbf{X^{a\prime}}\,\mathbf{M}_W\,\mathbf{y^a} \quad = \quad (1\,+\,\delta^2)\,\begin{pmatrix}
\sum_{h=1}^{K}\,\sigma_{1h}\,\beta_h\\
\sum_{h=1}^{K}\,\sigma_{2h}\,\beta_h\\
\vdots\\
\vdots\\
\sum_{h=1}^{K}\,\sigma_{Kh}\,\beta_h
\end{pmatrix} \qquad \text{(D-25)}$$

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

$$\begin{aligned}
\text{plim}_{n\to\infty}\,\hat{\boldsymbol{\beta}}^a &= \left(\text{plim}_{n\to\infty}\,\tfrac{1}{n}\,\mathbf{X^{a\prime}}\,\mathbf{M}_W\,\mathbf{X^a}\right)^{-1}\,\text{plim}_{n\to\infty}\,\tfrac{1}{n}\,\mathbf{X^{a\prime}}\,\mathbf{M}_W\,\mathbf{y^a} \\
&= (1+\delta^2)\begin{pmatrix}
(1+\delta^2)\sigma_1^2+\sigma_\varepsilon^2\left(\sigma_1^2+\mu_1^2\right) & (1\,+\,\delta^2)\,\sigma_{12} & \dots & (1\,+\,\delta^2)\,\sigma_{1k}\\
(1\,+\,\delta^2)\,\sigma_{21} & (1+\delta^2)\sigma_2^2+\sigma_\varepsilon^2\left(\sigma_2^2+\mu_2^2\right) & \dots & (1\,+\,\delta^2)\,\sigma_{2K}\\
\vdots & \vdots & \ddots & \vdots\\
(1\,+\,\delta^2)\,\sigma_{K1} & (1\,+\,\delta^2)\,\sigma_{K2} & \dots & (1+\delta^2)\sigma_K^2+\sigma_\varepsilon^2\left(\sigma_K^2+\mu_K^2\right)
\end{pmatrix}^{-1}\begin{pmatrix}
\sum_{h=1}^{K}\,\sigma_{1h}\,\beta_h\\
\sum_{h=1}^{K}\,\sigma_{2h}\,\beta_h\\
\vdots\\
\vdots\\
\sum_{h=1}^{K}\,\sigma_{Kh}\,\beta_h
\end{pmatrix}
\end{aligned} \qquad \text{(D-26)}$$

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 10.4.7.

D.5 Eine alternative Schreibweise

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

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

Dann können wir für (D-21) schreiben:

$$\text{plim}_{n\to\infty} \frac{1}{n} \mathbf{X^{a\prime}} \mathbf{M}_W \mathbf{X^a} \quad = \quad (1+\delta^2)\, cov[\mathbf{x}] \; + \; \sigma_\varepsilon^2 \begin{pmatrix} \sigma_1^2+\mu_1^2 & & & \\ & \sigma_2^2+\mu_2^2 & & \\ & & \ddots & \\ & & & \sigma_K^2+\mu_K^2 \end{pmatrix} \tag{D-27}$$

Entsprechend ergeben sich aus (D-23) und (D-26)

$$\text{plim}_{n\to\infty} \frac{1}{n} \mathbf{X^{a\prime}} \mathbf{M}_W \mathbf{y} \quad = \quad cov[\mathbf{x}]\, \boldsymbol{\beta} \tag{D-28}$$

bzw.

$$\text{plim}_{n\to\infty} \frac{1}{n} \mathbf{X^{a\prime}} \mathbf{M}_W \mathbf{y^a} \quad = \quad (1+\delta^2)\, cov[\mathbf{x}]\, \boldsymbol{\beta} \tag{D-29}$$

und damit für den naiven Schätzer im Fall der ausschließlichen Überlagerung der Regressoren

$$\underset{n\to\infty}{\text{plim}}\ \hat{\boldsymbol{\beta}^a}_W \quad = \quad \left((1+\delta^2)\, cov[\mathbf{x}] \; + \; \sigma_\varepsilon^2 \begin{pmatrix} \sigma_1^2+\mu_1^2 & & & \\ & \sigma_2^2+\mu_2^2 & & \\ & & \ddots & \\ & & & \sigma_K^2+\mu_K^2 \end{pmatrix} \right)^{-1} cov[\mathbf{x}]\, \boldsymbol{\beta} \tag{D-30}$$

