Schmollers Jahrbuch 130 (2010), 357 – 380
Duncker & Humblot, Berlin

IV-Schätzung eines linearen Panelmodells mit anonymisierten Betriebs- und Unternehmensdaten*

Von Gerd Ronning, Martin Rosemann und Elena Biewen

Abstract

Recently researchers in Germany were given the opportunity to use some of the business data provided by the Federal Statistical Office at their own workplace as Scientific Use Files (SUF). One of the anonymisation procedures used for the generation of SUF is multiplicative stochastic noise. Since this method leads to inconsistent estimates of linear panel models the use of correction procedures is necessary. This article analyses the instrument variables estimation as a correction method. As instrument variables we test a) a lagged variable, b) the difference between two lagged variables, and c) an anonymised variable from an additional anonymised data set. We conclude that only the last instrument leads to consistent estimators. We also study numerical problems which arise if the denominator of the estimator tends towards zero.

Zusammenfassung

Seit einiger Zeit ist es für Wissenschaftler / innen in Deutschland möglich geworden, einige Betriebs- und Unternehmensdaten der amtlichen Statistik am eigenen Arbeitsplatz als Scientific-Use-Files (SUFs) zu nutzen. Eines der Anonymisierungsverfahren, das bei der Erstellung von SUFs verwendet wird, ist die multiplikative stochastische Überlagerung. Da sie aber zu inkonsistenten Schätzungen linearer Panelmodelle führt, ist der Einsatz von Korrekturverfahren in Analysen notwendig. Dieser Beitrag untersucht, ob durch Instrumentvariable die Schätzverzerrung vermieden werden kann. Als Instrumente werden (a) die verzögerte Variable, (b) die Differenz von verzögerten Variablen und (c) eine (zusätzliche) anonymisierte Variable getestet. Wir kommen zu dem Ergebnis, dass lediglich das letzte Instrument zu konsistenten IV-Schätzern führt. Allerdings ergeben sich numerische Probleme, wenn Regressoren

* Der Beitrag präsentiert einige Ergebnisse des Projekts „Wirtschaftsstatistische Paneldaten und faktische Anonymisierung“, das 2006 bis 2008 vom Statistischen Bundesamt, den statistischen Landesämtern, dem Institut für Arbeitsmarkt- und Berufsforschung (IAB, Nürnberg) und dem Institut für Angewandte Wirtschaftsforschung (IAW, Tübingen) durchgeführt und vom Bundesministerium für Bildung und Forschung gefördert wurde.

eine hohe positive Autokorrelation aufweisen und die Wellenanzahl in einem Paneldatensatz gering ist.

JEL-Classification: C01, C13, C23

Received: January 25, 2010
Accepted: September 10, 2010

1. Einführung

1.1 Anonymisierung von Betriebs- und Unternehmensdaten

Mit dem Inkrafttreten des § 16, Abs. 6 des Bundesstatistikgesetzes (BStatG) im Jahr 1987 wurde der Zugang der wissenschaftlichen Forschung zu den Einzeldaten der amtlichen Statistik stark erleichtert. Wissenschaftler und Wissenschaflerinnen haben damit die Möglichkeit bekommen, mit Mikrodaten in Form faktisch anonymer Scientific-Use-Files (SUF) am eigenen Arbeitsplatz zu arbeiten. Faktische Anonymität bedeutet dabei, dass ein Rückschluss auf eine bestimmte Einheit (Person, Haushalt, Betrieb, Unternehmen) nur mit einem unverhältnismäßig großen Aufwand möglich ist. Im Fall der absoluten Anonymität hingegen ist die eindeutige Zuordnung von Einzelwerten aus einer Erhebung mit Sicherheit ausgeschlossen, was den Einsatz von besonders „starken“ Anonymisierungsmaßnahmen erfordern würde. Das Konzept der faktischen Anonymität erlaubt somit, ein besseres Analysepotenzial der Daten zu erreichen.

Nichtsdestotrotz blieb bis jetzt das Angebot an SUFs im Bereich der Betriebs- und Unternehmensdaten gering.[1] Der Grund dafür liegt vor allem darin, dass Unternehmen und Betriebe im Vergleich zu Personen und Haushalten ein deutlich höheres Reidentifikationsrisiko aufweisen. Bei personenbezogenen Daten reichen zur Sicherstellung der faktischen Anonymität in der Regel informationsreduzierende Verfahren aus, d. h. Verfahren, die sensible Informationen lediglich unterdrücken oder vergröbern. Bei Betriebs- und Unternehmensdaten kann der Datenschutz allein mit dieser Verfahrensgruppe nicht mehr gewährleistet werden. Dagegen wird bei Unternehmensdaten der Einsatz von datenverändernden Verfahren (d. h. stochastische oder systematische Veränderung der Einzelwerte) als notwendig angesehen (vgl. z. B. Ronning et al., 2005).

Eines der wichtigsten datenverändernden Anonymisierungsverfahren im Bereich der Betriebs- und Unternehmensdaten ist die multiplikative stochastische Überlagerung, bei der Originaldaten mit stochastisch erzeugten Fehlervariablen multipliziert werden. Dies führt dazu, dass große und kleine Werte

[1] Ein Überblick über die wirtschaftsstatistischen Mikrodaten, die als SUFs verfügbar sind, findet sich in Lenz / Zwick (2009).

jeweils um einen bestimmten Prozentsatz verändert werden. Die additive Variante trägt zwar zu einem guten Schutz der Kleinunternehmen bei, große Werte und somit besonders schutzbedürftige Großunternehmen werden dabei bei gleichem Anonymisierungsfehler viel weniger geschützt.

Die multiplikative Überlagerung aus einer unimodalen Verteilung (fortan einfache Überlagerung), i.d.R. einer Normalverteilung, hat jedoch den Mangel, dass viele anonymisierte Beobachtungen sehr nahe bei den Originalwerten liegen. Eine Vergrößerung der Fehlervarianzen würde in diesem Fall zwar mehr Unsicherheit erzeugen, jedoch würden dadurch die Ausgangsdaten besonders stark verändert, was zu unbrauchbaren Analysen führen könnte. Deshalb wurde für die faktische Anonymisierung von wirtschaftsstatistischen Mikrodaten ein Speziallfall der multiplikativen Überlagerung vorgeschlagen. Dabei handelt es sich um die Überlagerung mit einem konstanten Grundüberlagerungsfaktor (Höhne, 2008), bei der anstelle einer unimodalen Verteilung eine zweigipflige Mischungsverteilung verwendet wird (vgl. Roque, 2000). Die Fehlervariablen werden dabei aus einer Mischung zweier Normalverteilungen generiert. Auf diese Weise wird erreicht, dass anonymisierte Daten stärker von originalen Beobachtungen abweichen, wenn die gleiche Fehlervarianz wie bei der einfachen Überlagerung gewählt wird (Ronning, 2009).

1.2 Korrektur der Verzerrung bei multiplikativer Überlagerung

In der Literatur wurde bereits gezeigt, dass multiplikative (Mess)fehler zu inkonsistenten FE-Schätzungen führen. Stefanski (1985), Hwang (1986), Ronning et al. (2005), Rosemann (2006) zeigen z. B., dass der KQ-Schätzer im Fall von Querschnittsdaten verzerrt und inkonsistent ist. Ronning (2009), Schneeweiß / Ronning (2010) kommen zum Ergebnis, dass die FE-Schätzung auf Basis multiplikativ überlagerter Paneldaten ebenfalls inkonsistent ist. Um aussagekräftige Analysen auf Basis stochastisch überlagerter Daten machen zu können, müssen Korrekturverfahren zur Anwendung kommen.

Als Möglichkeiten zur Korrektur der Verzerrung in Fehler-in-den-Variablen-Modellen sind mehrere Varianten denkbar. Zum Beispiel kann ein Korrekturschätzer konstruiert werden (vgl. bei additiven Fehlervariablen: Griliches / Hausman (1986); Wansbeek / Koning (1991); Biørn (1996); bei multiplikativen Fehlern: Biewen (2008); Biewen / Ronning (2008); Schneeweiß / Ronning (2010). Dabei wird unterstellt, dass die Varianz der Fehlervariablen bekannt ist.

Ein relativ neues Korrekturverfahren ist die Simulation-Extrapolation-Methode (SIMEX), die zur Handhabung der Verzerrung aufgrund eines (additiven) Fehlers von Cook und Stefanski (1994) vorgeschlagen wurde. Da das Verfahren von der funktionalen Form eines Modells nicht abhängig ist, kann es sowohl in linearen als auch in nichtlinearen Modellen verwendet werden. Die

Voraussetzung ist jedoch auch hier, dass die Varianz der Fehlervariablen bekannt ist. Die Anwendbarkeit dieses Ansatzes bei multiplikativer Überlagerung wird in Rosemann (2006), Nolte (2007), Biewen / Nolte / Rosemann (2008) näher diskutiert.

