Biewen, Elena

Working Paper

Within-Schätzung bei anonymisierten Paneldaten

IAW Diskussionspapiere, No. 34

Provided in Cooperation with:
Institute for Applied Economic Research (IAW)

Suggested Citation: Biewen, Elena (2007) : Within-Schätzung bei anonymisierten Paneldaten, IAW Diskussionspapiere, No. 34, Institut für Angewandte Wirtschaftsforschung (IAW), Tübingen

This Version is available at:
https://hdl.handle.net/10419/36625

IAW-Diskussionspapiere

Discussion Paper

34

Within-Schätzung bei anonymisierten Paneldaten

Elena Biewen

November 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

---

Within-Schätzung bei anonymisierten Paneldaten

Elena Biewen*

Zusammenfassung

Dieser Beitrag befasst sich mit der Untersuchung der Auswirkungen der datenverändernden Anonymisierungsverfahren 'variablenspezifische abstandsorientierte Mikroaggregation' (vaabMA, im Englischen auch als Individual Ranking bezeichnet) und 'multiplikative stochastische Überlagerung' auf die 'Within'-Schätzung eines linearen Panelmodells mit Individualeffekten. Es wird gezeigt, dass der 'Within'-Schätzer auf der Grundlage der mittels vaabMA anonymisierten Daten konsistent bleibt. Bei der multiplikativen stochastischen Überlagerung wird neben der allgemeinen Form der Überlagerung eine spezielle Variante analysiert, bei der die Variablen zuerst mit einem konstantem Grundüberlagerungsfaktor und danach zusätzlich additiv überlagert werden. Es wird weiterhin gezeigt, dass beide Varianten der Überlagerung zu Inkonsistenz der 'Within'-Schätzer führen. Anschließend werden korrigierte Schätzer hergeleitet.

1 Einleitung

Dieser Beitrag entstand im Rahmen des vom Bundesministerium für Bildung und Forschung (BMBF) geförderten Projektes "Wirtschaftsstatistische Paneldaten und faktische Anonymisierung" und beschäftigt sich mit zwei datenverändernden Verfahren, die für die Anonymisierung von Paneldaten verwendet werden: variablenspezifische abstandsorientierte Mikroaggregation (fortan vaabMA) und multiplikative stochastische Überlagerung. Das Ziel ist die Analyse des Einflusses der beiden Anonymisierungsstrategien auf die 'Within'-Schätzung eines linearen Panelmodells mit Individualeffekten.

Die vaabMA ist eine Variante der getrennten Mikroaggregation, bei der die zu Gruppen zusammengefassten Beobachtungen durch die Gruppenmittelwerte ersetzt werden. Die Mikroaggregation bei Paneldaten wird getrennt sowohl für alle Merkmale als auch für alle Zeitpunkte ausgeführt. Andere Varianten der Mikroaggregation, allerdings für Querschnittsdaten, werden beispielsweise bei Schmid et al. (2005), Schmid (2006b) behandelt.

*Elena Biewen, Institut für Angewandte Wirtschaftsforschung, Ob dem Himmelreich 1, 72074 Tübingen, Deutschland. Tel.: (07071) 98 96-36, Fax -99, E-Mail: elena.biewen@iaw.edu. Ich danke Gerd Ronning und Martin Rosemann für viele sehr hilfreiche Hinweise.

Bei der stochastischen Überlagerung wird ein stochastisch erzeugter Fehler zu den Originaldaten hinzugefügt. Grundsätzlich unterscheidet man zwischen der additiven und der multiplikativen Überlagerung. Die multiplikative Überlagerung hat den Vorteil, dass sie alle Originalwerte proportional verändert, so dass für die Untersuchungseinheiten ein besserer Schutz gewährleistet ist. Außerdem bleiben die Nullen im Datensatz erhalten (Rosemann 2006). Aus diesen Gründen betrachtet die vorliegende Arbeit ausschließlich den multiplikativen Fall. Während der additive Fehler in der Literatur eingehend untersucht wurde, findet man zur multiplikativen Variante noch wenig. Stefanski (1985) leitet unter anderem einen korrigierten KQ-Schätzer für den multiplikativen Messfehler her. Iturria et al. (1999) schlagen einen konsistenten M-Schätzer für ein polynomiales Regressionsmodell vor. Lechner (2007), Ronning et al. (2005) und Rosemann (2006) untersuchen die SIMEX-Methode bei Vorliegen der multiplikativ überlagerten Daten.

In diesem Aufsatz wird zuerst die allgemeine Form der multiplikativen Überlagerung und anschließend eine spezielle Variante untersucht, bei der die Variablen zuerst mit einem konstanten Grundüberlagerungsfaktor multipliziert und danach zusätzlich additiv überlagert werden. Das letztere Verfahren schlug Höhne (2004) mit dem Ziel vor, die Verhältnisse zwischen unterschiedlichen Variablen nach der Anonymisierung zu erhalten (Ronning 2007).

Der Beitrag ist wie folgt gegliedert. Im Kapitel 2 wird das zu schätzende lineare Panelmodell definiert. Kapitel 3 untersucht nach einer kurzen Erläuterung des Verfahrens vaabMA die Auswirkungen desselben auf die 'Within'-Schätzung. Dies geschieht sowohl theoretisch als auch mithilfe eines Simulationsexperimentes. Kapitel 4 beschäftigt sich mit der Analyse der beiden Varianten der multiplikativen stochastischen Überlagerung. Eine Zusammenfassung der wichtigsten Ergebnisse schließt den Beitrag.

2 Das Modell

Das lineare Panelmodell mit Individualeffekten ist definiert als (Mundlak 1978)

$$\underset{(NT\times 1)}{\boldsymbol{Y}} = \underset{(NT\times N)(N\times 1)}{\boldsymbol{D}\,\boldsymbol{\alpha}} + \underset{(NT\times K)(K\times 1)}{\boldsymbol{X}\,\boldsymbol{\beta}} + \boldsymbol{\epsilon}, \tag{2-1}$$

$\boldsymbol{\alpha}$ enthält N Individualeffekte, die für jede Untersuchungseinheit nicht über die Zeit variieren, je nach Individuum jedoch verschieden sind. α's werden als stochastische Effekte mit Erwartungswert μ_α und Varianz σ_α^2 behandelt. Die Korrelation zwischen Regressoren und Individualeffekten ist daher möglich. $\boldsymbol{D}$ ist definiert als $\boldsymbol{D} = \boldsymbol{I_N} \otimes \boldsymbol{\iota_T}$, wobei $\boldsymbol{I_N}$ eine Einheitsmatrix der Dimension $(N \times N)$ und $\boldsymbol{\iota_T}$ ein T-dimensionaler Spaltenvektor mit Einsen sind. $\boldsymbol{X}$ enthält die K-Regressoren und $\boldsymbol{\beta}$ ist der Vektor mit entsprechenden Koeffizienten. Der Störterm des Modells ϵ_{it} hat die Eigenschaft:

$$\epsilon_{it} \sim iid\,(0, \sigma_\epsilon^2), \qquad i = 1, \ldots, N, \quad t = 1, \ldots, T \tag{2-2}$$

Multiplikation der beiden Seiten der Gleichung (2-1) mit der Matrix $\boldsymbol{Q} = \boldsymbol{I_{NT}} - \boldsymbol{D}(\boldsymbol{D'D})^{-1}\boldsymbol{D'}$ bewirkt das Abziehen der Mittelwerte über die Zeit von den Beobachtungen und führt damit zur Eliminierung der Individualeffekte ('Within'-Transformation).[1]

Der 'Within'-Schätzer[2] wird geschrieben als (Hsiao 2003)

$$\widehat{\boldsymbol{\beta_w}} = \left[\boldsymbol{X'QX}\right]^{-1} \boldsymbol{X'QY} \tag{2-3}$$

Der naive Schätzer, d.h. der Schätzer unter Verwendung der anonymisierten Daten, ist dann[3]

$$\widehat{\boldsymbol{\beta_w^a}} = \left[\boldsymbol{X^{a\prime}QX^a}\right]^{-1} \boldsymbol{X^{a\prime}QY^a} \tag{2-4}$$

bzw.

$$\widehat{\boldsymbol{\beta_w^a}} = \left[\sum_i^N \sum_t^T \left(\boldsymbol{x}_{it}^a - \bar{\boldsymbol{x}}_i^a\right)\left(\boldsymbol{x}_{it}^a - \bar{\boldsymbol{x}}_i^a\right)'\right]^{-1} \sum_i^N \sum_t^T (\boldsymbol{x}_{it}^a - \bar{\boldsymbol{x}}_i^a)(y_{it}^a - \bar{y}_i^a) \tag{2-5}$$

mit

$$\bar{\boldsymbol{x}}_i^a = \frac{1}{T}\sum_t^T \boldsymbol{x}_{it}^a, \qquad \bar{y}_i^a = \frac{1}{T}\sum_t^T y_{it}^a$$

Aus (2-5) wird deutlich, dass der Schätzer mithilfe der empirischen 'Within'-Varianz und 'Within'-Kovarianz ausgedrückt werden kann (Greene 2000, S.563-564):

$$\widehat{\boldsymbol{\beta_w^a}} = \left(\boldsymbol{S}_{x^a x^a}^w\right)^{-1} \boldsymbol{S}_{x^a y^a}^w$$

$$= \begin{pmatrix} S_{x_1^a}^{w\ 2} & S_{x_1^a x_2^a}^w & \cdots & S_{x_1^a x_K^a}^w \\ \vdots & & & \vdots \\ S_{x_K^a x_1^a}^w & \cdots & & S_{x_K^a}^{w\ 2} \end{pmatrix}^{-1} \begin{pmatrix} S_{x_1^a y^a}^w \\ \vdots \\ S_{x_K^a y^a}^w \end{pmatrix} \tag{2-6}$$

[1] Nicht nur Individualeffekte, sondern auch andere zeitkonstante Variablen werden durch diese Transformation gelöscht und ihr Effekt ist dann nicht mehr schätzbar. $\boldsymbol{\alpha}$ könnte man jedoch mittels einer Teilschätzung schätzen (siehe z.B. Greene 2000, Ronning 2007 (Abschnitt D.2)). Ein weiteres Schätzverfahren der Panelökonometrie, die sogenannte Random-Effects-Schätzung, erlaubt die Schätzung der zeitkonstanten Variablen, unterstellt jedoch die Annahme der Unkorreliertheit der Regressoren und der individuellen Effekte.