bzw. bei Überlagerung aller Variablen

$$\underset{n\to\infty}{\text{plim}}\ \hat{\boldsymbol{\beta}^a}_W \quad = \quad \left(cov[\mathbf{x}] \; + \; \frac{\sigma_\varepsilon^2}{(1+\delta^2)} \begin{pmatrix} \sigma_1^2+\mu_1^2 & & & \\ & \sigma_2^2+\mu_2^2 & & \\ & & \ddots & \\ & & & \sigma_K^2+\mu_K^2 \end{pmatrix} \right)^{-1} cov[\mathbf{x}]\, \boldsymbol{\beta} \tag{D-31}$$

Aus diesen Formeln lassen sich Korrekturschätzer bestimmen. Dafür sind wieder geeignete Schätzer für die Mittelwerte und Varianzen sowie Kovarianzen der Regressorvariablen abzuleiten. Dies wird im folgenden Unterabschnitt für den Fall der gemeinsamen Überlagerung aller Variablen im Detail dargestellt.

D.6 Korrekturschätzer

Die Ergebnisse aus Abschnitt D.5 lassen sich zur Konstruktion eines - konsistenten - Korrekturschätzers verwenden: Aus (D-31) ergibt sich direkt

$$\hat{\boldsymbol{\beta}}_W^{a,korr} \quad = \quad cov[\mathbf{x}]^{-1} \left(cov[\mathbf{x}] \; + \; \frac{\sigma_\varepsilon^2}{1+\delta^2} \begin{pmatrix} \sigma_1^2+\mu_1^2 & & & \\ & \sigma_2^2+\mu_2^2 & & \\ & & \ddots & \\ & & & \sigma_K^2+\mu_K^2 \end{pmatrix} \right) \hat{\boldsymbol{\beta}}_W^a \tag{D-32}$$

für den Fall der gemeinsamen Überlagerung aller Variablen.[53] Allerdings sind Erwartungswerte, Varianzen und Kovarianzen der Orginalvariablen des Vektors $\mathbf{x}$ nicht bekannt. Sie können aber wie folgt aus den Schätzungen der Originalvariablen bestimmt werden:[54] Aus (7-17) ergibt sich der Zusammenhang zwischen der Kovarianzmatrix eines multiplikativ überlagerten Zufallsvektors und der Kovarianzmatrix des entsprechenden Vektors der "Originalvariablen. Dies schreiben wir hier wie folgt:

$$\begin{aligned} \widehat{cov[\mathbf{x^a}]} &= \left(\sigma_\varepsilon^2\mathbf{I} + \delta^2\,\boldsymbol{\iota\iota}'\right) \odot \left(\widehat{cov[\mathbf{x}]} + \widehat{\boldsymbol{\mu}_x}\widehat{\boldsymbol{\mu}_x}'\right) + \widehat{cov[\mathbf{x}]} \\ &= \left(\sigma_\varepsilon^2\mathbf{I} + \delta^2\,\boldsymbol{\iota\iota}' + \boldsymbol{\iota\iota}'\right) \odot \widehat{cov[\mathbf{x}]} + \left(\sigma_\varepsilon^2\mathbf{I} + \delta^2\,\boldsymbol{\iota\iota}'\right) \odot \widehat{\boldsymbol{\mu}_x}\widehat{\boldsymbol{\mu}_x}' \quad \text{(D-33)} \end{aligned}$$

wobei das Symbol $\widehat{\ }$ andeutet, daß es sich hier um Schätzungen handelt.

Bei Auflösung nach $cov[\mathbf{x}]$ ergibt sich daraus

$$\begin{aligned} \widehat{cov[\mathbf{x}]} &= \left\{\widehat{cov[\mathbf{x^a}]} - \left(\sigma_\varepsilon^2\mathbf{I} + \delta^2\,\boldsymbol{\iota\iota}'\right) \odot \widehat{\boldsymbol{\mu}_x}\widehat{\boldsymbol{\mu}_x}'\right\} \div \left(\sigma_\varepsilon^2\mathbf{I} + (1+\delta^2)\,\boldsymbol{\iota\iota}'\right) \\ &= \left\{\widehat{cov[\mathbf{x^a}]} - \left(\sigma_\varepsilon^2\mathbf{I} + \delta^2\,\boldsymbol{\iota\iota}'\right) \odot \widehat{\boldsymbol{\mu}_x}\widehat{\boldsymbol{\mu}_x}'\right\} \odot \left(\frac{1}{\sigma_\varepsilon^2 + 1 + \delta^2}\mathbf{I} + \frac{1}{1+\delta^2}\left(\boldsymbol{\iota\iota}' - \mathbf{I}\right)\right), \end{aligned} \quad \text{(D-34)}$$

wobei $\div$ in der ersten Zeile die Hadamard-Division bezeichnet. In der zweiten Zeile wurde das entsprechende Hadamard-**Produkt** gebildet.