Obwohl die beiden oben genannten Verfahren zu guten Korrekturen führen können, ist ihre Anwendung im Kontext der Anonymisierung erschwert. Die Daten dürfen nur dann an Forscher weitergegeben werden, wenn die Anforderungen des Datenschutzes erfüllt sind. Die Freigabe der Informationen über den Anonymisierungsprozess, wie z. B. der Parameter von Anonymisierungsfehlern, würde das Reidentifikationsrisiko erhöhen.[2]

Als Alternative zu den oben genannten Korrekturvarianten findet in der Literatur die Instrumentvariablen-Schätzung bzw. die Verallgemeinerte Momentenmethode eine breite Anwendung (vgl. Wansbeek / Koning, 1991; Biørn (1992); Biørn / Klette (1998); Wansbeek (2001); Biørn / Krishnakumar (2008). Ein großer Vorteil dieses Verfahrens besteht darin, dass keine Informationen über den Überlagerungsprozess bekanntgegeben werden müssen. Bis jetzt wurde ausschließlich der Fall der additiven Fehler untersucht.

In dem vorliegenden Beitrag wird der für die Anonymisierungspraxis relevante Fall der Anonymisierung mittels der *multiplikativen stochastischen Überlagerung* vorgestellt, wobei ein lineares Panelmodell mit Individualeffekten mit der Schätzung unter Annahme fixer Effekte (kurz FE-Schätzung) geschätzt wird. Das Ziel ist es dabei zu untersuchen, ob und in welchem Umfang sich die IV-Methode als eine Korrekturmethode eignet. Dabei wird neben den standardmäßigen Instrumentvariablen (verzögerte überlagerte Variable und die Differenz der verzögerten überlagerten Variablen) noch eine Variante des Instruments herangezogen, die speziell im Fall der Anonymisierung anwendbar ist. Hier werden den Datennutzern zwei unabhängig voneinander generierte anonymisierte Datensätze zur Verfügung gestellt. Die Variablen des zweiten Datensates werden als Instrumente verwendet.

Der Beitrag ist wie folgt gegliedert. Zuerst wird im zweiten Abschnitt das zu schätzende lineare Panelmodell präsentiert. Die Varianten der multiplikativen Überlagerung werden kurz erläutert. Abschnitt 3 fasst die wichtigsten theoretischen Ergebnisse zur IV-Schätzung zusammen. In Abschnitt 4 wird eine Simulationsstudie vorgestellt. Abschnitt 5 zeigt ein empirisches Beispiel mit Paneldaten des Monatsberichts für Verarbeitendes Gewerbe für die Jahre 1995 bis 2004. In Abschnitt 6 werden Probleme diskutiert, die aufgrund einer hohen positiven Autokorrelation und kleiner Anzahl an Zeitpunkten im Datensatz auch bei konsistenter IV-Schätzung entstehen können. Abschnitt 7 fasst die Ergebnisse dieses Beitrags zusammen.

[2] Insbesondere die Freigabe des Überlagerungsfaktors bei der Höhne-Überlagerung wird in der amtlichen Statistik als kritisch eingeschätzt.

2. Methodische Grundlage

Ein lineares Panelmodell mit Individualeffekten ist wie folgt definiert:

$$y_{it} = c + \alpha_i + \boldsymbol{x}'_{it}\,\boldsymbol{\beta} + \epsilon_{it}, \qquad i = 1, \cdots, N, t = 1, \cdots, T\,. \tag{1}$$

y_{it} ist die abhängige Variable. c steht für ein Absolutglied, wobei $c = 0$ verwendet wird. Mit α_i wird der Individualeffekt bezeichnet. $\boldsymbol{\beta}$ ist der Vektor der Regressionskoeffizienten und ϵ_{it} ist der klassische Störterm (vgl. z. B. Greene 2008, 182).

Der Regressor x_{itk} ist autokorreliert und folgt einem autokorrelierten Prozess erster Ordnung (AR(1)-Prozess):

$$x_{itk} = \rho_0 + \rho x_{i,t-1,k} + \tau_{itk}, \qquad k = 1, \cdots, K. \tag{2}$$

ρ ist dabei der Autokorrelationsparameter. Es gelten die Annahmen: $|\rho| < 1$, $E(\tau_{itk}) = 0$ und $var(\tau_{itk}) = \sigma^2_\tau$. Man beachte, dass für alle N Untersuchungseinheiten ein identischer autoregressiver Parameter ρ gewählt wird.

Im Folgenden werden Modellvariablen mittels multiplikativer stochastischer Überlagerung anonymisiert. Zwei Varianten der Überlagerung werden dabei untersucht: einfache Überlagerung und Überlagerung mit einem konstanten Grundüberlagerungsfaktor. Da das letztere Verfahren von Höhne (2008) vorgeschlagen wurde, wird es im Folgenden auch als Höhne-Überlagerung bezeichnet.

Bei einfacher Überlagerung wird eine stochastisch erzeugte Fehlervariable zu der Originalvariable wie folgt hinzugefügt:

$$x^a_{itk} = x_{itk} * u_{itk}\,, \quad y^a_{it} = y_{it} * v_{it}\,. \tag{3}$$

Die Fehlervariablen u_{itk} und v_{it} haben den Erwartungswert Eins und die Varianz σ^2_u bzw. σ^2_v. Die Annahme bezüglich des Erwartungswertes stellt sicher, dass die anonymisierte Variable den gleichen Erwartungswert wie die Originalvariable hat. Des Weiteren wird unterstellt, dass Fehler mit Originalvariablen für alle i, t, k unkorreliert sind. Die Verteilung der Fehler ist beliebig, wobei in der Regel positive Verteilungen unterstellt werden.

Bei der Höhne-Überlagerung wird die Fehlervariable in zwei Schritten generiert:

$$\begin{aligned} x^a_{itk} &= x_{itk} * u^H_{itk} = x_{itk}\,(1 + \delta D_i + \varepsilon_{itk})\,, \\ y^a_{it} &= y_{it} * v^H_{it} = y_{it}(1 + \delta D_i + \varepsilon_{ity}) \end{aligned} \tag{4}$$

mit

$$D_i = \begin{cases} +1 \text{ mit Wahrscheinlichkeit } 0{,}5 \\ -1 \text{ mit Wahrscheinlichkeit } 0{,}5\,. \end{cases} \tag{5}$$

Im ersten Schritt werden alle Werte der *i*-ten Querschnittseinheit in eine gemeinsame Richtung mit dem Grundüberlagerungsfaktor $(1+\delta)$ oder $(1-\delta)$ verschoben. Im zweiten Schritt wird noch eine zufällig erzeugte Variable addiert, um eine zusätzliche Veränderung in den Daten zu erreichen.[3] Da bei der Überlagerung der Daten der *i*-ten Einheit die gleiche Variable D_i verwendet wird, sind die Fehlervariablen miteinander korreliert. Bei dieser Variante der Überlagerung haben die Fehlervariablen u_{itk}^H und v_{it}^H den Erwartungswert Eins, die Varianz $(\delta^2+\sigma_{\varepsilon_k}^2)$ bzw. $(\delta^2+\sigma_{\varepsilon_y}^2)$ und sind mit Originalvariablen unkorreliert. Die zusätzliche Fehlervariable $\varepsilon_{itj}(j=k,y)$ ist normalverteilt $\left(\varepsilon_{itj} \sim N(0,\sigma_{\varepsilon_j}^2)\right)$ und mit allen anderen Variablen unkorreliert.

Im Vergleich zur einfachen Überlagerung trägt dieses Verfahren zu einem besseren Schutz der Unternehmensdaten bei, da bei gleicher Varianz des Fehlers anonymisierte Variablen weiter von den Originalwerten entfernt sind (vgl. Ronning 2009, 13).