[2] Die Bezeichnung 'Within' kommt daher, dass dieser Schätzer nur die Variation innerhalb der Gruppen ausnutzt (Hsiao 203, S.33).

[3] Mit dem Index a wird im Folgenden die anonymisierte Variable bzw. der naive Schätzer bezeichnet.

$\widehat{\boldsymbol{\beta}_{\boldsymbol{w}}^{\boldsymbol{a}}}$ ist dann konsistent, wenn die einzelnen 'Within'-Varianzen und 'Within'-Kovarianzen in (2-6) konsistent sind, genauer gesagt, wenn sie in Wahrscheinlichkeit gegen die entsprechenden wahren theoretischen Momente konvergieren.[4]

Im Folgenden wird untersucht, wie die Anonymisierung die Konsistenz der 'Within'-Schätzer beeinflusst. Dabei wird nur der Einfluss auf $\boldsymbol{\beta}$ untersucht.

3 Variablenspezifische abstandsorientierte Mikroaggregation

3.1 Das Verfahren

Mit der vaabMA werden die Variablen wie folgt anonymisiert (Schmid 2006a). Den Datensatz sortiert man zuerst nach der ersten Variable, es kann sowohl x als auch y sein. Nachdem die Anzahl der Elemente je Gruppe (A) festgelegt ist, werden alle Beobachtungen dieser ersten Variable in Gruppen zusammengefasst und ihre Werte durch die Gruppendurchschnitte ersetzt. Die anderen Variablen bleiben dabei unverändert. Die gleiche Prozedur wird anschließend mit der zweiten, dritten u.s.w. Variable durchgefürt. Im Fall von Paneldaten wird die Mikroaggregation auf jeden Zeitpunkt getrennt angewandt.

Wir haben beispielsweise zwei Variablen, die zu zwei Zeitpunkten beobachtet werden (Tabelle a). Zuerst wird nach der ersten Variable (z.B. x) im Zeitpunkt $t = 1$ sortiert. Danach wird diese Variable mikroaggregiert (Tabelle b). Im nächsten Schritt erfolgt die Sortierung nach der zweiten Variable (y) in $t = 1$, und wiederum wird nur diese Variable mirkroaggregiert (Tabelle c).

a

t=1		t=2	
x	y	x	y
0,5	0,3	0,6	5,4
0,9	0,2	2,3	1,2
0,7	0,7	4,2	3,2
1,4	1,1	0,2	1,5
1,3	0,6	2,2	0,3
0,3	0,1	0,7	3,1

b

t=1		t=2	
$\boldsymbol{x}$	y	x	y
0,5	0,1	0,7	3,1
0,5	0,3	0,6	5,4
0,5	0,7	4,2	3,2
1,2	0,2	2,3	1,2
1,2	0,6	2,2	0,3
1,2	1,1	0,2	1,5

c

t=1		t=2	
x	$\boldsymbol{y}$	x	y
0,5	0,2	0,7	3,1
1,2	0,2	2,3	1,2
0,5	0,2	0,6	5,4
1,2	0,8	2,2	0,3
0,5	0,8	4,2	3,2
1,2	0,8	0,2	1,5

Die gleiche Vorgehensweise wird weiter für den zweiten Zeitpunkt vorgenommen. Anschließend erhalten wir den Datensatz, in dem alle Variablen mittels vaabMA anonymisiert wurden (Tabelle d).

[4]Da bei den Datensätzen, die im Projekt untersucht werden, die Anzahl der Unternehmen sehr groß, die Anzahl der Perioden hingegen gering ist, wird in diesem Beitrag der Wahrscheinlichkeitsgrenzwert für $N \to \infty$ untersucht.

d

t=1		t=2	
x	y	x	y
1,2	0,8	2,9	1,0
1,2	0,2	2,9	1,0
1,2	0,8	0,5	1,0
0,5	0,2	0,5	3,9
0,5	0,8	2,9	3,9
0,5	0,2	0,5	3,9

3.2 Auswirkungen auf die 'Within'-Schätzung

Dieses Kapitel befasst sich mit den Auswirkungen von vaabMA auf die 'Within'-Schätzung. Die folgenden Darstellungen basieren auf den Ausführungen von Matthias Schmid (Schmid 2006a), der die Konsistenz des Kleinstquadrate-Schätzers im Fall der Querschnittsdaten bewiesen hat.

Mit $S^w_{x^a_k x^a_l}$ wird weiter die empirische 'Within'-Kovarianz zwischen zwei mikroaggregierten Regressoren x^a_k und x^a_l bezeichnet.[5] $S^{w\,2}_{x^a_k}$ ist die Varianz des k-ten anonymisierten Regressors.

Die 'Within'-Momente kann man umschreiben als[6] [7]

$$S^{w\,2}_{x^a_k} = \frac{1}{T}\sum_t \left(S^2_{x^a_{tk}} + (\bar{x}^a_{tk})^2\right) - \frac{1}{T^2}\sum_t\sum_s \left(S_{x^a_{tk}x^a_{sk}} + \bar{x}^a_{tk}\bar{x}^a_{sk}\right) \qquad (3\text{-}7)$$

$$S^w_{x^a_k x^a_l} = \frac{1}{T}\sum_t \left(S_{x^a_{tk}x^a_{tl}} + \bar{x}^a_{tk}\bar{x}^a_{tl}\right) - \frac{1}{T^2}\sum_t\sum_s \left(S_{x^a_{tk}x^a_{sl}} + \bar{x}^a_{tk}\bar{x}^a_{sl}\right) \qquad (3\text{-}8)$$

$$S^w_{x^a_k y^a} = \frac{1}{T}\sum_t \left(S_{x^a_{tk}y^a_t} + \bar{x}^a_{tk}\bar{y}^a_t\right) - \frac{1}{T^2}\sum_t\sum_s \left(S_{x^a_{tk}y^a_s} + \bar{x}^a_{tk}\bar{y}^a_s\right) \qquad (3\text{-}9)$$

$$t,s = 1,\ldots,T, \quad k = 1,\ldots,K, \quad l = 1,\ldots,L$$

Die 'Within'-Kovarianzen/-Varianzen lassen sich durch die Gesamtvarianzen/-kovarianzen und Mittelwerte in den einzelnen Zeitpunkten t und s darstellen.

Die Mittelwerte bleiben nach der Durchführung der vaabMA erhalten. Für die Kovarianzen/Varianzen lässt sich der Beweis von Schmid (2006a, Lemma 1 und 2) auf den Fall von Paneldaten übertragen. Bei Gültigkeit der Annahmen von Schmid gilt für jedes feste Paar (t,s) asymptotisch:

[5] Analog für $S^w_{x^a_k y^a}$

[6] Herleitung im Appendix A

[7] Ich danke Hans Schneeweiß für wichtige Hinweise bei der Herleitung der Formel.

$$\underset{N\to\infty}{plim}\ S_{x^a_{tk}x^a_{sk}} = \sigma_{x_{tk}x_{sk}} \tag{3-10}$$

$$\underset{N\to\infty}{plim}\ S_{x^a_{tk}x^a_{sl}} = \sigma_{x_{tk}x_{sl}} \tag{3-11}$$

$$\underset{N\to\infty}{plim}\ S_{x^a_{tk}y^a_{s}} = \sigma_{x_{tk}y_{s}} \tag{3-12}$$

$$t, s = 1, \ldots, T, \quad k = 1, \ldots, K, \quad l = 1, \ldots, L$$

Daraus folgt, dass (3-7)-(3-9) ebenfalls konsistent bleiben. Folglich konvergieren $\boldsymbol{S}^w_{x^a x^a}$, $\boldsymbol{S}^w_{x^a y^a}$ in Wahrscheinlichkeit gegen die wahren $\boldsymbol{\Sigma}^w_{xx}$, $\boldsymbol{\Sigma}^w_{xy}$.

Als ein wichtiges Ergebnis lässt sich somit festhalten, dass der 'Within'-Schätzer auf der Grundlage der mit vaabMA anonymisierten Daten konsistent bleibt. Dabei hat der geführte Beweis keine Unabhängigkeitsannahme von $\boldsymbol{X}$ unterstellt. Die Konsistenz gilt auch dann, wenn die Regressoren über die Zeit korreliert sind.