Man beachte, daß sich daraus im Spezialfall nur eines einzigen Regressors x die Formel (10-42) aus Abschnitt 10.5 ergibt, d.h.

$$\widehat{\sigma_x^2} = \frac{s_{x^a}^2 - (\delta^2 + \sigma_\varepsilon^2)\,\overline{x}^{a\,2}}{1 + \delta^2 + \sigma_\varepsilon^2} \quad .$$

Dabei soll

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

und

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

gelten.

Im obigen multiplen Regressionsmodell sind außerdem die Kovarianzen aus $\widehat{cov[\mathbf{x^a}]}$ nach der üblichen Formel für empirische Momente zu bestimmen. Der unbekannte Vektor $\boldsymbol{\mu}_x$ sollte durch den entsprechenden Vektor der arithmetischen Mittel der anonymisierten Variablen geschätzt werden, d.h.

$$\widehat{\boldsymbol{\mu}_x} = \begin{pmatrix} \overline{x}_1^a \\ \overline{x}_2^a \\ \vdots \\ \overline{x}_{K-1}^a \\ \overline{x}_K^a \end{pmatrix} \quad .$$

Da bei Anonymisierung δ und σ_ε^2 bekannt sind, läßt sich der Korrekturschätzer (D-32) nun berechnen.

[53] Entsprechend läßt sich aus (D-30) der Korrekturschätzer für den Fall der Überlagerung nur der Regressoren bestimmen. Wir gehen darauf hier nicht weiter ein.

[54] Siehe auch die entsprechenden Ergebnisse für den Fall eines einzigen Regressors in Abschnitt 10.5.

E Beweise zu Abschnitt 10.6.2 (Überlagerung allgemein im Panelfall)

Wir betrachten

$$x^a_{it} \;=\; x_{it}\, u_{itx} \quad \text{und} \quad y^a_{it} \;=\; y_{it}\, u_{ity}$$

mit

$$V[u_{itx}] \;=\; \gamma_x^2 \quad \text{und} \quad V[u_{ity}] \;=\; \gamma_y^2 \quad \text{sowie} \quad COV[u_x\, u_y \;=\; \gamma_{xy}$$

und erhalten

$$x^a_{it} \;-\; \overline{x}^a_{i\bullet} \quad = \quad x_{it}\, u_{itx} \;-\; \overline{xu}_{i\bullet x}$$

sowie

$$y^a_{it} \;-\; \overline{y}^a_{i\bullet} \quad = \quad (\alpha \;+\; \tau_i)\,(u_{ity} \;-\; \overline{u}_{i\bullet y}) \;+\; \beta\,(x_{it}\, u_{ity} \;-\; \overline{xu}_{i\bullet y}) \;+\; (\eta_{it}\, u_{ity} \;-\; \overline{\eta u}_{i\bullet y})$$

Man beachte, daß für den Fall $\gamma_{xy} \neq 0$ die beiden Überlagerungsvariablen u_{itx} und u_{ity} miteinander korreliert sind. Im folgenden soll außerdem zugelassen sein, daß x und τ korreliert sind (Nichtexogenität des Regressors). Die entsprechende Kovarianz bezeichnen wir mit $\sigma_{\tau x}$.