Im Fall der Anonymisierung haben Wissenschaftler i.d.R. keine Informationen über Originalvariablen, können aber anonymisierte Variablen beobachten. Die naive FE-Schätzung des Modells (1)

$$\widehat{\beta_{FE}^a} = \left(\boldsymbol{X}^{a\prime}\boldsymbol{Q}\boldsymbol{X}^a\right)^{-1}\boldsymbol{X}^{a\prime}\boldsymbol{Q}\boldsymbol{Y}^a \tag{6}$$

ergibt eine verzerrte und inkonsistente Schätzung (vgl. Biewen / Ronning, 2008). $\boldsymbol{X}^a$ ist dabei die $(NT \times K)$-Matrix der anonymisierten Regressoren. $\boldsymbol{Q}$ ist die $(NT \times NT)$-dimensionale Matrix, die die Within-Transformation ausführt, d. h. von jeder Beobachtung wird der Mittelwert dieser Beobachtung über die Zeit hinweg abgezogen:

$$\boldsymbol{Q} = \begin{pmatrix} \boldsymbol{I} - \frac{1}{T}\iota\iota' & 0 & \dots & 0 \\ 0 & \boldsymbol{I} - \frac{1}{T}\iota\iota' & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \dots & 0 & \boldsymbol{I} - \frac{1}{T}\iota\iota' \end{pmatrix}, \tag{7}$$

[3] Das Höhne-Verfahren wird ausführlich in Ronning (2009) und Höhne (2008) beschrieben. Ronning (2009) zeigt, dass es sich bei der Höhne-Überlagerung um eine Mischungsverteilung handelt.

wobei $\boldsymbol{I}$ eine $(NT \times NT)$ Einheitsmatrix und ι einen T-dimensionalen Einser-Vektor bezeichnen.

Auf diese Weise werden die in Gleichung (1) vorhandenen Individualleffekte eliminiert.

3. Zur Theorie der IV-Schätzung

3.1 Instrumente

Als ein Verfahren zur Korrektur der Verzerrung wenden wir die IV-Schätzung an. Die Untersuchung in dieser Arbeit beschränkt sich auf den Fall, dass nur eine Instrumentvariable z für den fehlerhaften Regressor verwendet wird. Ein gültiges Instrument soll dabei drei Eigenschaften erfüllen (vgl. Carroll et al. 1995, 129):

A(1): Die Korrelation zwischen dem Instrument und der fehlerfreien Variable soll möglichst hoch sein.

A(2): Das Instrument und der Störterm der Gleichung sollen unkorreliert sein.

A(3): Der Anonymisierungsfehler und das Instrument sollen unkorreliert sein.

Paneldaten bieten wegen ihrer Mehrdimensionalität eine größere Auswahl an Instrumentvariablen als reine Querschnitts- bzw. Zeitreihendaten. Wenn Überlagerungsfehler miteinander unkorreliert sind und die Originalvariable x eine Autokorrelation aufweist, könnte jede Beobachtung von x^a zu einem anderen Zeitpunkt als ein Instrument verwendet werden (Sevestre et al. 1996). Das würde bedeuten, dass für jede Periode $(T-1)$ Instrumente verfügbar sind. In diesem Fall ist jedoch Vorsicht geboten. Im Kontext der Paneldaten sollte zuerst überprüft werden, ob die verzögerte Variable tatsächlich ein gültiges Instrument ist. Die Korrelation der Individualeffekte mit den Regressoren und die zur Eliminierung des Individualeffekts notwendige Transformation können zur Verletzung der Validität solch eines Instruments führen (Griliches et al. 1986, 94).

Als ein weiteres „internes“ Instrument bietet sich die Variable an, die als Differenz zwischen zwei Zeitpunkten gebildet wurde ($z_{i(t\theta)k} = x^a_{itk} - x^a_{i\theta k}\,,\ \ t \neq \theta$). Auch in diesem Fall sollte zuerst untersucht werden, ob im zu untersuchenden Modell die Annahmen eines gültigen Instruments nicht verletzt werden.

Im Rahmen der Anonymisierung wäre noch eine weitere Alternative denkbar. Die Daten bereitstellende Institution erzeugt mindestens zwei oder sogar mehrere anonymisierte Versionen eines Datensatzes, die den Datennutzern zur Verfügung gestellt werden. Da die überlagerte Variable aus der Original-

variable (und einem stochastisch erzeugten Fehler) besteht, ist eine Korrelation zwischen den gleichen Variablen aller anonymisierten Datenversionen gegeben. Auf diese Weise kann die gleiche Variable eines anderen Datensatzes als Instrument verwendet werden. Die Bereitstellung zusätzlicher anonymisierter Datensätze bedeutet jedoch ein höheres Re-Identifikationsrisiko. Aus Gründen des Datenschutzes ist es deswegen zu empfehlen, die weiteren Datensätze stärker zu anonymisieren. Dabei ist jedoch zu beachten, dass mit steigender Überlagerung die Korrelation zwischen dem Regressor und der Instrumentvariable geringer werden kann und dadurch das Problem der schwachen Instrumentalisierung entstehen könnte (vgl. z. B. Bound et al., 1995; Staiger et al., 1997).

3.2 Wahrscheinlichkeitsgrenzwert des IV-Schätzers

Im Folgenden werden Wahrscheinlichkeitsgrenzwerte des IV-Schätzers für die jeweilige Intsrumentvariable bestimmt.[4] Dabei werden alle Grenzwerte für $N \to \infty$ hergeleitet, da Paneldatensätze über Betriebe und Unternehmen in der Regel eine hohe Anzahl an Querschnittsbeobachtungen und wenig Zeitpunkte enthalten. Der Einfachheit halber wird ein einfaches lineares Panelmodell betrachtet.

Der Wahrscheinlichkeitsgrenzwert des IV-Schätzers lässt sich wie folgt schreiben:

$$\begin{aligned} plim\,\widehat{\beta^{IV}_{FE}} &= \frac{plim\frac{1}{N}\sum_{i=1}^{N}\frac{1}{N}\sum_{t=1}^{T}(z_{it}-\overline{z_i})(y^a_{it}-\overline{y^a_i})}{plim\frac{1}{N}\sum_{i=1}^{N}\frac{1}{N}\sum_{t=1}^{T}(z_{it}-\overline{z_i})(x^a_{it}-\overline{x^a_i})} \\ &= \frac{E[S_{zx,i}]}{E[S_{zx^a,i}]}\beta. \end{aligned} \tag{8}$$

$E[S_{zx,i}]$ ist hier der Erwartungswert der empirischen Within-Kovarianz zwischen Instrument z und Regressor x der i-ten Querschnittseinheit (analog für $E[S_{zx^a,i}]$):[5]

$$S_{zx,i} = \frac{1}{T}\sum_{t=1}^{T}(z_{it}-\overline{z_i})(x_{it}-\overline{x_i})$$

und $\overline{z_i}$ der Mittelwert der i-ten Beobachtung über die Zeit:

[4] Für detaillierte Herleitungen siehe z.B. Ronning (2009), Biewen/Ronning/Rosemann (2009).

[5] Zur Erinnerung: hier wurde unterstellt, dass alle N Beobachtungen demselben AR(1)-Prozess folgen.

$$\overline{z_i} = \frac{1}{T}\sum_{t=1}^{T} z_{it} \quad \text{(analog für andere Variablen)}.$$

Der IV-Schätzer ist somit dann konsistent, wenn der Erwartungswert der empirischen Within-Kovarianz zwischen z und x^a ($E[S_{zx^a}]$) demjenigen zwischen z und x ($E[S_{zx}]$) entspricht. Dieses wird im nächsten Schritt überprüft.

3.2.1 Verzögerte Variable als Instrument

Als Instrumentvariable wird zuerst die verzögerte anonymisierte Variable verwendet:

$$z_{it} = x^a_{it-1} = x_{it-1}u_{it-1}\,. \tag{9}$$

Für den Wahrscheinlichkeitsgrenzwert des IV-Schätzers im Fall einfacher Überlagerung lässt sich herleiten (vgl. Ronning 2009, 125, Gl. (12 – 24)):

$$plim\,\widehat{\beta^{IV}_{FE}} = \frac{E\left[S_{zx,i}\right]}{E[S_{zx,i}] - \frac{(T-2)}{(T-1)^2}\sigma^2_u\left(\gamma_0 + \left(\frac{\rho_0}{1-\rho}\right)^2\right)}\beta\,. \tag{10}$$

γ_0 steht für die Varianz des Originalregressors. ρ_0 und ρ sind Parameter des AR(1)-Prozesses (vgl. Gl. (2)).

Für $T \geq 3$ ist die Instrumentvariablen-Schätzung nicht konsistent. Die Verzerrung resultiert dabei aufgrund der Within-Transformation. Da der Nenner kleiner als der Zähler ausfällt, ist der IV-Schätzer nach oben verzerrt $\left(plim\,\widehat{\beta^{IV}_{FE}} > \beta\right)$.