Simulationsergebnisse

Die Auswirkungen von vaabMA auf die 'Within'-Schätzung wurden anhand einer Monte-Carlo-Studie überprüft. Es wurde das lineare Panelmodell (2-1) mit zwei Regressoren geschätzt. Die Regressoren erfüllen die iid-Annahme und folgen der Lognormalverteilung mit

$$X_1 \sim L\left(4,35; 1,75^2\right), \quad X_2 \sim L\left(3,45; 1,4^2\right)$$

Der Störterm des Modells ϵ_{it} ist standardnormalverteilt. Der Individualeffekt α ist mit dem zweiten Regressor korreliert und wird nach der Formel von Biørn (1996) generiert:

$$\alpha_i = \left(\bar{x}_{i,2} - E\left[x_2\right]\right)\lambda + \epsilon_{\alpha,i} \qquad i = 1, \ldots, N$$

mit $\bar{x}_{i,2} = \frac{1}{T}\sum_t x_{it2}$ (Mittelwert des zweiten Regressors über die Zeit für das i-te Unternehmen), $\lambda = 1$, $\epsilon_{\alpha,i} \sim N\left(0,1\right)$.

In der vorliegenden Studie wird mit $N = 1035$ Beobachtungen und $t = 4$ Zeitpunkten gearbeitet. Sowohl die abhängige Variable als auch alle Regressoren werden mittels vaabMA anonymisiert. Die Gruppen werden mit $A = 3$ und 5 Elementen gebildet. Die Anzahl der Monte-Carlo-Replikationen beträgt 500. Die wahren Parameter sind auf $\beta_1 = 1,0$ und $\beta_2 = -2,5$ gesetzt.

Der 'Within'-Schätzer berechnet mit Originaldaten (Original) wird mit dem naiven 'Within'-Schätzer (Naiv) verglichen. Neben den durchschnittlichen Schätzern[8] und empirischen Standardabweichungen[9] werden noch weitere Größen berechnet. Die durchschnittliche Verzerrung zeigt an, wie stark

[8] Durchschnitt der Schätzer über alle Monte-Carlo-Wiederholungen

[9] Standardabweichung der Schätzer über alle Monte-Carlo-Wiederholungen

der mittlere Schätzer von dem wahren Parameter abweicht. Eine geringe Verzerrung kann jedoch mit einer hohen Varianz des Schätzers einhergehen. Root Mean Squared Error (RMSE: Wurzel des mittleren quadratischen Fehlers) berücksichtigt sowohl die Verzerrung als auch die Varianz (Greene 2000, S.104). Ein weiteres Maß ist die relative Standardabweichung (RELSE). Sie wird berechnet als "Durchschnitt der Standardfehler aller Monte-Carlo-Wiederholungen geteilt durch die empirische Standardabweichung der 'Within'-Schätzer über alle Monte-Carlo-Replikationen". Die Idee ist, dass mit steigender Anzahl der Replikationen die empirische Standardabweichung gegen den wahren Standardfehler konvergiert (Lechner et al. 2003). Wenn RELSE kleiner/größer 1 ist, ist der geschätzte Standardfehler nach unten/oben verzerrt. Der relative Standardfehler macht damit die Aussagen über die Genauigkeit der geschätzten asymptotischen Standardfehler der Schätzer.[10]

Tabelle 3/1 zeigt die Ergebnisse für die vaabMA. Es ist zu sehen, dass die Anonymisierung der Daten mit vaabMA die 'Within'-Schätzung nicht verändert. Bei beiden Gruppengrößen sind die durchschnittliche Verzerrung und RMSE gering. Weiter fällt auf, dass RELSE sehr nahe bei 1 liegt. Das würde implizieren, dass die vaabMA zu keinem Effizienverlust führt.

Tabelle 3/1: Variablenspezifische abstandsorientierte Mikroaggregation

A = 3	Ø Schätzer	Std.Abweichung	Ø Verzerrung	RMSE	RELSE
Original	1,000	0,011	0,000	0,011	0,987
	-2,501	0,012	-0,001	0,013	1,045
Naiv	0,999	0,011	-0,001	0,011	0,982
	-2,499	0,013	0,001	0,015	1,031

A = 5	Ø Schätzer	Std.Abweichung	Ø Verzerrung	RMSE	RELSE
Original	1,000	0,010	0,000	0,010	1,051
	-2,500	0,012	0,000	0,012	1,056
Naiv	0,998	0,010	-0,002	0,010	1,038
	-2,496	0,013	0,004	0,013	1,028

[10] zum RELSE siehe auch Lechner et al. 2003

4 Multiplikative stochastische Überlagerung

4.1 Allgemeine Überlagerung

Im Folgenden erfolgt die Analyse des Effekts der multiplikativ überlagerten Variablen auf die 'Within'-Schätzung.[11] Das lineare Panelmodell ist (2-1). Alle Regressoren und die abhängige Variable werden multiplikativ stochastisch überlagert:

$$\boldsymbol{X^a} = \boldsymbol{X} \odot \boldsymbol{U} \tag{4-13}$$

und

$$\boldsymbol{Y^a} = \boldsymbol{Y} \odot \boldsymbol{V} \tag{4-14}$$

$\odot$ bezeichnet das Hadamard-Produkt (elementenweise Multiplikation). $\boldsymbol{U}$ und $\boldsymbol{V}$ sind $(NT \times K)$-dimensionale Zufallsmatrizen.

Der Regressor ist generiert als[12]

$$x_{itk} \sim iid\ (\mu_k, \sigma_k^2), \qquad i = 1, \ldots, N,\ \ t = 1, \ldots, T, \quad k = 1, \ldots, K \tag{4-15}$$

Die einzelnen x_{itk} und y_{it} sind von den Fehlervariablen für alle i, t, k stochastisch unabhängig.

In der Literatur zu Messfehlern wird häufig unterstellt, dass die Fehlervariablen miteinander unkorreliert sind. In dieser Arbeit wird lediglich die Annahme getroffen, dass die Überlagerungen zu unterschiedlichen Zeitpunkten unkorreliert sind:

$$\begin{array}{rcl} cov\,(u_{itk}, v_{is}) & = & 0,\ t \neq s \\ cov\,(u_{itk}, u_{isl}) & = & 0,\ t \neq s, k \neq l \end{array} \qquad \left\{ \begin{array}{l} i = 1, \ldots, N, \\ t = 1, \ldots, T, \\ k = 1, \ldots, K. \end{array} \right. \tag{4-16}$$

Die Annahme

$$E[u_{itk}] = E[v_{it}] = 1 \tag{4-17}$$

stellt sicher, dass die anonymisierte und die Originalvariable den gleichen Erwartungswert haben.

Der Schätzer auf Basis der anonymisierten Daten wurde in (2-4) bzw. (2-6) definiert. Im Folgenden ist zu untersuchen, ob die einzelnen Elemente

[11] Stefanski (1985, S.586f.) hat bereits gezeigt, dass der KQ-Schätzer bei Vorliegen des multiplikativen Messfehlers inkonsistent ist, und hat einen korrigierten Schätzer hergeleitet.

[12] Diese Annahme kann bei Paneldaten verletzt sein, erleichert jedoch für den Anfang die Analyse.

in (2-6) im Fall der multiplikativen stochastischen Überlagerung konsistente Schätzer der wahren Momente sind.

Zuerst haben wir die Varianz des k-ten Regressors ${S^w_{x^a_k}}^2$ angeschaut. Sie lässt sich schreiben als[13]

$$\underset{N\to\infty}{plim}\ {S^w_{x^a_k}}^2 = \left(1 - \frac{1}{T}\right)\left[\left(\sigma_k^2 + \mu_k^2\right)\sigma_{u_k}^2 + \sigma_k^2\right] \tag{4-18}$$

mit σ_k^2, μ_k^2 als Varianz bzw. Erwartungswert des Originalregressors k und $\sigma_{u_k}^2$ als Varianz der Fehlervariable von x_k.

Für die 'Within'-Kovarianz $S^w_{x^a_k x^a_l}$ zwischen zwei Regressoren k und l erhält man[14]

$$\underset{N\to\infty}{plim}\ S^w_{x^a_k x^a_l} = \left(1 - \frac{1}{T}\right)\left[\left(\sigma_{kl} + \mu_k\mu_l\right)\sigma_{u_k u_l} + \sigma_{kl}\right] \tag{4-19}$$

σ_{kl} bezeichnet die Kovarianz der beiden wahren Regressoren k und l. $\sigma_{u_k u_l}$ steht für die Kovarianz der Fehlervariablen der jeweiligen Regressoren.

Anschließend kann man zeigen, dass[15]

$$\underset{N\to\infty}{plim}\ S^w_{x^a_k y^a}$$

$$= \left(1 - \frac{1}{T}\right)\left(\sum_l \beta_l \sigma_{kl} + \left(\sigma_{\alpha k} + \mu_\alpha\mu_k + \sum_l \beta_l\left(\sigma_{kl} + \mu_k\mu_l\right)\right)\sigma_{u_k v}\right) \tag{4-20}$$

gilt.