Zur Bestimmung des Wahrscheinlichkeitsgrenzwert von

$$\hat{\beta}^a_W \quad = \quad \frac{\frac{1}{T}\,\frac{1}{n}\,\sum_{t=1}^{T}\sum_{i=1}^{n}(x^a_{it} - \overline{x}^a{}_{i\bullet})(y^a_{it} - \overline{y^a}_{i\bullet})}{\frac{1}{T}\,\frac{1}{n}\,\sum_{t=1}^{T}\sum_{i=1}^{n}(x^a_{it} - \overline{x}^a{}_{i\bullet})^2}$$

benötigen wir insbesondere die relevanten Ausdrücke für den Zähler, die im folgenden abgeleitet werden.[55] Dabei wird der Faktor $1/T$ sowie die Summation über alle t unterdrückt, weil der Grenzwert für $n \to \infty$ betrachtet wird.

$$\text{plim}_{n\to\infty}\,\frac{1}{n}\,\sum_{i=1}^{n}\,\alpha\,(x_{it}\, u_{itx} \;-\; \overline{xu}_{i\bullet x})\,(\,(u_{ity} \;-\; \overline{u}_{i\bullet y}) \quad = \quad \alpha\,\mu_x\,\gamma_{xy} \tag{E-1}$$

Beweis: Der Ausdruck stellt ein Stichprobenmment dar und konvergiert gegen das entsprechende theoretische Moment, das folgende Form hat:

$$\begin{aligned}
\alpha\,E[(X\,U_x \;-\; E[X\,U_x])\,(U_y \;-\; E[U_y])] \;&=\; \alpha\,E[(X\,U_x \;-\; \mu_x)\,(U_y \;-\; 1)] \\
&=\; \alpha\,E[X\,U_x\,U_y \;-\; \mu_x\,U_y \;+\; \mu_x \;-\; X\,U_y] \\
&=\; \alpha\,(\mu_x\,(\gamma_{xy} \;+\; 1) \;-\; \mu_x \;+\; \mu_x \;-\; \mu_x) \\
&=\; \alpha\,\mu_x\,\gamma_{xy}
\end{aligned}$$

$$\text{plim}_{n\to\infty}\,\frac{1}{n}\,\sum_{i=1}^{n}\,\tau_i\,(x_{it}\, u_{itx} \;-\; \overline{xu}_{i\bullet x})\,(u_{ity} \;-\; \overline{u}_{i\bullet y}) \quad = \quad \sigma_{\tau x}\gamma_{xy} \tag{E-2}$$

Beweis: Das entsprechende theoretische Moment lautet in diesem Fall

$$\begin{aligned}
E[\tau\,(X\,U_x \;-\; E[X\,U_x])\,(U_y \;-\; E[U_y])] \;&=\; E_{\tau,X}[E_{U_xU_y}[\tau\,X\,(U_x \;-\; 1)\,(U_y \;-\; 1)|\,\tau\,X]\,] \\
&=\; E_{\tau,X}[\tau\,X\,\gamma_{xy}] \\
&=\; \sigma_{\tau x}\gamma_{xy}
\end{aligned}$$

[55] Der entsprechende Grenzwert für den Nenner wurde bereits in Abschnitt 10.6.2 bestimmt.

$$\text{plim}_{n\to\infty}\,\frac{1}{n}\sum_{i=1}^{n}\beta\,(x_{it}\,u_{itx}-\overline{xu}_{i\bullet x})\,(x_{it}\,u_{ity}-\overline{xu}_{i\bullet y}) \quad = \quad \beta(\sigma_x^2+\gamma_{xy}(\mu_x^2+\sigma_x^2)) \qquad \text{(E-3)}$$

Beweis: In diesem Fall lautet das theoretische Moment

$$\begin{aligned}\beta\,E[(X\,U_x\,-\,E[X\,U_x])\,(X\,U_y\,-\,E[X\,U_y])] &= \beta\,E[(X\,U_x\,-\,\mu_x)\,(X\,U_y\,-\,\mu_x)]\\ &= \beta\,E[X^2\,U_x\,U_y\,-\,\mu_x\,X\,U_x\,-\,\mu_x\,X\,U_y\,+\,\mu_x^2]\\ &= \beta\,((\mu_x^2\,+\,\sigma_x^2)\,(\gamma_{xy}\,+\,1)\,-\,\mu_x^2\,-\,\mu_x^2\,+\,\mu_x^2)\\ &= \beta\,(\mu_x^2\,+\,\sigma_x^2\,+\,(\mu_x^2\,+\,\sigma_x^2)\,\gamma_{xy}\,-\,\mu_x^2)\\ &= \beta(\sigma_x^2\,+\,\gamma_{xy}(\mu_x^2\,+\,\sigma_x^2))\end{aligned}$$

$$\text{plim}_{n\to\infty}\,\frac{1}{n}\sum_{i=1}^{n}(x_{it}\,u_{itx}\,-\,\overline{xu}_{i\bullet x})\,(\eta_{it}\,u_{ity}\,-\,\overline{\eta u}_{i\bullet y}) \quad = \quad 0 \qquad \text{(E-4)}$$

Beweis: In diesem Fall folgt das Resultat unmittelbar aus der Tatsache, daß

$$E[\eta\,|\,X\,,\,U_x\,,\,U_y] \quad = \quad 0$$

gilt.