Im Fall der Höhne-Überlagerung wird die folgende Instrumentvariable verwendet:

$$z^H_{it} = x^a_{it-1} = x_{it-1}\left(1 + \delta D_i + \varepsilon_{i,t-1,x}\right)\,. \tag{11}$$

Für den Wahrscheinlichkeitsgrenzwert des IV-Schätzers resultiert (vgl. Ronning 2008, 126, Gl. (12 – 25))

$$plim\,\widehat{\beta^{IV}_{FE}} = \frac{E\left[S_{zx,i}\right]}{E[S_{zx,i}] - \frac{(T-2)}{(T-1)^2}\left(\delta^2 + \sigma^2_{\epsilon_x}\right)\left(\gamma_0 + \left(\frac{\rho_0}{1-\rho}\right)^2\right)}\beta\,. \tag{12}$$

Dabei ist zu beachten, dass aufgrund unterschiedlicher Konstruktion der Fehlervariablen die Erwartungswerte $E[S_{zx,i}]$ in Gl. (10) und (12) nicht identisch sind.

Die IV-Schätzung in diesem Fall ist ebenfalls inkonsistent. Dieses Ergebnis ist allerdings nicht verwunderlich, da die Fehlervariable $u_{it}^H = 1 + \delta D_i + \varepsilon_{itx}$ und das Instrument $z_{it}^H = x_{it-1}\left(1 + \delta D_i + \varepsilon_{i,t-1,x}\right)$ über den gemeinsamen Term δD_i korreliert sind. Dieses führt zur Verletzung der Annahme A(3), die Unkorreliertheit zwischen dem Instrument und der Überlagerungsvariable unterstellt.

3.2.2 Erste Differenz der verzögerten Variablen als Instrument

Des Weiteren wird als Instrument die Variable verwendet, die aus der Differenz der verzögerten anonymisierten Variable gebildet wird:[6]

$$z_{it} = x_{it}^a - x_{it-1}^a = x_{it}u_{it} - x_{it-1}u_{it-1} \tag{13}$$

bzw. bei der Höhne-Überlagerung:

$$z_{it}^H = x_{it}\left(1 + \delta D_i + \varepsilon_{itx}\right) - x_{i,t-1}(1 + \delta D_i + \varepsilon_{i,t-1,x}) \ . \tag{14}$$

Es kann gezeigt werden, dass die Verwendung der Differenzvariable als Instrument für beide Varianten der Überlagerung zur Verletzung der Annahme A(3) führt. Für die Kovarianz zwischen Instrument und Fehlervariable erhält man bei der einfachen Überlagerung:

$$\begin{aligned} cov(z_{it}, u_{it}) &= cov(x_{it}u_{it} - x_{it-1}u_{it-1}, u_{it}) \\ &= E[x_{it}]var(u_{it}) - E[x_{it-1}]cov(u_{it}, u_{it-1}) \ . \end{aligned} \tag{15}$$

Die Annahme A(3) ist dann erfüllt, wenn die Fehlervariable nicht autokorreliert ist und gleichzeitig $var(u_{it}) = 0$ gilt. Dies wäre jedoch im Modell mit überlagerten Variablen nicht möglich, da $\sigma_u^2 = 0$ den Fall ohne Anonymisierung darstellen würde.

Für die Höhne-Überlagerung gilt die gleiche Argumentation wie bei der verzögerten Variable. Die Annahme A(3) ist a priori aufgrund eines gemeinsamen Faktors in der Fehler- und Instrumentvariable verletzt bzw. die Kovarianz ist

[6] $t > 1$. In diesem Beitrag wird nur die erste Differenz vorgestellt. Diese Ergebnisse lassen sich aber auch auf den Fall höherer Differenzen übertragen.

$$
\begin{aligned}
cov(z_{it}, u_{it}) &= cov(x_{it}(1+\delta D_i+\varepsilon_{itx}) - x_{it-1}(1+\delta D_i+\varepsilon_{i,t-1,x}), 1+\delta D_i+\varepsilon_{itx}) \\
&= E[x_{it}](\delta^2+\sigma^2_{\varepsilon_x}) - E[x_{it-1}]\delta^2 \neq 0\,.
\end{aligned} \tag{16}
$$

Damit ist eine der Bedingungen für eine konsistente IV-Schätzung nicht mehr erfüllt. Für die einfache Überlagerung erhält man den Grenzwert (vgl. Biewen / Ronning / Rosemann 2009, 41 f.):

$$
plim\,\widehat{\beta^{IV}_{FE}} = \frac{E\left[S_{zx,i}\right]}{E[S_{zx,i}] - \frac{1}{(T-2)^2}\sigma^2_u\left(\gamma_0 + \left(\frac{\rho_0}{1-\rho}\right)^2\right)}\beta \tag{17}
$$

bzw. für die Höhne-Überlagerung:

$$
plim\,\widehat{\beta^{IV}_{FE}} = \frac{E\left[S_{zx,i}\right]}{E[S_{zx,i}] - \frac{1}{(T-2)^2}(\delta^2+\sigma^2_{\epsilon_x})\left(\gamma_0 + \left(\frac{\rho_0}{1-\rho}\right)^2\right)}\beta\,, \tag{18}
$$

wobei $E\left[S_{zx,i}\right]$ in Gl. (17) und (18) unterschiedlich ist.

3.2.3 Zusätzliche anonymisierte Variable als Instrument

Das Instrument im Fall der einfachen Überlagerung ist wie folgt definiert:

$$
z_{it} = x_{it}\nu_{it}\,. \tag{19}
$$

ν_{it} ist die Fehlervariable, mit der die Originaldaten überlagert werden. ν_{it} hat den Erwartungswert Eins und die Varianz σ^2_ν und ist mit dem Fehler aus dem ersten anonymisierten Datensatz (u_{it}) für alle i,t unkorreliert. Des Weiteren wird der Fall betrachtet, bei dem ν_{it} nicht autokorreliert ist, d. h. $cov(\nu_{it}, \nu_{is}) = 0\,(t \neq s)$.

Für den Wahrscheinlichkeitsgrenzwert des IV-Schätzers resultiert (vgl. Biewen / Ronning / Rosemann 2009, 29, Gl. (5 – 82)):

$$
plim\,\widehat{\beta^{IV}_{FE}} = \frac{E\left[S_{zx,i}\right]}{E[S_{zx,i}]}\beta = \frac{\gamma_0\left(1-\frac{1}{T}\right) - \frac{2}{T^2}\sum_{j=1}^{T-1}(T-j)\gamma_j}{\gamma_0\left(1-\frac{1}{T}\right) - \frac{2}{T^2}\sum_{j=1}^{T-1}(T-j)\gamma_j}\beta = \beta \tag{20}
$$

mit γ_j als Autokovarianz des Regressors. Aufgrund der Gleichheit des Zählers und Nenners in Gl. (20) erhält man somit einen konsistenten Schätzer.

Bei der Höhne-Variante wird das Instrument

$$z_{it}^{H} = x_{it}\left(1 + \delta^{z} D_i^{z} + \varepsilon_{itx}^{z}\right) \tag{21}$$

verwendet.

An dieser Stelle ist anzumerken, dass jetzt die Fehlervariable des Instruments – insbesondere D_i^z – unabhängig von dem Fehler in der ursprünglichen anonymisierten Variable (insbesondere D_i) erzeugt wird. Daher ist die im Fall beider anderen Instrumente verletzte Annahme erfüllt, d. h. das Instrument und die Fehlervariable sind unkorreliert.

Für den Wahrscheinlichkeitsgrenzwert des IV-Schätzers folgt (vgl. Biewen / Ronning / Rosemann 2009, 51, Gl. (6 – 165)):

$$plim\, \widehat{\beta_{FE}^{IV}} = \frac{E[S_{zx,i}]}{E[S_{zx,i}]}\beta = \frac{\gamma_0\left(1-\frac{1}{T}\right) - \frac{2}{T^2}\sum_{j=1}^{T-1}(T-j)\gamma_j^{H}}{\gamma_0\left(1-\frac{1}{T}\right) - \frac{2}{T^2}\sum_{j=1}^{T-1}(T-j)\gamma_j^{H}}\beta = \beta\,. \tag{22}$$

γ_j^H ($j = 1, \ldots, T-1$) ist die Autokovarianz der Variable, die mit der Höhne-Überlagerung anonymisiert wurde.

Dieser IV-Schätzer ist somit konsistent.

4. Simulationsstudie

4.1 Simulationsdesign

Die theoretischen Ergebnisse werden mithilfe einer Simulationsstudie überprüft. Dabei wird das einfache lineare Panelmodell (1) mit der FE-Schätzung geschätzt.