Dabei ist $\sigma_{u_k v}$ die Kovarianz zwischen der Fehlervariable des Originalregressors x_k und der Fehlervariable von y, $\sigma_{\alpha k}$ die Kovarianz des k-ten Regressores und des Individualeffekts. μ_α bezeichnet den Erwartungswert des Individualeffekts.

Während bei der 'Within'-Schätzung ohne Anonymisierung sich Individualeffekte rauskürzen lassen, wird aus (4-20) deutlich, dass im Fall der Anonymisierung durch allgemeine Überlagerung Individualeffekte bleiben.

Nachdem die einzelnen Elemente hergeleitet wurden, kann man schreiben:

$$\underset{N\to\infty}{plim}\ \boldsymbol{S}^w_{x^a x^a}\frac{1}{(1 - 1/T)}$$

[13] Herleitung im Appendix B.1

[14] Appendix B.2

[15] Appendix B.3

$$= \underbrace{\begin{pmatrix} \sigma_{u_1}^2(\sigma_1^2+\mu_1^2)+\sigma_1^2 & \sigma_{u_1u_2}(\sigma_{12}+\mu_1\mu_2)+\sigma_{12} & \cdots & \sigma_{u_1u_K}(\sigma_{1K}+\mu_1\mu_K)+\sigma_{1K} \\ \sigma_{u_2u_1}(\sigma_{21}+\mu_2\mu_1)+\sigma_{21} & \sigma_{u_2}^2(\sigma_2^2+\mu_2^2)+\sigma_2^2 & & \sigma_{u_2u_K}(\sigma_{2K}+\mu_1\mu_K)+\sigma_{2K} \\ \vdots & & \ddots & \\ \sigma_{u_Ku_1}(\sigma_{K1}+\mu_K\mu_1)+\sigma_{K1} & & & \sigma_{u_K}^2(\sigma_K^2+\mu_K^2)+\sigma_K^2 \end{pmatrix}}_{=\boldsymbol{A}},$$

(4-21)

$$\underset{N\to\infty}{plim}\ \boldsymbol{S}_{x^a y^a}^{w} \frac{1}{(1-1/T)}$$

$$= \begin{pmatrix} \sum_l \beta_l\sigma_{1l} + (\sigma_{\alpha 1}+\mu_\alpha\mu_1+\sum_l \beta_l(\sigma_{1l}+\mu_1\mu_l))\,\sigma_{u_1v} \\ \vdots \\ \sum_l \beta_l\sigma_{Kl} + (\sigma_{\alpha K}+\mu_\alpha\mu_K+\sum_l \beta_l(\sigma_{Kl}+\mu_K\mu_l))\,\sigma_{u_Kv} \end{pmatrix}$$

$$= \underbrace{\begin{pmatrix} \sigma_{u_1v}(\sigma_{1\alpha}+\mu_1\mu_\alpha) \\ \vdots \\ \sigma_{u_Kv}(\sigma_{K\alpha}+\mu_K\mu_\alpha) \end{pmatrix}}_{=\boldsymbol{B}} + \underbrace{\begin{pmatrix} \sigma_1^2+\sigma_{u_1v}\left(\sigma_1^2+\mu_1^2\right) & \cdots & \sigma_{1K}+\sigma_{u_Kv}(\sigma_{1K}+\mu_1\mu_K) \\ & \vdots & \\ \sigma_{K1}+\sigma_{u_Kv}(\sigma_{K1}+\mu_K\mu_1) & \cdots & {\sigma_K}^2+\sigma_{u_Kv}\left(\sigma_K^2+\mu_K^2\right) \end{pmatrix}}_{=\boldsymbol{C}} \underbrace{\begin{pmatrix} \beta_1 \\ \vdots \\ \beta_K \end{pmatrix}}_{=\boldsymbol{\beta}}$$

(4-22)

Der Wahrscheinlichkeitsgrenzwert des naiven 'Within'-Schätzers sieht dann wie folgt aus:

$$\underset{N\to\infty}{plim}\ \widehat{\boldsymbol{\beta}_{\boldsymbol{w}}^{\boldsymbol{a}}} = \boldsymbol{A}^{-1}(\boldsymbol{B}+\boldsymbol{C})\,\boldsymbol{\beta} \tag{4-23}$$

Aus (4-23) lässt sich ein korrigierter 'Within'-Schätzer herleiten:

$$\widehat{\boldsymbol{\beta}_{\boldsymbol{w}}^{\boldsymbol{korr}}} = \widehat{\boldsymbol{C}}^{-1}\left(\widehat{\boldsymbol{A}}\widehat{\boldsymbol{\beta}_{\boldsymbol{w}}^{\boldsymbol{a}}} - \boldsymbol{B}\right) \tag{4-24}$$

Mit Dach werden die geschätzten Matrizen/Vektoren bezeichnet.

Jetzt spielt es eine Rolle, wie die Fehlervariablen generiert werden. Bei korrelierten Fehlern, d.h. $\sigma_{u_kv} \neq 0$, erscheint in der Korrekturformel die Kovarianz zwischen Originalregressoren und Individualeffekten. Die Korrektur ist nur dann möglich, wenn diese Kovarianz bekannt bzw. schätzbar ist. Bei Unkorreliertheit der Fehlervariablen verschwindet $\boldsymbol{B}$ und $\boldsymbol{C}$ kürzt sich zur Varianz-Kovarianz-Matrix der Regressoren. Die Korrektur der Verzerrung ist damit ohne Kenntnis von $\sigma_{\alpha k}$ möglich.

Schätzung der Momente

In der Praxis stehen oft nur anonymisierte Daten zur Verfügung, deswegen müssen die einzelnen Momente in (4-24) geschätzt werden. Es wird jedoch davon ausgegangen, dass die Varianz-Kovarianz-Matrix der Fehlervariablen mit den Daten mitgeliefert wurde.

Wegen Erwartungstreue

$$E[x_{itk}^a] = E[x_{itk}u_{itk}] = E[x_{itk}] \tag{4-25}$$

ist der Erwartungswert des wahren Regressors durch den Mittelwert der anonymisierten Daten schätzbar:

$$\widehat{\mu_k} = \overline{x}_k^a = \frac{1}{NT} \sum_i \sum_t x_{itk}^a \tag{4-26}$$

Für die einzelnen Varianzen lässt sich herleiten:

$$Var(x_{itk}^a) = Var(x_{itk} u_{it}) = \sigma_k^2 \left(\sigma_{u_k}^2 + 1\right) + \mu_k^2 \sigma_{u_k}^2 \tag{4-27}$$

Daraus folgt[16]

$$\hat{\sigma}_k^2 = \frac{s_{x_k^a}^2 - \overline{x}_k^{a2} \sigma_{u_k}^2}{\sigma_{u_k}^2 + 1} \tag{4-28}$$

Für die Kovarianz zwischen den Regressoren erhält man

$$Cov(x_k^a, x_l^a) = \sigma_{kl} \left(\sigma_{u_k u_l} + 1\right) + \sigma_{u_k u_l} \mu_k \mu_l \tag{4-29}$$

Es resultiert

$$\widehat{\sigma_{kl}} = \frac{s_{x_k^a x_l^a} - \sigma_{u_k u_l} \overline{x}_k \overline{x}_l}{\sigma_{u_k u_l} + 1} \tag{4-30}$$

Bei korrelierten Fehlervariablen werden die Kovarianzen zwischen Regressoren und Individualeffekten benötigt. Bis jetzt ist uns leider noch unklar, wie sie aus den anonymisierten Daten geschätzt werden könnten.

Simulationsergebnisse

In diesem Abschnitt werden die Ergebnisse der Simulationsstudie für den Fall unkorrelierter Überlagerungsvariablen wiedergegeben.[17]

Das Simulationsdesign bleibt weitgehend wie im Kapitel 3.2 bei vaabMA. Die Regressoren wurden identisch und unabhängig verteilt (iid)[18]

$$X_1 \sim L\left(4,35; 1,75^2\right), \quad X_2 \sim L\left(3,45; 1,4^2\right)$$

Alle Überlagerungsvariablen sind iid verteilt und wurden aus einer Lognormalverteilung mit Erwartungswert 1 erzeugt. Neben der naiven Schätzung

[16] $s_{x^a}^2$ ist die empirische Varianz des anonymisierten Regressors:

$$s_{x^a}^2 = \frac{1}{NT} \sum_i \sum_t \left(x_{it}^a - \overline{x}^a\right)^2$$

[17] Es wurden ebenfalls Simulationen für korrelierte Fehler und bekannte Kovarianz zwischen Regressor und Individualeffekt durchgeführt. Die Ergebnisse decken sich mit den theoretischen Ausführungen. Auf ihre Darstellung wird hier jedoch aus Platzgründen verzichtet.

[18] Eine Simulationsstudie mit normalverteilten und autokorrelierten (AR(1)-Prozess) Regressoren hat zu ähnlichen Ergebnissen geführt, die hier deswegen nicht berichtet werden.

wird der korrigierte 'Within'-Schätzer untersucht. Die Korrektur wurde sowohl mit wahren als auch mit geschätzten Momenten für Originalregressoren durchgeführt. Da zwischen den beiden kein Unterschied gefunden werden konnte, werden hier nur die Ergebnisse mit geschätzten Werten berichtet.

