Schätzung ökonometrischer Modelle auf der Grundlage anonymisierter Daten

Sandra Lechner
Universität Konstanz

Winfried Pohlmeier*
Universität Konstanz
CoFE, ZEW

April 2003

Zusammenfassung

Die Anonymisierung von sensiblen Individualdaten führt zu einem Konflikt zwischen dem Ziel der Minimierung des Reidentifikationsrisikos und der Qualität ökonometrischer Schätzungen. Der durch Anonymisierung bedingte Verlust an Effizienz und/oder der Konsistenz eines Schätzers wirft die grundsätzliche Frage auf, inwieweit anonymisierte Individualdaten überhaupt für die wissenschaftliche Nutzung geeignet sind.

Deshalb gehen wir in dieser Arbeit der Frage nach, welchen Einfluss Anonymisierungsverfahren auf die Eigenschaften von ökonometrischen Schätzern haben. Zunächst untersuchen wir die Auswirkungen gängiger Anonymisierungsverfahren auf lineare ökonometrische Schätzer in endlichen Stichproben. Im zweiten Schritt untersuchen wir, inwieweit sich die Selektionseffekte durch Anonymisierung aufgrund von Data Blanking mit Hilfe von semiparametrischen Verfahren korrigieren lassen. Die quantitative Evidenz beruht auf Monte-Carlo Simulationen und einer illustrativen Anwendung für einen Querschnitt der Kostenstrukturerhebung.

JEL Klassifikation: C81, C21, C24, C25

Schlüsselwörter: Mikroaggregation, stochastische Überlagerung, Data Blanking, IV-Schätzung, semiparametrisches Selektionsmodell

*Korrespondierender Autor: Fachbereich Wirtschaftswissenschaften, Fach D 124, Universität Konstanz, 78457 Konstanz, Tel.: +49-7531-88-2660, Fax.: 88-4450, e-mail: winfried.pohlmeier@uni-konstanz.de. Eine vorläufige Version dieses Papers wurde auf der Nutzertagung 'Faktische Anonymisierung wirtschaftsstatistischer Einzeldaten' des Statistischen Bundesamtes und des IAW, 20./21. März 2003 in Tübingen vorgetragen. Unser Dank gilt Gerd Ronning für hilfreiche Kommentare und Anregungen bei der Erstellung dieser Arbeit sowie für die organisatorische Unterstützung durch das IAW. Für die finanzielle Förderung bedanken wir uns bei der Deutschen Forschungsgemeinschaft.

1. Einleitung

In den letzten dreißig Jahren ist die Nachfrage nach Mikrodaten durch die empirische Wirtschaftsforschung stark angestiegen. Diese Nachfrage, die sich ursprünglich auf Haushalts- und Personaldaten bezog, erweiterte sich rasch auf Firmendaten. Individualdaten und hier insbesondere Firmendaten beinhalten oftmals sensible Informationen, deren Vertraulichkeit im Interesse der Beobachtungseinheit, aber auch im Interesse der datenerhebenden Institution und der Datennutzer es zu schützen gilt. Damit stehen die datenerhebenden Institutionen vor einem Konflikt zwischen dem Ziel der Gewährleistung einer maximalen Vertraulichkeit der Daten und dem Ziel der Weitergabe maximaler Information.

Um die Möglichkeit der Reidentifikation individueller Angaben aus Mikrodaten zu minimieren und die von der datenerhebenden Institution gemachte Vertraulichkeitszusage zu garantieren, werden in der Praxis unterschiedliche Anonymisierungverfahren verwendet[1)], die sich im Ausmaß der Anonymisierung und ihres Effektes auf die Effizienz und die Konsistenz des verwendeten ökonometrischen Schätzverfahrens unterscheiden. Im allgemeinen stellt ein Anonymisierungsverfahren nichts anderes als einen Datenfilter dar, der den wahren datengenerierenden Prozess verändert. Für den empirischen Wirtschaftsforscher ergibt sich hieraus die Frage, inwieweit sich der wahre datengenerierende Prozess auf der Grundlage der anonymisierten (gefilterten) Daten schätzen lässt. Letztlich stellt die Anonymisierung von Individualdaten den Nutzer vor die grundsätzliche Frage, wie erheblich der durch die Anonymisierung bedingte Verlust an Information ist und unter welchen Umständen überhaupt konsistente Schätzungen des wahren datengenerierenden Prozesses möglich sind. Selbst wenn ein gegebenes Anonymisierungsverfahren nicht zum Verlust der Konsistenzeigenschaft des Schätzverfahrens führt, stellt sich die Frage der Relevanz der erzielten empirischen Ergebnisse, denn statistisch als insignifikant gefundene Zusammenhänge mögen schlichtweg das Resultat des Informationsverlustes durch Anonymisierung sein. Ist der durch Anonymisierung bedingte Informationsverlust erheblich, wird der anonymisierte Datensatz für den mit statistischen Inferenzmethoden arbeitenden wissenschaftlichen Nutzer unbrauchbar.