Der Modellstörterm ϵ_{it} ist normalverteilt mit Erwartungswert Null und Varianz $\sigma_\epsilon^2 = 0{,}25$. α_i ist ein Individualeffekt, der mit x_{it} korreliert ist. Die Korrelation zwischen α_i und x_{it} wird nach der Formel von Biørn (1996, 260 f.) mit $\lambda = 1$ und $\xi_i \sim N(0,1)$ erzeugt:

$$\alpha_i = \left(\overline{x_i} - E[x]\right)\lambda + \xi_i\,. \tag{23}$$

Der Regressor folgt einem AR(1)-Prozess mit $\rho_0 = 4{,}35$ und $\tau_{it} \sim N(0,\, 1)$ (vgl. Gl. (2)). ρ nimmt unterschiedliche Werte an: $\rho = 0{,}1$, 0,5 und 0,9. Der wahre Parameter β ist $-2{,}5$.

Die Originalvariablen x und y werden mittels der multiplikativen Überlagerung anonymisiert. Die Fehlervariablen bei der einfachen Überlagerung

sind normalverteilt mit Erwartungswert Eins und Varianz $0{,}114^2$, wobei die Fehler miteinander unkorreliert sind. In der Höhne-Variante werden die Parameter $\delta = 0{,}11$ und $\sigma_{\varepsilon_j} = 0{,}03$ $(j = x, y)$ verwendet.[7] Die Anonymisierungsniveaus in beiden Überlagerungsvarianten sind somit identisch.

Im Fall der IV-Schätzung werden drei Instrumentvariablen getestet: (1) die verzögerte anonymisierte Variable (verz.Var.), (2) die Variable, die aus Differenz der verzögerten anonymisierten Variablen gebildet wurde, (Dif.Var.) und (3) die Variable, die aus einem anderen anonymisierten Datensatz erzeugt wurde (anon.Var.). Im letzten Fall wird der neue Datensatz mit dem Ziel des Datenschutzes stärker überlagert. Der Fehler der Instrumentvariable wird dabei bei einfacher Überlagerung mit $\sigma_u = \sigma_v = 0{,}206$ erzeugt. Bei der Höhne-Methode wird er mit $\delta = 0{,}2$ und $\sigma_{\varepsilon_j} = 0{,}05$ $(j = x, y)$ generiert.

Es werden $N = 1.000$ Querschnittseinheiten simuliert, die Anzahl der Wellen des Panels ist alternativ $T = 4$ und $T = 10$. Die Anzahl der Monte-Carlo-Replikationen beträgt 500, in jeder Zeitreihe werden 200 erste Beobachtungen herausgenommen, um den Effekt des Startwertes zu eliminieren.

Man beachte, dass die Korrelation zwischen Instrument *z* und Regressor *x*, die oftmals als Gütemaß für ein Instrument betrachtet wird, bei dem verwendeten Simulationsdesign abnimmt, wenn der autoregressive Parameter ρ wächst. Wir illustrieren dies hier für das Instrument ‚anon. Var'. Im Fall der einfachen Überlagerung erhalten wir als Korrelationsmaß

$$
\begin{aligned}
korr[x,z] &= \frac{\sigma_x^2}{\sqrt{\sigma_x^2 \sigma_z^2}} \\
&= \frac{\frac{\sigma_\tau^2}{1-\rho^2}}{\sqrt{\frac{\sigma_\tau^2}{1-\rho^2}\left[\frac{\sigma_\tau^2}{1-\rho^2} + \sigma_u^2\left\{\frac{\sigma_\tau^2}{1-\rho^2} + \left(\frac{\rho_0}{1-\rho}\right)^2\right\}\right]}} \,.
\end{aligned} \tag{24}
$$

Aus der Formel ergibt sich, dass vor allem wegen des Terms $\mu_x^2 = \left(\frac{\rho_0}{1-\rho}\right)^2$ die Korrelation sinkt, wenn ρ steigt.

Bei Bestimmung der Korrelation im Fall der Höhne-Überlagerung ist zu beachten, dass die Zuschlagsvariable D_i^z gemäß Gl. (5) und Gl. (21) für jedes Unternehmen nur einmal gezogen wird, d. h. der Zu- bzw. Abschlag δ ist konstant für alle Zeitpunkte. Für Untersuchungseinheiten mit einem positiven Zuschlag ergibt sich als (bedingte) Korrelation

[7] Diese Parameter wurden im Projekt „Wirtschaftsstatistische Paneldaten und faktische Anonymisierung" unter Berücksichtigung des Trade-off zwischen Analysepotenzial und Sicherheitsrisiko in Simulationsstudien ermittelt.

$$(25) \qquad korr[x,z|D=+1] = \frac{(1+\delta)\sigma_x^2}{\sqrt{\sigma_x^2\left\{\sigma_x^2 * (1+\delta^2) + [\mu_x^2+\sigma_x^2]\sigma_\varepsilon^2\right\}}}$$

und für Unternehmen mit einem Abschlag erhalten wir

$$(26) \qquad korr[x,z|D=-1] = \frac{(1-\delta)\sigma_x^2}{\sqrt{\sigma_x^2\left\{\sigma_x^2 * \left(1-\delta^2\right) + \left[\mu_x^2+\sigma_x^2\right]\sigma_\varepsilon^2\right\}}}$$

mit

$$\mu_X = \frac{\rho_0}{1-\rho} \quad \text{und} \quad \sigma_x^2 = \frac{\sigma_\tau^2}{1-\rho^2}\,.$$

Auch hier wird durch die Abhängigkeit dieser beiden Momente von ρ die Korrelation für wachsendes ρ geringer.

Die Tabelle 1 zeigt die Korrelation zwischen Regressor und dem Instrument ‚anon. Var.' für ausgewählte Parameter, die denen in der Simulationsstudie entsprechen. Für die Höhne-Überlagerung wird auch das arithmetische Mittel aus den bedingten Korrelationen dargestellt.

Tabelle 1

Korrelation zwischen *x* und *z* – variables Niveau $\left(\mu_x = \frac{\rho_0}{1-\rho}\right)$

ρ	–0,900	–0,500	–0,100	0,000	0,100	0,500	0,900
μ_x	2,289	2,900	3,954	4,350	4,833	8,700	43,500
σ_x^2	5,263	1,333	1,010	1,000	1,010	1,333	5,263
Einfache Überl.: $korr[x,z]$	0,960	0,873	0,767	0,736	0,702	0,538	0,247
Höhne-Überl.:							
$korr[x,z\|D=+1]$	0,998	0,993	0,985	0,983	0,979	0,953	0,784
$korr[x,z\|D=-1]$	0,996	0,986	0,969	0,963	0,955	0,903	0,644
korr(Durchschnitt)	0,997	0,989	0,977	0,973	0,967	0,928	0,714

Bemerkungen: Ergebnisse für das Instrument ‚anon. Var.'
$\rho_0 = 4{,}35$, $\sigma_\tau^2 = 1$; $\delta = 0{,}2$, $\sigma_\varepsilon^2 = 0{,}05^2$; $\sigma_u^2 = \delta^2 + \sigma_\varepsilon^2$

Vor allem bei der einfachen Überlagerung nimmt die Korrelation zwischen *x* und *z* für wachsendes ρ dramatisch ab! Allerdings ist dies, wie bereits weiter oben bemerkt, ein Effekt, der durch das steigende Niveau von μ_x bewirkt wird. Würde man dagegen ein konstantes Niveau von μ_x für alle Werte von ρ unter-

stellen, so würde sich ein deutlich anderes Bild ergeben, da jetzt nur die Varianz σ_x^2 variiert. Siehe dazu Tabelle 2: Für (absolut betrachtet) großes ρ ergibt sich nun eine stärkere Korrelation zwischen Instrument und Regressor, d. h. das Instrument ist umso schwächer, je geringer die Autokorrelation ist, was auch intuitiv einleuchtet.