Tabelle 4/2 zeigt die Ergebnisse. Der naive Schätzer weist eine Verzerrung auf, die mit größerer Überlagerung stark wächst. Die Korrektur reduziert diese Verzerrung auch bei hoher Anonymisierung ($\sigma_{Fehler} = 0,2$) und verbessert wesentlich RMSE. Je höher die Überlagerung, desto schlechter fällt der relative Standardfehler (RELSE) aus. Dabei erleidet der korrigierte Schätzer einen größeren Effizienzverlust als die naive Schätzung.

Tabelle 4/2: 'Within'-Schätzung bei allgemeiner Überlagerung

$\boldsymbol{\sigma_{u_1} = \sigma_{u_2} = \sigma_v = 0,06}$ **:**

	Ø Schätzer	Std.Abweichung	Ø Verzerrung	RMSE	RELSE
Original	1,001	0,011	0,001	0,011	0,971
	-2,500	0,013	0,000	0,013	1,004
Naiv	0,975	0,013	-0,025	0,028	0,968
	-2,438	0,018	0,062	0,065	0,841
Korrigiert	1,000	0,013	0,000	0,013	0,959
	-2,500	0,019	0,000	0,019	0,821

$\boldsymbol{\sigma_{u_1} = \sigma_{u_2} = \sigma_v = 0,1}$ **:**

	Ø Schätzer	Std.Abweichung	Ø Verzerrung	RMSE	RELSE
Original	1,000	0,010	0,000	0,010	1,002
	-2,500	0,013	0,000	0,013	0,980
Naiv	0,934	0,015	-0,066	0,068	0,994
	-2,334	0,026	0,166	0,168	0,726
Korrigiert	1,000	0,017	0,001	0,017	0,951
	-2,500	0,028	0,000	0,028	0,707

$\boldsymbol{\sigma_{u_1} = \sigma_{u_2} = \sigma_v = 0,2}$ **:**

	Ø Schätzer	Std.Abweichung	Ø Verzerrung	RMSE	RELSE
Original	0,999	0,010	-0,001	0,010	1,069
	-2,500	0,013	0,000	0,013	0,997
Naiv	0,775	0,023	-0,225	0,226	0,961
	-1,952	0,039	0,549	0,550	0,696
Korrigiert	0,999	0,030	-0,001	0,030	0,852
	-2,506	0,053	-0,006	0,054	0,610

4.2 Überlagerung mit einem konstanten Grundüberlagerungsfaktor

Eine spezielle Variante der multiplikativen Überlagerung wurde von Höhne (2004) vorgeschlagen. Sie hat zwei wesentliche Vorteile gegenüber dem allgemeinen Fall. Erstens zielt das Verfahren darauf ab, die Verhältnisse zwischen den Merkmalen nach der Anonymisierung zu erhalten (Ronning 2007). Zweitens wird aufgrund der anderen Modellierung der Fehlervariablen die Korrektur der Verzerrung auch ohne Kenntnis der Kovarianz zwischen Individualeffekten und Regressoren möglich. Die folgenden Ausführungen folgen Ronning (2007).

Die Regressoren und die abhängige Variable werden multiplikativ wie folgt überlagert:

$$x^a_{itk} = x_{itk}(1 \ + \ \delta d_i + \varepsilon_{itk}) \tag{4-31}$$

$$y^a_{it} = y_{it}(1 \ + \ \delta d_i + \varepsilon_{ity}), \tag{4-32}$$

$$i = 1, ..., N, \quad t = 1, ...T \text{ und } k = 1, ...K$$

δ ist ein Parameter, der den Abschlag oder Zuschlag zu Originalvariablen definiert. Die Variable d_i ist

$$d_i = \begin{cases} +1 & \text{mit Wahrscheinlichkeit } 0,5 \\ -1 & \text{mit Wahrscheinlichkeit } 0,5 \end{cases} \tag{4-33}$$

Ein i-tes Unternehmen erhält denselben Abschlag- oder Zuschlagparameter δ für alle Mermale und alle Zeitpunkte. Dadurch bleibt die Proportionalität zwischen zwei Merkmalen annähernd erhalten (Ronning 2007).[19] Über die gemeinsame Variable d_i sind die Fehlervariablen der Regressoren und der abhängigen Variablen miteinander korreliert.

Jede einzelne Beobachtung wird noch mit einem eigenen Zufallsfehler ε zusätzlich überlagert, wobei i.d.R.

$$\varepsilon_{itk} \sim N \ \left(0, \sigma^2_\varepsilon\right) \tag{4-34}$$

ist.

d_i, ε_{itk} und ε_{ity} sind miteinander und mit x_{itk} und y_{it} unkorreliert für alle i, t und k.

Weiter gilt

$$E\left[1 + \delta d_i + \varepsilon_{it}\right] = 1 \tag{4-35}$$

In der Arbeit von Ronning (2007) wird das Höhne-Verfahren ausführlich erläutert. Hier werden nur die wichtigsten Ergebnisse dargestellt.

[19] $E[\frac{X}{Y}] \quad \approx \quad E[\frac{X^a}{Y^a}]$

Für den Grenzwert des naiven Schätzers ergibt sich (Ronning 2007, Formel C-13):

$$\underset{N\to\infty}{plim}\,\widehat{\boldsymbol{\beta}_{\boldsymbol{w}}^{\boldsymbol{a}}} = \left(\underset{N\to\infty}{plim}\,\tfrac{1}{NT}(\boldsymbol{X^a})'\boldsymbol{Q}\boldsymbol{X^a}\right)^{-1}\underset{N\to\infty}{plim}\,\tfrac{1}{NT}(\boldsymbol{X^a})'\boldsymbol{Q}\boldsymbol{y^a}$$

$$= (1+\delta^2)\underbrace{\begin{pmatrix} (1+\delta^2)\sigma_1^2+\sigma_\varepsilon^2\left(\sigma_1^2+\mu_1^2\right) & (1+\delta^2)\sigma_{12} & \cdots & (1+\delta^2)\sigma_{1k} \\ (1+\delta^2)\sigma_{21} & (1+\delta^2)\sigma_2^2+\sigma_\varepsilon^2\left(\sigma_2^2+\mu_2^2\right) & \cdots & (1+\delta^2)\sigma_{2K} \\ \vdots & \vdots & \ddots & \vdots \\ (1+\delta^2)\sigma_{K1} & (1+\delta^2)\sigma_{K2} & \cdots & (1+\delta^2)\sigma_K^2+\sigma_\varepsilon^2\left(\sigma_K^2+\mu_K^2\right) \end{pmatrix}}_{=\boldsymbol{E}}^{-1} \times$$

$$\underbrace{\begin{pmatrix} \widehat{\sigma_1^2} & \cdots & \widehat{\sigma_{1K}} \\ \vdots & \ddots & \vdots \\ \widehat{\sigma_{K1}} & & \widehat{\sigma_K^2} \end{pmatrix}}_{=\boldsymbol{\Sigma_{xx}}}\underbrace{\begin{pmatrix} \beta_1 \\ \vdots \\ \beta_K \end{pmatrix}}_{=\boldsymbol{\beta}} \tag{4-36}$$

$$= (1+\delta^2)\boldsymbol{E}^{-1}\boldsymbol{\Sigma_{xx}}\boldsymbol{\beta}$$

$\boldsymbol{\Sigma_{xx}}$ ist die Kovarianzmatrix der Originalregressoren.
Der korrigierte 'Within'-Schätzer ist dann

$$\widehat{\boldsymbol{\beta}_{\boldsymbol{w}}^{\boldsymbol{korr}}} = \frac{1}{(1+\delta^2)}\widehat{\boldsymbol{S_{xx}}}^{-1}\widehat{\boldsymbol{E}}\widehat{\boldsymbol{\beta}_{\boldsymbol{w}}^{\boldsymbol{a}}}, \tag{4-37}$$

Aufgrund der Modellierung der Fehlervariablen in (4-31) und (4-32) lassen sich im Gegensatz zur allgemeinen Überlagerung Individualeffekte bei der Durchführung der Within-Transformation eliminieren und man erhält eine Korrekturformel, in der alle Momente schätzbar sind.

Schätzung der Momente

Da der Erwartungswert der anonymisierten Variable gleich 1 ist, wird $\widehat{\mu_x} = \overline{x}^a$ verwendet.