F 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

$$\mathbf{A}\odot\mathbf{B} \quad = \quad \begin{pmatrix} a_{11}b_{11} & a_{12}b_{12} & \dots & a_{1,n-1}b_{1,n-1} & a_{1n}b_{1n}\\ a_{21}b_{21} & a_{22}b_{22} & \dots & a_{2,n-1}b_{2,n-1} & a_{2n}b_{2n}\\ \vdots & \vdots & \vdots & \vdots & \vdots\\ a_{m-1,1}b_{m-1,1} & a_{m-1,2}b_{m-1,2} & \dots & a_{m-1,,n-1}b_{m-1,n-1} & a_{m-1,n}b_{m-1,n}\\ a_{m,1}b_{m,1} & a_{m,2}b_{m,2} & \dots & a_{m,,n-1}b_{m,n-1} & a_{m,n}b_{m,n}\end{pmatrix}\quad , \qquad \text{(F-1)}$$

d.h. die resultierende Matrix hat ebenfalls m Zeilen und n Spalten. Mit $\div$ wird die entsprechende elementweise **Division** bezeichnet. Insbesondere gilt für beliebige Rechtecksmatriz

$$\mathbf{A}\,\div\,\mathbf{A} \quad = \quad \begin{pmatrix} 1 & 1 & \dots & 1 & 1\\ 1 & 1 & \dots & 1 & 1\\ \vdots & \vdots & \vdots & \vdots & \vdots\\ 1 & 1 & \dots & 1 & 1\\ 1 & 1 & \dots & 1 & 1\end{pmatrix} \quad = \boldsymbol{\iota}\,\boldsymbol{\iota}' \quad \text{und} \quad \mathbf{A}\,\odot\,\boldsymbol{\iota}\,\boldsymbol{\iota}' \quad = \quad \mathbf{A} \qquad , \qquad \text{(F-2)}$$

und für beliebigen Vektor $\mathbf{a}$ gilt

$$\mathbf{a}\,\div\,\mathbf{a} \quad = \quad \begin{pmatrix} 1\\ 1\\ \vdots\\ 1\\ 1\end{pmatrix} \quad = \quad \boldsymbol{\iota} \quad \text{und} \quad \mathbf{a}\,\odot\,\boldsymbol{\iota} \quad = \quad \mathbf{a} \qquad . \qquad \text{(F-3)}$$

Unmittelbar aus (F-1) folgt[56]

$$vec(\mathbf{A} \odot \mathbf{B}) \quad = \quad vec(\mathbf{A}) \odot vec(\mathbf{B}) \tag{F-4}$$

Für m-dimensionale Vektoren $\mathbf{a}$, $\mathbf{c}$ und n-dimensionale Vektoren $\mathbf{b}, \mathbf{d}$ sind manchmal folgende Ergebnisse nützlich:

$$\mathbf{a} \odot \mathbf{b} \quad = \quad \mathbf{D}_a \, \mathbf{b} \quad = \quad \mathbf{D}_b \, \mathbf{a} \tag{F-5}$$

Dabei sind $\mathbf{D}_a$ und $\mathbf{D}_b$ Diagonalmatrizen mit den Elementen des jeweiligen Vektors auf der Diagonalen. Ferner gilt

$$(\mathbf{a} \odot \mathbf{b}) \, (\mathbf{c} \odot \mathbf{d})' \quad = \quad \mathbf{a} \, \mathbf{c}' \odot \mathbf{b} \, \mathbf{d}' \tag{F-6}$$

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

$$\mathbf{A}' \mathbf{A} \quad = \quad \sum_{j=1}^{n} \left[(\mathbf{I}_n \otimes \mathbf{e}_j) \; vec(\mathbf{A}') \right] \left[(\mathbf{I}_n \otimes \mathbf{e}_j) \; vec(\mathbf{A}') \right]' \tag{F-7}$$