Tabelle 2

Korrelation zwischen *x* und *z* – konstantes Niveau ($\mu_x = \text{konst.}$)

ρ	–0,900	–0,500	–0,100	0,000	0,100	0,500	0,900
μ_x	4,350	4,350	4,350	4,350	4,350	4,350	4,350
σ_x^2	5,263	1,333	1,010	1,000	1,010	1,333	5,263
Einfache Überl.: $korr[x,z]$	0,914	0,779	0,737	0,736	0,737	0,779	0,914
Höhne-Überl.: $korr[x,z\|D=+1]$	0,996	0,987	0,983	0,983	0,983	0,987	0,996
$korr[x,z\|D=-1]$	0,991	0,971	0,963	0,963	0,963	0,971	0,991
korr(Durchschnitt)	0,993	0,979	0,973	0,973	0,973	0,979	0,993

Bemerkungen: Ergebnisse für das Instrument ‚anon. Var.'
$\rho_0 = 4{,}35$, $\sigma_\tau^2 = 1$; $\delta = 0{,}2$, $\sigma_\varepsilon^2 = 0{,}05^2$; $\sigma_u^2 = \delta^2 + \sigma_\varepsilon^2$

4.2 Simulationsergebnisse

Tabelle 3 stellt durchschnittliche Schätzer und in Klammern ihre Standardabweichungen über alle Monte-Carlo-Wiederholungen dar. Sowohl die verzögerte als auch die Differenzvariable führen zu verzerrten Schätzergebnissen. Außerdem wächst mit steigender Autokorrelation des Regressors die Verzerrung der naiven FE-Schätzung. Beides entspricht den theoretischen Ergebnissen aus Abschnitt 3 sowie in Ronning (2009, Abschnitt 12). Da die Within-Transformation neben den Individualeffekten auch konstante Faktoren eliminiert und die Fehlervariable bei der Höhne-Überlagerung einen konstanten Grundüberlagerungsfaktor enthält, fällt die Verzerrung in diesem Fall viel kleiner aus. Mit steigendem T werden die Ergebnisse der naiven Schätzung etwas besser. Auch dies entspricht den theoretischen Ergebnissen.

Im Fall des kleinen und mittleren ρ ($\rho = 0{,}1$ und $0{,}5$) führt die IV-Schätzung nur dann zu zufriedenstellenden Ergebnissen, wenn als Instrument der Regressor aus einem anderen anonymisierten Datensatz verwendet wird. Auch für sehr kleines T ($T = 4$) erhält man IV-Schätzer, die nahe beim wahren Schätzer liegen. Dabei schneidet die Höhne-Überlagerung in termini der Standardabweichungen besser ab.

Tabelle 3

IV-Schätzung bei multiplikativer Überlagerung

	$T = 4$				$T = 10$		
	$\rho = 0{,}1$	$\rho = 0{,}5$	$\rho = 0{,}9$		$\rho = 0{,}1$	$\rho = 0{,}5$	$\rho = 0{,}9$
Orig.		–2,4999 (0,0045)				–2,4999 (0,0026)	
Einfache Überl.:							
Naiv	–1,8775 (0,0324)	–1,1539 (0,0430)	–0,0824 (0,0473)		–1,8918 (0,0179)	–1,3059 (0,0238)	–0,1521 (0,0281)
IV: verz.Var.	–1,7634 (0,1366)	–0,1674 (0,2805)	0,0456 (0,1734)		–0,9448 (0,5129)	–3,7091 (0,3322)	1,3672 (0,4983)
IV: Dif.Var.	–1,8458 (0,0482)	–0,9376 (0,0666)	–0,0411 (0,0686)		–1,8521 (0,0283)	–0,9912 (0,0381)	–0,0489 (0,0388)
IV: anon.Var.	–2,5017 (0,0560)	–2,5020 (0,1581)	3,4578 (66,6270)		–2,5001 (0,0329)	–2,5011 (0,0739)	–2,8006 (1,4403)
Höhne-Überl.:							
Naiv	–2,4450 (0,0110)	–2,3159 (0,0187)	–0,8269 (0,0467)		–2,4458 (0,0066)	–2,3525 (0,0100)	–1,2135 (0,0274)
IV: verz.Var.	–2,4345 (0,0510)	–1,2798 (5,1383)	0,7846 (0,3839)		–2,1640 (0,7830)	–2,5601 (0,0332)	–3,1477 (0,1477)
IV: Dif.Var.	–2,4419 (0,0166)	–2,2468 (0,0317)	-0,4942 (0,0681)		–2,4409 (0,0090)	–2,2626 (0,0169)	–0,5636 (0,0382)
IV: anon.Var.	–2,4994 (0,0119)	–2,4994 (0,0214)	–2,5132 (0,2468)		–2,5003 (0,0063)	–2,5004 (0,0107)	–2,5024 (0,0791)

Weiter fällt auf, dass das Instrument ‚anon.Var.' bei einer hohen Autokorrelation ($\rho = 0{,}9$) Probleme aufweist. Ein Extremfall tritt bei einfacher Überlagerung mit $\rho = 0{,}9$ und $T = 4$ auf. Der IV-Schätzer verändert das Vorzeichen und hat einen sehr hohen Wert der Standardabweichung (66,63). Die unzureichenden IV-Korrekturen gehen mit sehr hohen Standardabweichungen einher. Der Grund liegt darin, dass in einigen Monte-Carlo-Wiederholungen starke Ausreißer (vom wahren Parameter stark abweichende Schätzwerte) resultieren, die dementsprechend den in der Tabelle ausgewiesenen durchschnittlichen Schätzer stark verzerren. Diesem Problem wird in Abschnitt 6 nachgegangen.

Im folgenden Abschnitt wird zunächst ein empirisches Beispiel betrachtet.

5. Empirisches Beispiel

5.1 Datengrundlage

Die bisher hergeleiteten Ergebnisse zur Instrumentvariablen-Schätzung werden nun anhand der Daten des Monatsberichts für Betriebe im Verarbeitenden Gewerbe und Bergbau (als nicht balanciertes Panel für 1995 bis 2004) überprüft. Diese Statistik ist eine Vollerhebung und umfasst alle Industriebetriebe, die mindestens 20 Personen beschäftigen.[8] Die Analyse basiert auf dem Modell von Wagner (2007). Wagner untersucht auf Basis des Monatsberichts den Zusammenhang zwischen der Exporttätigkeit und der Produktivität bei deutschen Unternehmen.

Das empirische Untersuchungsmodel lautet:

$$\ln AP_{it} = \alpha_i + \beta\, expant_{it} + \boldsymbol{Kontroll}'_{it}\,\boldsymbol{\gamma} + \eta_{it}\,, \qquad i = 1,\ldots,N, \quad t = 1,\ldots,T\,. \tag{27}$$

ln AP bezeichnet die logarithmierte Arbeitsproduktivität. Die Exportvariable *expant* ist definiert als Exportanteil am Gesamtumsatz. Weitere Kontrollvariablen sind: *pers* (Beschäftigtenanzahl), *perssq* (quadrierte Beschäftigtenanzahl) und *hc* (Humankapitalintensität).

Zur Illustration werden alle stetigen Variablen mittels der stochastischen Überlagerung anonymisiert. Die Fehlervariablen haben den Erwartungswert Eins. Bei einfacher Überlagerung sind sie miteinander unkorreliert und haben die Varianz $0{,}114^2$. Die Fehler der Höhne-Variante erhalten die Parameter $\delta = 0{,}11, \sigma_\varepsilon = 0{,}03$. Das Instrument ‚anon.Var.' wird mit Varianz $0{,}36^2$ (einfache Überlagerung) und $\delta = 0{,}2, \sigma_\varepsilon = 0{,}3$ generiert.[9]

5.2 Ergebnisse

Tabelle 4 zeigt nur die Ergebnisse für Modelle ohne Ausreißer für Westdeutschland. Da andere Varianten (Modell für Ostdeutschland, Modelle mit Ausreißern) zu ähnlichen Ergebnissen führen, werden sie hier nicht berichtet. Folgend Wagner (2007) sind unter Außreißern Beobachtungen gemeint, deren Arbeitsproduktivität unter dem 1 %-Quantil der Arbeitsproduktivitätsverteilung liegt oder das 99 %-Quantil übersteigt.

Die IV-Schätzung mit Verwendung der verzögerten Variable bzw. der Differenzvariable als Instrument führt zu starken Verzerrungen. Auffällig ist aber

[8] 2007 wurde die Abschneidegrenze auf 50 Beschäftigte angehoben.