Weiter wird noch der Schätzer für die Varianz der wahren Regressoren benötigt.

$$Var(x_{itk}^a) = E\left[x_{itk}^{a2}\right] - \left(E\left[x_{itk}^a\right]\right)^2$$

$$= E\left[x_{itk}^2\left(1+\delta d_i+\varepsilon_{itk}\right)^2\right] - \left(E\left[x_{itk}+x_{itk}\delta d_i+x_{itk}\varepsilon_{itk}\right]\right)^2$$

Wegen (4-35), (4-34), Modellannahmen und $E\left[d_i^2\right]=1$
gelangt man zu

$$Var(x_{itk}^a) = \sigma_k^2\left(1+\delta^2+\sigma_\varepsilon^2\right)+\mu_k^2\left(\delta^2+\sigma_\varepsilon^2\right) \tag{4-38}$$

Damit wird die wahre Varianz geschätzt mit

$$\widehat{\sigma_k^2} = \frac{s_{k^a}^2 - (\delta^2 + \sigma_\varepsilon^2)\overline{x}_k^{a2}}{1 + \delta^2 + \sigma_\varepsilon^2} \tag{4-39}$$

Die geschätzte Kovarianz zwischen zwei Regressoren k und l wird wie folgt hergeleitet.

$$Cov\left(x_{itk}^a, x_{itl}^a\right) = E\left[x_{itk}^a x_{itl}^a\right] - E\left[x_{itk}^a\right] E\left[x_{itl}^a\right]$$

$$= E\left[x_{itk} x_{itl}\right]\left(1 + \delta^2\right) - E\left[x_{itk}\right] E\left[x_{itl}\right]$$

$$= \sigma_{kl}\left(1 + \delta^2\right) + \delta^2 \mu_k^a \mu_l^a \tag{4-40}$$

Unter Ausnutzung von (4-40) gelangt man zu dem Schätzer der Kovarianz:[20]

$$\widehat{\sigma}_{kl} = \frac{s_{k^a l^a} - \delta^2 \overline{x}_k^a \overline{x}_l^a}{(1 + \delta^2)} \tag{4-41}$$

Simulationsergebnisse

In einer Simulationsstudie wurde die Auswirkung der Anonymisierung mit dem Höhne-Verfahren, nämlich der Effekt unterschiedlicher δ-Parameter und Standardabweichungen der stochastischen Fehlervariable ε, auf die 'Within' - Schätzung überprüft und die Leistungsfähigkeit des Korrekturverfahrens getestet. Das Design der Simulation bleibt grundsätzlich wie im Kapitel 4.1. Die Überlagerungsvariablen wurden jedoch folgend den theoretischen Erläuterungen aus der Normalverteilung erzeugt. Trotz der Normalverteilung bleiben die Fehler bei gewählten Größen von δ und ε im positiviven Bereich.[21]

Die Ergebnisse sind in der Tabelle 4/3 präsentiert. Die naive Schätzung verhält sich, wie es zu erwarten war. Die Verzerrung steigt und die Effizienz sinkt mit größerer Anonymisierung. Dabei beeinträchtigt die Erhöhung von σ_ε stärker die Ergebnisse als die Wahl eines Zu- bzw. Abschlagsparameters δ. Die korrigierte Schätzung gelingt gut unabhängig von der Höhe der Überlagerung, erleidet jedoch einen Effizienzverlust.

[20] $s_{k^a l^a}$ ist die empirische Kovarianz der anonymisierten Regressoren.

[21] Die für die Anonymisierung empfohlenen Werte sind $\delta = 0,11$ und $\varepsilon = 0,03$. Das würde der Standardabweichung der Fehlervariablen bei allgemeiner Überlagerung von ca. 0,11 entsprechen.

Tabelle 4/3: 'Within'-Schätzung bei Höhne-Überlagerung

$\boldsymbol{\delta = 0,05,\ \sigma_{u_1} = \sigma_{u_2} = \sigma_v = 0,1:}$

	Ø Schätzer	Std.Abweichung	Ø Verzerrung	RMSE	RELSE
Original	1,001	0,010	0,001	0,010	0,997
	-2,501	-0,001	0,013	0,013	1,007
Naiv	0,933	0,016	-0,067	0,069	0,959
	-2,336	0,024	0,164	0,166	0,809
Korrigiert	1,000	0,018	0,000	0,018	0,901
	-2,501	0,026	-0,001	0,026	0,758

$\boldsymbol{\delta = 0,11,\ \sigma_{u_1} = \sigma_{u_2} = \sigma_v = 0,03:}$

	Ø Schätzer	Std.Abweichung	Ø Verzerrung	RMSE	RELSE
Original	1,000	0,010	0,000	0,010	1,023
	-2,500	0,013	0,000	0,013	0,981
Naiv	0,994	0,011	-0,006	0,013	0,993
	-2,484	0,015	0,016	0,022	0,897
Korrigiert	1,000	0,011	0,000	0,011	0,987
	-2,500	0,015	0,000	0,015	0,893

$\boldsymbol{\delta = 0,11,\ \sigma_{u_1} = \sigma_{u_2} = \sigma_v = 0,1:}$

	Ø Schätzer	Std.Abweichung	Ø Verzerrung	RMSE	RELSE
Original	1,000	0,010	0,000	0,010	0,986
	-2,500	0,013	-0,001	0,013	0,974
Naiv	0,933	0,016	-0,067	0,069	0,965
	-2,339	0,025	0,161	0,163	0,760
Korrigiert	1,000	0,017	0,000	0,017	0,924
	-2,503	0,028	-0,003	0,028	0,715

5 Zusammenfassung

Dieser Beitrag untersuchte den Effekt der Anonymisierungsverfahren *variablenspezifische abstandsorientierte Mikroaggregation (vaabMA)* und *multiplikative stochastische Überlagerung* auf die 'Within'-Schätzung eines linearen Panelmodells.

Zurückgreifend auf die Ergebnisse von Matthias Schmid (2007) konnte gezeigt werden, dass die vaabMA die 'Within'-Schätzung nicht beeinflusst und der naive Schätzer konsistent bleibt. Wenn die Daten mittels vaabMA anonymisiert werden, könnten damit in den empirischen Analysen naive Schätzer verwendet werden.

Wird hingegen der Paneldatensatz multiplikativ stochastisch überlagert, führt die Verwendung der naiven Schätzer zu verzerrten Ergebnissen. Es ist aber möglich, einen korrigierten 'Within'-Schätzers herzuleiten. Im Fall der

allgemeinen Überlagerung mit korrelierten Fehlervariablen ist die Korrektur der Verzerrung nur dann möglich, wenn die Kovarianz zwischen Regressor und Individualeffekt bekannt ist. Die Korrektur ist ohne Kenntnis dieser Kovarianz jedoch bei allgemeiner Überlagerung mit unkorrelierten Fehlern und bei Anonymisierung nach dem Höhne-Verfahren durchführbar.

In der vorliegenden Arbeit wurde bei der Untersuchung der multiplikativen stochastischen Überlagerung von den identisch und unabhängig verteilten Regressoren ausgegangen. Da ein Paneldatensatz Informationen über die Untersuchungseinheiten zu mehreren Zeitpunkten anbietet, wird die zeitliche Korrelation eine Rolle spielen. Ein weiterer Arbeitsschritt wäre die Aufhebung der iid-Annahme.

Eine wichtige Frage ist, wie die beschriebenen Anonymisierungsverfahren Analysen mit empirischen Daten beeinflussen. Der nächste Schritt wird sich mit der Untersuchung der Auswirkungen von variablenspezifischer abstandsorientierter Mikroaggregation und multiplikativer stochastischer Überlagerung auf die 'Within'-Schätzung unter Heranziehen der dem Projekt zur Verfügung stehenden empirischen Daten befassen.

Literatur

[1] Biørn, E. (1996) Panel Data with Measurement Errors. In: Mátyás, L., Sevestre, P. The Econometrics of Panel Data, 236-279.

[2] Greene, W. H. (2000) Econometric Analysis, 4th ed. Prentice Hall International.

[3] Höhne, J. (2004) Varianten der Zufallsüberlagerung. Arbeitspapier des Projekts "Faktische Anonymisierung wirtschaftsstatistischer Einzeldaten".

[4] Hsiao, C. (2003) Analysis of Panel Data, 2nd ed. Cambridge: Cambridge University Press.

[5] Iturria, S. J., Carroll, R. J. and Firth, D. (1999) Regression and Estimating Functions in the Presence of Multiplicative Measurement Error. Journal of the Royal Statistical Society. Series B (Statistical Methodology), 61, 547 – 561.

[6] Lechner, S., Pohlmeier, W.(2003) Schätzung ökonometrischer Modelle auf der Grundlage anonymisierten Daten. In: Gnoss, R. und Ronning, G. (Hrsg.), Anonymisierung wirtschaftsstatistischer Einzeldaten, Bd.42 von Forum der Bundesstatistik, Wiesbaden, 115-137.

[7] Lechner, S.(2007) The Multiplicative Simulation-Extrapolation Approach. Center for Quantitative Methods and Survey Research. University of Konstanz. Working Paper.

[8] Mundlak, Y.(1978) On the Pooling of Time Series and Cross-Section Data. Econometrica, 46, 69-84.

[9] Ronning, G., Sturm, R., Hoehne, J., Lenz, R., Rosemann, M., Scheffler, M., und Vorgrimler, D. (2005) Handbuch zur Anonymisierung wirtschafsstatistischer Mikrodaten. Statistisches Bundesamt: Statistik und Wissenschaft, Band 4.

[10] Ronning, G. (2007) IAW-Diskussionspapiere. Discussion Paper 30. Stochastische Überlagerung mit Hilfe der Mischungsverteilung. April 2007.

[11] Rosemann, M. (2006) Auswirkungen datenverändernder Anonymisierungsverfahren auf die Analyse von Mikrodaten. IAW-Forschungsberichte, 66.

[12] Schmid, M., Schneeweiß, H. (2005) The Effect of Microaggregation Procedures on the Estimation of Linear Models: A Simulation Study. In: Econometrics of Anonymized Micro Data, ed. by W. Pohlmeier, G. Ronning, and J. Wagner. Jahrbücher für Nationalökonomie und Statistik, 225, No. 5, Stuttgart: Lucius & Lucius, 529-543.