Dabei ist $\mathbf{e}_j$ ein n-dimensionaler Vektor und bezeichnet die j-te Spalte der Einheitsmatrix

Beweis: Zunächst schreiben wir die Matrix $\mathbf{A}$ wie folgt:

$$\mathbf{A} \; = \; \begin{pmatrix} \mathbf{a}_1' \\ \mathbf{a}_2' \\ \vdots \\ \mathbf{a}_{m-1}' \\ \mathbf{a}_m' \end{pmatrix} \qquad \text{mit} \quad \mathbf{a}_j' \; = \; (a_{j1} \, , \, \ldots \, , \, a_{jn})$$

und damit

$$\mathbf{A}' \mathbf{A} \quad = \quad \sum_{j=1}^{m} \mathbf{a}_j \mathbf{a}_j'$$

Ferner gilt

$$\mathbf{A}' \quad = \quad \left(\; \mathbf{a}_1 \, , \, \mathbf{a}_2 \, \ldots \, \mathbf{a}_{m-1} \, , \, \mathbf{a}_m \; \right) \quad \text{und} \quad vec(\mathbf{A}') \quad = \quad \begin{pmatrix} \mathbf{a}_1 \\ \vdots \\ \vdots \\ \mathbf{a}_m \end{pmatrix}$$

sowie

$$\left(\mathbf{I}_n \otimes \mathbf{e}_j' \right) \; vec(\mathbf{A}') \quad = \left(\; \mathbf{0} \, , \, \ldots \mathbf{I_n} \ldots \, , \, \mathbf{0} \; \right) \begin{pmatrix} \mathbf{a}_1 \\ \vdots \\ \mathbf{a}_j \\ \vdots \\ \mathbf{a}_m \end{pmatrix} \quad = \quad \mathbf{a}_j \qquad .$$

[56] Der vec-Operator stapelt die Spalten einer Matrix übereinander. $vec(\mathbf{A})$ ergibt einen mn-dimensionalen Vektor.

Daraus folgt (F-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:

$$\mathbf{A}\,\mathbf{x} \quad = \quad \left(\mathbf{I}_m \otimes \mathbf{x}'\right)\, vec(\mathbf{A}') \tag{F-8}$$

Beweis: Zunächst gilt

$$\mathbf{A}\,\mathbf{x} \quad = \quad \begin{pmatrix} \mathbf{a}_1'\,\mathbf{x} \\ \vdots \\ \mathbf{a}_m'\,\mathbf{x} \end{pmatrix}$$

$$\mathbf{I}_m \otimes \mathbf{x}' \quad = \quad \begin{pmatrix} \mathbf{x}' & & & & \\ & \mathbf{x}' & & & \\ & & \ddots & & \\ & & & \mathbf{x}' & \\ & & & & \mathbf{x}' \end{pmatrix}$$

und damit

$$\left(\mathbf{I}_m \otimes \mathbf{x}'\right)\, vec(\mathbf{A}') \quad = \quad \begin{pmatrix} \mathbf{x}'\,\mathbf{a}_1 \\ \vdots \\ \mathbf{x}'\,\mathbf{a}_m \end{pmatrix} \quad ,$$

womit (F-8) bewiesen ist.

IAW-Diskussionspapiere

Bisher erschienen:

Nr. 1 (September 2001)
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 (Dezember 2001)
Die Einkommensteuerreform 1990 als natürliches Experiment. Methodische und konzeptionelle Aspekte zur Schätzung der Elastizität des zu versteuernden Einkommens
Peter Gottfried / Hannes Schellhorn

Nr. 3 (Januar 2001)
Gut betreut in den Arbeitsmarkt? Eine mikroökonomische Evaluation der Mannheimer Arbeitsvermittlungsagentur
Jürgen Jerger / Christian Pohnke / Alexander Spermann

Nr. 4 (Dezember 2001)
Das IAW-Einkommenspanel und das Mikrosimulationsmodell SIMST
Peter Gottfried / Hannes Schellhorn

Nr. 5 (April 2002)
A Microeconometric Characterisation of Household Consumption Using Quantile Regression
Niels Schulze / Gerd Ronning

Nr. 6 (April 2002)
Determinanten des Überlebens von Neugründungen in der baden-württembergischen Industrie – eine empirische Survivalanalyse mit amtlichen Betriebsdaten
Harald Strotmann