[9] Aus Datenschutzaspekten werden hier im Vergleich zur Simulationsstudie für das Instrument höhere Überlagerungsparameter verwendet.

auch, dass die Schätzer bei dem Instrument ‚anon.Var.' bei einigen Variablen ebenfalls stark verzerrt sind, insbesondere bei *pers* und *perssq*. Bei diesen problematischen Variablen verändert sich in einigen Fällen die Richtung der Verzerrung und die Signifikanz der IV-Schätzer. Der Grund für diese in Widerspruch zur Theorie stehenden empirischen Ergebnisse wird im nächsten Abschnitt näher untersucht. (Tabelle 4 S. 19)

6. Probleme bei der konsistenten Schätzung

Der Wahrscheinlichkeitsgrenzwert des IV-Schätzers im Fall der zusätzlichen anonymisierten Variable als Instrument zeigt, dass der IV-Schätzer konsistent ist (vgl. Gl. (20) und (22)). Simulationsexperimente und Studien mit empirischen Daten haben allerdings gezeigt, dass bei diesem Instrument die IV-Schätzer – auch bei großem Stichprobenumfang – verzerrt sein können. Eine Erklärung dieses Phänomens liegt darin begründet, dass – trotz der Gleichheit des Zählers und Nenners im Grenzwert des IV-Schätzers und der daraus resultierenden Konsistenz – der Nenner bei einigen Parameterkonstellation Werte nahe bei Null annimmt. Dies führt dazu, dass der gesamte Ausdruck „explodieren" und zu unverhältnismäßig großen Schätzwerten führen kann.

Zur Verdeutlichung betrachten wir den Erwartungswert des Nenners für diesen Schätzer im Fall eines AR(1)-Prozesses. Wegen $\gamma_j = \gamma_0 \rho^j$ kann man schreiben (vgl. Gl. (20)):[10]

$$\gamma_0 \left((T-1) - \frac{2}{T^2} \sum_{j=1}^{T-1} (T-j)\rho^j \right) . \tag{28}$$

Für verschiedene Werte von ρ wird die Funktion (28) in Abbildung 1 für $T = 2$ bis $T = 100$ dargestellt. Da γ_0 sehr unterschiedlich sein kann, wird dieser als identisch 1 unterstellt. In der ersten Grafik sind die Kurven für unterschiedliche Werte der *positiven* Autokorrelation und in der zweiten Grafik für *negative* Autokorrelation zu sehen. Die Kurvenverläufe sind in allen Fällen sehr ähnlich. Der simulierte Term liegt für kleines T näher bei Null. Mit steigender Anzahl der Zeitperioden entfernt er sich von Null. Im Fall der positiven Autokorrelation verschiebt sich die in der Grafik abgebildete Funktion nach unten (Richtung Nullwert), wenn ρ größer wird. Die Werte für $\rho = 0{,}9$ und kleines T liegen besonders kritisch nahe bei Null. Dies bedeutet, dass der Nennerausdruck, der um diesen Erwartungswert herum variiert, für kleines T und großes (positives) ρ Werte nahe bei Null annehmen kann und auch die Umkehr des geschätzten Vorzeichens ergeben kann.

[10] Ähnliche Implikationen ergeben sich auch für die Höhne-Überlagerung.

Tabelle 4

Empirische Studie: Westdeutschland ohne Ausreißer

Einfache Überlagerung:

	Original		Anonymisiert		IV verz.Var.		IV Dif.Var.		IV anon.Var.	
	$\widehat{\beta_{FE}}$	t-Wert	$\widehat{\beta^{a}_{FE}}$	t-Wert	$\widehat{\beta^{IV}_{FE}}$	t-Wert	$\widehat{\beta^{IV}_{FE}}$	t-Wert	$\widehat{\beta^{IV}_{FE}}$	t-Wert
expant	0,003	48,43	0,003	44,59	0,002	14,83	0,002	22,27	0,003	32,49
pers	–0,046	–10,23	–0,015	–5,28	–0,089	–6,52	0,009	1,58	–0,021	–1,40
perssq	0,106	9,01	0,024	3,45	0,160	5,09	0,009	0,64	0,016	0,49
hc	0,165	139,57	0,156	130,82	0,259	22,82	0,150	22,36	0,168	110,08

Höhne-Überlagerung:

	Original		Anonymisiert		IV verz.Var.		IV Dif.Var.		IV anon.Var.	
	$\widehat{\beta_{FE}}$	t-Wert	$\widehat{\beta^{a}_{FE}}$	t-Wert	$\widehat{\beta^{IV}_{FE}}$	t-Wert	$\widehat{\beta^{IV}_{FE}}$	t-Wert	$\widehat{\beta^{IV}_{FE}}$	t-Wert
expant	0,003	48,43	0,003	48,04	0,002	16,41	0,003	24,67	0,003	35,86
pers	–0,046	–10,23	–0,037	–8,93	–0,064	–9,48	0,000	0,03	–0,029	–2,31
perssq	0,106	9,01	0,083	7,84	0,111	7,71	0,073	2,08	0,054	1,92
hc	0,165	139,57	0,163	137,99	0,261	24,02	0,160	23,13	0,161	115,44

Im Fall der negativen Autokorrelation verhält es sich genau umgekehrt. Am weitesten vom Nullpunkt entfernt ist die Kurve für $\rho = -0{,}9$. Die am nächsten bei Null liegende Kurve ergibt sich für $\rho = -0{,}1$. Allerdings ist anzumerken, dass die Situation bei negativer Autokorrelation nicht so problematisch erscheint, da der kleinste Wert weiter von Null entfernt liegt als bei positivem ρ.

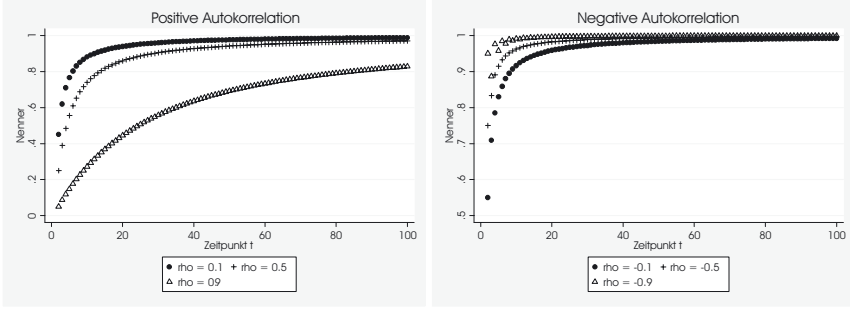

Abbildung 1: Erwartungswert des Nenners des IV-Schätzers – siehe Gl. (28)

Weiter wird eine empirische Untersuchung mit Daten des Monatsberichts[11] durchgeführt, wobei dieses Mal der Datensatz balanciert wird, d. h. es werden nur die Unternehmen berücksichtigt, die in allen Jahren keine fehlenden Angaben aufweisen. Dann wird die IV-Schätzung mit der zusätzlichen anonymisierten Variable als Instrument durchgeführt, wobei die Anonymisierung 2.500 mal erfolgt. Die Schätzer, die in Tabelle 5 ausgewiesen werden, sind die Durchschnitte der Schätzer über alle 2.500 Monte-Carlo-Wiederholungen. Das Instrument wird mit den Überlagerungsparametern $\sigma_u = 0{,}539$ (einfache Überlagerung), $\delta = 0{,}5$ und $\sigma_\epsilon = 0{,}2$ (Höhne-Überlagerung) erzeugt.[12]

Tabelle 5

Zusätzliche anonymisierte Variable; Westdeutschland ohne Ausreißer

		Naiv		IV	
	Original	**Einf.**	**Höhne**	**Einf.**	**Höhne**
expant	0,025	0,022	0,024	0,025	0,025
pers	–0,602	–0,244	–0,546	–1,384	–0,406
perssq	1,281	0,429	1,147	3,134	0,713
hc	1,666	1,589	1,661	1,662	1,667

[11] Nur das Modell für Westdeutschland ohne Ausreißer wird geschätzt.

[12] Hier werden die Ergebnisse für eine sehr hohe Anonymisierung dargestellt. Die angesprochenen Probleme resultieren aber auch bei kleineren Anonymisierungsparametern.

Tabelle 5 zeigt die Ergebnisse. Zwar gelingt die Korrektur bei zwei Variablen *expant* und *hc* gut, bei *pers* und *perssq* schneidet die naive Schätzung jedoch wesentlich besser ab. Die minimalen und maximalen Werte der einzelnen IV-Schätzer über alle Monte-Carlo-Wiederholungen zeigt Tabelle 6.

Tabelle 6

Zusätzliche anonymisierte Variable: Ergänzung

	Einfach		**Höhne**	
	Minimum	**Maximum**	**Minimum**	**Maximum**
expant	–0,217	1,850	–0,252	0,072
pers	–2450,459	329,722	–70,836	468,408
perssq	–614,619	5275,269	–1272,714	169,113
hc	–8,054	2,548	1,217	4,541

Sowohl bei der einfachen als auch bei der Höhne-Überlagerung resultieren in einzelnen Monte-Carlo-Wiederholungen IV-Schätzer, die sehr stark vom wahren Wert abweichen. Diese Situation tritt insbesondere bei *pers* und *perssq* auf, was sich in schlechten mittleren IV-Schätzern widerspiegelt. Dieses deutet darauf hin, dass der Nenner des IV-Schätzers in diesen Fällen kritisch nahe bei Null liegt.