[13] Schmid, M. (2006a) Estimation of a Linear Model under Microaggregation by Individual Ranking. Allgemeines Statistisches Archiv, 90, 419-438.

[14] Schmid, M. (2006b) Estimation of a Linear Regression with Microaggregated Data. München : Dr. Hut.

[15] Stefanski, L. (1985) The Effects of Measurement Error on Parameter Estimation. Biometrica, 72, 583-592.

A Variablenspezifische abstandsorientierte Mikroaggregation

Die empirische 'Within'-Kovarianz $S^w_{x^a_k x^a_l}$ kann umgeschrieben werden als

$$S^w_{x^a_k x^a_l} = \frac{1}{N}\sum_i \frac{1}{T}\sum_t \left(x^a_{itk} - \bar{x}^a_{ik}\right)\left(x^a_{itl} - \bar{x}^a_{il}\right)$$

$$= \frac{1}{N}\sum_i \left(\frac{1}{T}\sum_t x^a_{itk}x^a_{itl} - \bar{x}^a_{il}\frac{1}{T}\sum_t x^a_{itk} - \bar{x}^a_{ik}\frac{1}{T}\sum_t x^a_{itl} + \frac{1}{T}\sum_t \bar{x}^a_{ik}\bar{x}^a_{il}\right)$$

$$= \frac{1}{N}\sum_i \left(\frac{1}{T}\sum_t x^a_{itk}x^a_{itl} - \bar{x}^a_{ik}\bar{x}^a_{il}\right)$$

$$= \frac{1}{N}\sum_i \left(\frac{1}{T}\sum_t x^a_{itk}x^a_{itl} - \frac{1}{T}\sum_t x^a_{itk}\frac{1}{T}\sum_s x^a_{isl}\right)$$

$$= \frac{1}{T}\sum_t \left(\frac{1}{N}\sum_i x^a_{itk}x^a_{itl}\right) - \frac{1}{T^2}\sum_t\sum_s \left(\frac{1}{N}\sum_i x^a_{itk}x^a_{isl}\right) \qquad \text{(A-42)}$$

Die empirische Gesamtkovarianz zu festen Zeitpunkten t und s ist

$$S_{x^a_{tk}x^a_{sl}} = \frac{1}{N}\sum_i x^a_{itk}x^a_{isl} - \bar{x}^a_{tk}\bar{x}^a_{sl}, \qquad \text{(A-43)}$$

wobei $t = s$ sein kann und $\bar{x}^a_{tk} = \frac{1}{N}\sum_i x^a_{itk}$ bzw. $\bar{x}^a_{tl} = \frac{1}{N}\sum_i x^a_{itl}$ sind. Einsetzen von (A-43) in (A-42) resultiert in

$$S^w_{x^a_k x^a_l} = \frac{1}{T}\sum_t \left(S_{x^a_{tk}x^a_{tl}} + \bar{x}^a_{tk}\bar{x}^a_{tl}\right) - \frac{1}{T^2}\sum_t\sum_s \left(S_{x^a_{tk}x^a_{sl}} + \bar{x}^a_{tk}\bar{x}^a_{sl}\right) \qquad \text{(A-44)}$$

Auf analoge Weise lassen sich ${S^w_{x^a_k}}^2$ und $S^w_{x^a_k y^a}$ herleiten.

B Multiplikative stochastische Überlagerung

B.1 'Within'-Varianz eines Regressors

Die 'Within'-Varianz des k-ten anonymisierten Regressors ist

$$S_{x_k}^{w\,2} = \frac{1}{NT}\sum_i\sum_t \left(x_{itk}^a - \overline{x}_{ik}^a\right)^2$$

$$= \frac{1}{NT}\sum_i\sum_t {x_{itk}^a}^2 - \frac{1}{N}\sum_i\left(2\frac{1}{T}\sum_t x_{itk}^a\overline{x}_{ik}^a - \frac{1}{T}\sum_t {\overline{x}_{ik}^a}^2\right)$$

$$= \frac{1}{NT}\sum_i\sum_t {x_{itk}^a}^2 - \frac{1}{N}\sum_i {\overline{x}_{ik}^a}^2 \tag{B-45}$$

mit
$\overline{x}_{ik}^a = \frac{1}{T}\sum_t x_{itk}^a$

Wegen Umformung in (B-45) und unter Verwendung der Regeln für das Rechnen mit Wahrscheinlichkeitsgrenzwerten[22] ergibt sich

$$\underset{N\to\infty}{plim}\ \frac{1}{N}\sum_i\frac{1}{T}\sum_t \left(x_{itk}^a - \overline{x}_{ik}^a\right)^2$$

$$= \underset{N\to\infty}{plim}\ \frac{1}{N}\sum_i\frac{1}{T}\sum_t {x_{itk}^a}^2 - \underset{N\to\infty}{plim}\ \frac{1}{N}\sum_i {\overline{x}_{ik}^a}^2 \tag{B-46}$$

Da der Stichprobenmittelwert gegen das theoretische Moment in Wahrscheinlichkeit konvergiert, müssen die Erwartungswerte der beiden Ausdrücke auf der rechten Seite von (B-46) ausgerechnet werden.

$$E\left[\frac{1}{T}\sum_t {x_{itk}^a}^2\right] = var\left(x_{itk}^a\right) + \left(E\left[x_{itk}^a\right]\right)^2 \tag{B-47}$$

$$E\left[{\overline{x}_{ik}^a}^2\right] = \frac{var\left(x_{itk}^a\right)}{T} + \left(E\left[x_{itk}^a\right]\right)^2 \tag{B-48}$$

Unter Ausnutzung der Annahmen

$$x_{itk} \sim iid\left(\mu_k, \sigma_k^2\right) \quad \text{und} \quad u_{itk} \sim iid\left(1, \sigma_{u_k}^2\right)$$

erhält man weiter

$$\underset{N\to\infty}{plim}\ \frac{1}{N}\sum_i\frac{1}{T}\sum_t \left(x_{itk}^a - \overline{x}_{ik}^a\right)^2 = \left(1 - \frac{1}{T}\right) var\left(x_{itk}^a\right) \tag{B-49}$$

[22] siehe z.B. Greene 2000, S.113

$$= \left(1 - \frac{1}{T}\right) \left(E\left[\left(x_{itk} u_{itk}\right)^2\right] - \left(E\left[x_{itk} u_{itk}\right]\right)^2\right)$$

$$= \left(1 - \frac{1}{T}\right) \left(\sigma_k^2 + \left(\sigma_k^2 + \mu_k^2\right) \sigma_{u_k}^2\right), \tag{B-50}$$

wobei σ_k^2 und μ_k^2 die Varianz und der Erwartungswert des k-ten Regressors sind. $\sigma_{u_k}^2$ bezeichnet die Varianz der Fehlervariable des k-ten Regressors.

B.2 'Within'-Kovarianz zweier Regressoren

Die 'Within'-Kovarianz zwischen zwei anonymisierten Regressoren k und l ist definiert als

$$S_{x_k x_l}^w = \frac{1}{NT} \sum_i \sum_t \left(x_{itk}^a - \overline{x}_{ik}^a\right) \left(x_{itl}^a - \overline{x}_{il}^a\right)$$

$$= \frac{1}{NT} \sum_i \sum_t x_{itk}^a x_{itl}^a - \frac{1}{N} \sum_i \left(\overline{x}_{il}^a \frac{1}{T} \sum_t x_{itk}^a - \overline{x}_{ik}^a \frac{1}{T} \sum_t x_{itl}^a + \frac{1}{T} \sum_t \overline{x}_{ik}^a \overline{x}_{il}^a\right)$$

$$= \frac{1}{NT} \sum_i \sum_t x_{itk}^a x_{itl}^a - \frac{1}{N} \sum_i \overline{x}_{ik}^a \overline{x}_{il}^a \tag{B-51}$$

Damit ist der folgende Grenzwert zu bestimmen:

$$\underset{N\to\infty}{plim}\ \frac{1}{N} \sum_i \frac{1}{T} \sum_t \left(x_{itk}^a - \overline{x}_{ik}^a\right) \left(x_{itl}^a - \overline{x}_{il}^a\right)$$

$$= \underset{N\to\infty}{plim}\ \frac{1}{N} \sum_i \frac{1}{T} \sum_t x_{itk}^a x_{itl}^a - \underset{N\to\infty}{plim}\ \frac{1}{N} \sum_i \overline{x}_{ik}^a \overline{x}_{il}^a \tag{B-52}$$

Für die Erwartungswerte von (B-52) lässt sich schreiben

$$E\left[\frac{1}{T} \sum_t x_{itk}^a x_{itl}^a\right] = cov\left(x_{itk}^a, x_{itl}^a\right) + E\left[x_{itk}^a\right] E\left[x_{itl}^a\right], \tag{B-53}$$

$$E\left[\overline{x}_{ik}^a \overline{x}_{il}^a\right] = cov\left(\overline{x}_{ik}^a, \overline{x}_{il}^a\right) + E\left[\overline{x}_{ik}^a\right] E\left[\overline{x}_{il}^a\right]$$

$$= cov\left(\frac{1}{T} \sum_t x_{itk}^a, \frac{1}{T} \sum_s x_{isl}^a\right) + E\left[\frac{1}{T} \sum_t x_{itk}^a\right] E\left[\frac{1}{T} \sum_t x_{itl}^a\right]$$

$$= \frac{1}{T^2} \sum_t \sum_s cov\left(x_{itk}^a, x_{isl}^a\right) + E\left[\frac{1}{T} \sum_t x_{itk}^a\right] E\left[\frac{1}{T} \sum_t x_{itl}^a\right]$$

$$= \frac{1}{T^2}\left(\sum_t cov\left(x^a_{itk}, x^a_{itl}\right) + \sum_t \sum_{\substack{s \\ t \neq s}} cov\left(x^a_{itk}, x^a_{isl}\right)\right) + E\left[\frac{1}{T}\sum_t x^a_{itk}\right] E\left[\frac{1}{T}\sum_t x^a_{itl}\right]$$

$$\overset{iid}{=} \frac{1}{T^2}\sum_t cov\left(x^a_{itk}, x^a_{itl}\right) + E\left[x^a_{itk}\right] E\left[x^a_{itl}\right]$$

$$= \frac{cov\left(x^a_{itk}, x^a_{itl}\right)}{T} + E\left[x^a_{itk}\right] E\left[x^a_{itl}\right] \tag{B-54}$$

(B-54) wurde unter der Annahme hergeleitet, dass die Variablen über alle i und t identisch und unabhängig verteilt sind.