Nr. 7 (November 2002)
Die Baulandausweisungsumlage als ökonomisches Steuerungsinstrument einer nachhaltigkeitsorientierten Flächenpolitik
Raimund Krumm

Nr. 8 (März 2003)
Making Work Pay: U.S. American Models for a German Context?
Laura Chadwick, Jürgen Volkert

IAW-Diskussionspapiere

Nr. 9 (Juni 2003)
Erste Ergebnisse von vergleichenden Untersuchungen mit anonymisierten und nicht anonymisierten Einzeldaten am Beispiel der Kostenstrukturerhebung und der Umsatzsteuerstatistik
Martin Rosemann

Nr. 10 (August 2003)
Randomized Response and the Binary Probit Model
Gerd Ronning

Nr. 11 (August 2003)
Creating Firms for a New Century: Determinants of Firm Creation around 1900
Joerg Baten

Nr. 12 (September 2003)
Das fiskalische BLAU-Konzept zur Begrenzung des Siedlungs-flächenwachstums
Raimund Krumm

Nr. 13 (Dezember 2003)
Generelle Nichtdiskontierung als Bedingung für eine nachhaltige Entwicklung?
Stefan Bayer

Nr. 14 (Februar 2003)
Die Elastizität des zu versteuernden Einkommens. Messung und erste Ergebnisse zur empirischen Evidenz für die Bundesrepublik Deutschland.
Peter Gottfried / Hannes Schellhorn

Nr. 15 (Februar 2004)
Empirical Evidence on the Effects of Marginal Tax Rates on Income – The German Case
Peter Gottfried / Hannes Schellhorn

Nr. 16 (Juli 2004)
Shadow Economies around the World: What do we really know?
Friedrich Schneider

IAW-Diskussionspapiere

Nr. 17 (August 2004)
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 (Januar 2005)
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 (März 2005)
Legislative Malapportionment and the Politicization of Germany's Intergovernmental Transfer System
Hans Pitlik / Friedrich Schneider / Harald Strotmann

Nr. 20 (April 2005)
Implementation ökonomischer Steuerungsansätze in die Raumplanung
Raimund Krumm

Nr. 21 (Juli 2005)
Determinants of Innovative Activity in Newly Founded Knowledge Intensive Business Service Firms
Andreas Koch / Harald Strotmann

Nr. 22 (Dezember 2005)
Impact of Opening Clauses on Bargained Wages
Wolf Dieter Heinbach

Nr. 23 (Januar 2006)
Hat die Einführung von Gewinnbeteiligungsmodellen kurzfristige positive Produktivitätswirkungen? – Ergebnisse eines Propensity-Score-Matching-Ansatzes
Harald Strotmann

Nr. 24 (März 2006)
Who Goes East? The Impact of Enlargement on the Pattern of German FDI
Claudia M. Buch / Jörn Kleinert

Nr. 25 (Mai 2006)
Estimation of the Probit Model from Anonymized Micro Data
Gerd Ronning / Martin Rosemann

IAW-Diskussionspapiere

Nr. 26 (Oktober 2006)
Bargained Wages in Decentralized Wage-Setting Regimes
Wolf Dieter Heinbach

Nr. 27 (Januar 2007)
A Capability Approach for Official German Poverty and Wealth Reports: Conceptual Background and First Empirical Results
Christian Arndt / Jürgen Volkert

Nr. 28 (Februar 2007)
Typisierung der Tarifvertragslandschaft – Eine Clusteranalyse der tarifvertraglichen Öffnungsklauseln
Wolf Dieter Heinbach / Stefanie Schröpfer

Nr. 29 (März 2007)
International Bank Portfolios: Short- and Long-Run Responses to the Business Cycles
Sven Blank / Claudia M. Buch

Nr. 30 (April / revidiert Oktober 2007)
Stochastische Überlagerungen mit Hilfe der Mischungsverteilung
Gerd Ronning

Nr. 31 (Mai 2007)
Openness and Growth: The Long Shadow of the Berlin Wall
Claudia M. Buch / Farid Toubal

Nr. 32 (Mai 2007)
International Banking and the Allocation of Risk
Claudia M. Buch / Gayle DeLong / Katja Neugebauer

Nr. 33 (Juli 2007)
Multinational Firms and New Protectionisms
Claudia M. Buch / Jörn Kleinert