Einer der beiden (anonymen) Gutachter wies darauf hin, dass möglicherweise die ‚Schwäche' der verwendeten Instrumente bei den aufgetretenen Problemen eine Rolle spielt. In der aktuellen Fassung haben wir die Korrelation zwischen Regressor *x* und Instrument *z* beispielhaft für das Instrument ‚anon. Var.' dargestellt, für das sich gemäß Theorie eine konsistente Schätzung ergeben sollte (siehe Abschnitt 4.1). Zwar nimmt mit wachsendem ρ die Korrelation vor allem im Fall der einfachen Überlagerung stark ab (siehe Tabelle 1), jedoch erachten wir auch Korrelationen in der Größenordnung von 0,25 für $\rho = 0{,}9$ als nicht zu ‚schwach'. Insofern ist die Schwäche des Instruments als Ursache auszuschließen. Abschließend sei noch darauf hingewiesen, dass die durch einen Nenner des Schätzers nahe Null ausgelöste Problematik natürlich auch die Höhne-Überlagerung betrifft. Allerdings ist sie bei der in der Simulationsstudie verwendeten Parameterkonstellation nicht relevant. Siehe auch Ronning (2009, Abschnitte 12.4.4 und 12.5.3).

7. Zusammenfassung

Der vorliegende Beitrag untersuchte die Anwendbarkeit der Instrumentvariablen-Methode als Korrekturverfahren im Fall der multiplikativen stochastischen Überlagerung. Dabei wurden als Instrumente die verzögerte anonymisierte Variable, die Differenz von verzögerten anonymisierten Variablen und eine zusätzliche anonymisierte Variable verwendet.

Es zeigte sich, dass lediglich das Instrument „zusätzliche anonymisierte Variable“ zu konsistenten IV-Schätzern führt. Diese Ergebnisse wurden in einer Simulationsstudie und einer empirischen Studie mit Daten des Monatsberichts für Verarbeitendes Gewerbe (Panel für 1995 – 2005) illustriert. Allerdings zeigte sich im empirischen Beispiel, dass der konsistente IV-Schätzer mit der zusätzlichen anonymisierten Variable als Instrument manchmal verzerrte Ergebnisse erzeugt. Der Grund dafür sind numerische Probleme, die bei kleiner Anzahl der Zeitperioden und einer hohen Autokorrelation entstehen können. Da empirische Ergebnisse oft zu sehr großen und somit unplausiblen Schätzern führen, könnte eine Prüfung auf die Plausibilität der Ergebnisse dem Forscher / der Forscherin einen Hinweis auf die numerische Probleme geben.

Im Gegensatz zu echten Messfehlern kann die Auswahl der Überlagerungsparameter im Fall der Anonymisierung direkt kontrolliert werden. Dadurch kann aber auch die Schätzverzerrung beeinflusst werden. Durch weitere Optimierung der Anonymisierungsparameter – unter Beachtung der Sicherheitsakpekte – könnte die Verzerrung noch mehr reduziert werden. Dieses Ziel lässt sich bei der Überlagerung mit Faktorstruktur besser als bei einfacher Überlagerung erreichen, da die Verzerrung aufgrund der Konstruktion der Fehlervariable hier kleiner ausfällt.

Literatur

Biewen, E. (2008): Within-Schätzung bei anonymisierten Paneldaten, AStA Wirtschafts- und Sozialstatistisches Archiv 2(3), 277 – 297.

Biewen, E. / *Nolte,* S. / *Rosemann,* M. (2008): Perturbation by Multiplicative Noise and the Simulation Extrapolation Method, AStA Advances in Statistical Analysis 92(4), 391 – 404.

Biewen, E. / *Ronning,* G. (2008): Estimation of Linear Models with Anonymised Panel Data, AStA Advances in Statistical Analysis 92(4), 423 – 438.

Biewen, E. / *Ronning,* G. / *Rosemann,* M. (2009): IV-Schätzung eines linearen Panelmodells mit stochastisch überlagerten Betriebs- und Unternehmensdaten, IAW-Diskussionspapier 53.

Biørn, E. (1992): The Bias of Some Estimators for Panel Data Models with Measurement Errors, Empirical Economics 17, 51 – 66.

Biørn, E. (1996): Panel Data with Measurement Errors, in: L. Mátyás / P. Sevestre (eds.), The Econometrics of Panel Data, Dordrecht / Boston / London.

Biørn, E. / *Klette,* T. (1998): Panel Data with Errors-in-Variables: Essential and Redundant Orthogonality Conditions in GMM-Estimation, Economics Letters 59, 275 – 282.

Biørn, E. / *Krishnakumar,* J. (2008): Measurement Errors and Simultaneity, in: L. Mátyás / P. Sevestre (eds.), The Econometrics of Panel Data. Fundamentals and Recent Developments in Theory and Practice, Springer-Verlag, N.Y., 323 – 367.

Bound, J. / *Jaeger,* D. / *Baker,* R. (1995): Problems with Instrumental Variables Estimation When the Correlation Between the Instruments and the Endogeneous Explanatory Variable is Weak, Journal of the American Statistical Association 90 (430), 443 – 450.

Carroll, R. J. / *Ruppert,* D. / *Stefanski,* L. A. (1995): Measurement Error in Nonlinear Models, London.

Cook, J. / *Stefanski,* L. A. (1994): Simulation-Extrapolation Estimation in Parametric Measurement Error Models, Journal of the American Statistical Association 89 (428), 1314 – 1328.

Greene, W. H. (2008): Econometric Analysis, New Jersey.

Griliches, Z. / *Hausman,* J. A. (1986): Errors in Variables in Panel Data, Journal of Econometrics 31, 93 – 118.

Höhne, J. (2008): Anonymisierungsverfahren für Paneldaten, AStA Wirtschafts- und Sozialstatistisches Archiv 2 (3), 259 – 276.

Hwang, J. T. (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.

Lenz, R. / *Zwick,* M. (2009): Business Microdata in Germany: Linkage and Anonymisation, Schmollers Jahrbuch 129 (4).

Nolte, S. (2007): The Multiplicative Simulation Extrapolation Approach, University of Konstanz, Working Paper.

Ronning, G. (2009): Stochastische Überlagerung mit Hilfe der Mischungsverteilung. Schätzung linearer (Panel-)Modelle auf Basis anonymisierter Daten, IAW-Diskussionspapier 48. http: // www.iaw.edu / RePEc / iaw / pdf / iaw_dp_48.pdf.

Ronning, G. / *Sturm,* R. / *Hoehne,* J. / *Lenz,* R. / *Rosemann,* M. / *Scheffler,* M. / *Vorgrimler,* D. (2005): Handbuch zur Anonymisierung wirtschaftsstatistischer Mikrodaten, Statistisches Bundesamt: Statistik und Wissenschaft, Band 4, Wiesbaden.

Roque, G. (2000): Masking Microdata Files with Mixtures of Multivariate Normal Distribution, Unpublished Ph. D. dissertation, Department of Statistics, University of California-Riverside.

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

Schneeweiß, H., / *Ronning*, G. (2010): Multiple Linear Panel Regression with Multiplicative Random Noise, Statistical Modelling and Regression Structures. Festschrift in Honour of Ludwig Fahrmeier (Eds.) Th.Kneib, G. Tutz, 399 – 417.

Sevestre, P. / *Trognon*, A. (1996): Linear Models with Random Regressors, in: L. Mátyás / P. Sevestre (eds.), The Econometrics of Panel Data, Dordrecht / Boston / London.

Staiger, D. / *Stock*, J. (1995): Instrumental Variables Regression with Weak Instruments, Econometrica, 65 (3), 557 – 586.

Stefanski, L. (1985): The Effects of Measurement Error on Parameter Estimation, Biometrica 72, 583 – 592.

Wagner, J. (2007): Exports and Productivity in Germany, Applied Economics Quarterly 53 (4), 353 – 373.

Wansbeek, T. (2001): GMM Estimation in Panel Data Models with Measurement Error, Journal of Econometrics 104, 259 – 268.

Wansbeek, T. / *Koning*, R. (1991): Measurement Error and Panel Data, Statistica Nederlandia 45, 85 – 92.