Abziehen (B-54) von (B-53) führt zu

$$\underset{N\to\infty}{plim}\ \frac{1}{N}\sum_i \frac{1}{T}\sum_t \left(x^a_{itk} - \overline{x}^a_{ik}\right)\left(x^a_{itl} - \overline{x}^a_{il}\right) = \left(1 - \frac{1}{T}\right) cov\left(x^a_{itk}, x^a_{itl}\right)$$

$$= \left(1 - \frac{1}{T}\right)\left(E\left[x_{itk}u_{itk}x_{itl}u_{itl}\right] - E\left[x_{itk}u_{itk}\right] E\left[x_{itl}u_{itl}\right]\right)$$

$$= \left(1 - \frac{1}{T}\right)\left(\sigma_{kl} + \left(\sigma_{kl} + \mu_k\mu_l\right)\sigma_{u_k u_l}\right) \tag{B-55}$$

mit σ_{kl} als Kovarianz der Originalregressoren k und l und $\sigma_{u_k u_l}$ Kovarianz der Fehlervariablen von X_k und X_l.

B.3 'Within'-Kovarianz eines Regressors und der abhängigen Variable

Des Weiteren wird der Grenzwert von $S^w_{x_k y}$ benötigt.

Die Differenzen der einzelnen Beobachtungen von den Mittelwerten über die Zeit kann man schreiben als

$$x^a_{itk} - \bar{x}^a_{ik} = x_{itk}u_{itk} - \overline{xu}_{ik},$$

$$y^a_{it} - \bar{y}^a_i = \alpha_i\left(v_{it} - \overline{v}_i\right) + \sum_{l=1}^{K}\beta_l\left(x_{itl}v_{it} - \overline{x_l v}_i\right) + \left(\epsilon_{it}v_{it} - \overline{\epsilon v}_i\right)$$

mit $\overline{xu}_{ik} = \frac{1}{T}\sum_t x_{itk}u_{itk}$, $\overline{v}_i = \frac{1}{T}\sum_t v_{it}$, $\overline{x_l v}_i = \frac{1}{T}\sum_t x_{itl}v_{it}$ und $\overline{\epsilon v}_i = \frac{1}{T}\sum_t \epsilon_{it}v_{it}$.

Anschließend ist zu berechnen

$$\underset{N\to\infty}{plim}\ \frac{1}{N}\sum_i \frac{1}{T}\sum_t \left(x^a_{itk}-\overline{x}^a_{ik}\right)\left(y^a_{it}-\overline{y}^a_i\right)$$

$$=\underset{N\to\infty}{plim}\ \frac{1}{N}\sum_i \frac{1}{T}\sum_t \left[\alpha_i\left(x_{itk}u_{itk}-\overline{xu}_{ik}\right)\left(v_{it}-\overline{v}_i\right)\right]$$

$$+\underset{N\to\infty}{plim}\ \frac{1}{N}\sum_i \frac{1}{T}\sum_t \left[\left(x_{itk}u_{itk}-\overline{xu}_{ik}\right)\sum_l \beta_l\left(x_{itl}v_{it}-\overline{x_l v}_i\right)\right]$$

$$+\underset{N\to\infty}{plim}\ \frac{1}{N}\sum_i \frac{1}{T}\sum_t \left[\left(x_{itk}u_{itk}-\overline{xu}_{ik}\right)\left(\epsilon_{it}v_{it}-\overline{\epsilon v}_i\right)\right] \tag{B-56}$$

Analog zur Herleitung der 'Within'-Kovarianz zwischen zwei Regressoren ergibt sich

$$\underset{N\to\infty}{plim}\ \frac{1}{N}\sum_i \frac{1}{T}\sum_t \left[\left(\alpha_i x_{itk}u_{itk}-\overline{\alpha_i xu}_{ik}\right)\left(v_{it}-\overline{v}_i\right)\right]=\left(1-\frac{1}{T}\right)cov\left(\alpha_i x_{itk}u_{itk},v_{it}\right)$$

$$=\left(1-\frac{1}{T}\right)\left(E\left[\alpha_i x_{itk}u_{itk}v_{it}\right]-E\left[\alpha_i x_{itk}u_{itk}\right]E\left[v_{it}\right]\right)$$

$$=\left(1-\frac{1}{T}\right)\left(\sigma_{\alpha k}+\mu_\alpha\mu_k\right)\sigma_{u_k v} \tag{B-57}$$

$\sigma_{\alpha k}$ ist die Kovarianz zwischen Individualeffekt und Regressor k. μ_α bezeichnet den Erwartungswert der Individualeffekte.

$$\underset{N\to\infty}{plim}\ \frac{1}{N}\sum_i \frac{1}{T}\sum_t \left[\left(x_{itk}u_{itk}-\overline{xu}_{ik}\right)\sum_l \beta_l\left(x_{itl}v_{it}-\overline{x_l v}_i\right)\right]$$

$$=\left(1-\frac{1}{T}\right)cov\left(x_{itk}u_{itk},\sum_l \beta_l x_{itl}v_{it}\right)$$

$$=\left(1-\frac{1}{T}\right)\left(E\left[x_{itk}u_{itk}\sum_l \beta_l x_{itl}v_{it}\right]-E\left[x_{itk}u_{itk}\right]E\left[\sum_l \beta_l x_{itl}v_{it}\right]\right)$$

$$=\left(1-\frac{1}{T}\right)\left(\sum_l \beta_l\sigma_{kl}+\sum_l \beta_l\left(\sigma_{kl}+\mu_k\mu_l\right)\sigma_{u_k v}\right) \tag{B-58}$$

Anschließend erhält man

$$\underset{N\to\infty}{plim}\ \frac{1}{N}\sum_i \frac{1}{T}\sum_t \left[\left(x_{itk}u_{itk}-\overline{xu}_{ik}\right)\left(\epsilon_{it}v_{it}-\overline{\epsilon v}_i\right)\right]$$

$$= \left(1 - \frac{1}{T}\right) cov\left(x_{itk} u_{itk}, \epsilon_{it} v_{it}\right) = 0 \tag{B-59}$$

Hier wird ausgenutzt, dass der Störterm ϵ_{it}, der Regressor x_{itk} und die Fehlervariablen u_{itk} sowie v_{it} stochastisch unabhängig sind und $E\left[\epsilon_{it}\right] = 0$ gilt.

Die in (B-57), (B-58) und (B-59) hergeleiteten Ergebnisse führen zu dem Grenzwert[23]

$$\underset{N\to\infty}{plim} \frac{1}{N} \sum_i \frac{1}{T} \sum_t \left(x_{itk}^a - \overline{x}_{ik}^a\right)\left(y_{it}^a - \overline{y}_i^a\right)$$

$$= \left(1 - \frac{1}{T}\right)\left(\sum_l \beta_l \sigma_{kl} + \left(\sigma_{\alpha k} + \mu_\alpha \mu_k + \sum_l \beta_l \left(\sigma_{kl} + \mu_k \mu_l\right)\right) \sigma_{u_k v}\right) \tag{B-60}$$

[23] Wird ein Panelmodell mit einem Absolutglied definiert

$$y_{it} = c + \alpha_i + \sum_l x_{itl} \beta_l + \epsilon_{it}$$

mit $\mu_\alpha = 0$,
so erhält man statt (B-60)

$$\left(1 - \frac{1}{T}\right)\left(\sum_l \beta_l \sigma_{kl} + \left(\sigma_{\alpha k} + c\mu_k + \sum_l \beta_l \left(\sigma_{kl} + \mu_k \mu_l\right)\right) \sigma_{u_k v}\right)$$

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 Siedlungsflächenwachstums
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

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

IAW-Diskussionspapiere

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

Nr. 26 (Oktober 2006)
Bargained Wages in Decentralized Wage-Setting Regimes
Wolf Dieter Heinbach

IAW-Diskussionspapiere

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 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

Nr. 34 (November 2007)
Within-Schätzung bei anonymisierten Paneldaten
Elena Biewen