In dieser Arbeit untersuchen wir deshalb den Einfluss von Anonymisierungsverfahren auf die Eigenschaften von ökonometrischen Schätzern. Im Mittelpunkt unseres Interesses stehen insbesondere Auswirkungen von Anonymisierungsverfahren auf die Eigenschaften ökonometrischer Schätzer in endlichen Stichproben. Anhand von Monte-Carlo Simulationen sollen dabei die Auswirkung der Anonymisierungsverfahren auf die ökonometrische Schätzung quantifiziert werden.

Die Arbeit ist wie folgt aufgebaut. In Abschnitt 2 arbeiten wir die Konsequenzen der faktischen Anonymisierung durch Mikroaggregation und der stochastische Überlagerungen für die Schätzung des linearen Regressionsmodells heraus.[2)] Wir zeigen anhand einer Monte-Carlo-Studie, dass selbst im einfachen Fall des linearen Modells Anonymisierung nicht unproblematisch ist und den Nutzen für den Anwender erheblich einschränken kann. In Abschnitt 3 stellen wir einen zweistufigen semiparametrischen Selektionsschätzer vor, der der Selektionsver-

---

1) siehe z. B. Gottschalk (2002) für eine Übersicht

2) Weitere Verfahren jenseits der faktischen Anonymisierung werden in Brand (2000) vorgestellt.

zerrung durch Data Blanking oder partieller Aggregation Rechnung trägt. Dieses Verfahren beruht auf dem semiparametrischen Schätzer von Klein und Spady (1993) für binäre Auswahlmodelle in der ersten Stufe und dem semiparametrischen Reihenapproximationsschätzer von Newey (1999). Abschnitt 4 illustriert am Beispiel einer Regression, basierend auf den Daten der Kostenstrukturerhebung des Statistischen Bundesamtes, inwieweit sich Schätzergebnisse aufgrund der verwendeten Anonymisierungsmethode im Vergleich zu einer Analyse auf Basis der Originaldaten ändern. Abschnitt 5 gibt einen Ausblick auf die zukünftige Forschung.

2. Klassische Anonymisierungsverfahren: Einige Konsequenzen

Um den Effekt der Anonymisierung auf die Eigenschaften des KQ-Schätzers zu bewerten, gehen wir vom linearen Regressionsmodell unter vollen idealen Bedingungen aus:

$$Y = X\beta + \varepsilon, \tag{2.1}$$

mit

$$E[\varepsilon] = 0,$$
$$V[\varepsilon] = \sigma^2 I_N,$$
$$\text{plim}\,\frac{1}{\mathrm{N}} X'X = \Sigma_{XX},$$

wobei X eine $N \times K-$Regressionsmatrix fester erklärender Variablen und Y der $N \times 1$-Vektor der zu erklärenden Variablen ist. Das idealtypische Design für die Originaldaten erlaubt uns, die Auswirkungen des Anonymisierungsverfahrens auf die stochastischen Eigenschaften verschiedener Schätzer gegenüber dem Idealfall des besten linearen unverzerrten Schätzers zu vergleichen.

Mikroaggregation im linearen Modell

Bei der Mikroaggregation werden die Variablenausprägungen durch vorher ermittelte Mittelwerte von jeweils ähnlichen Datensätzen ersetzt (Paaß and Wauschkuhn, 1984). Hier sei nur der Fall der listenweisen Aggregation betrachtet, bei der jeweils die Variablen von A Beobachtungen zu entsprechenden Gruppenmittelwerten zusammengefasst werden. In diesem Fall ergeben sich $M = N / A$ unterschiedliche (aggregierte) Beobachtungen. Zur Vereinfachung der Notation sei angenommen, dass M ganzzahlig ist. Üblicherweise werden in der Praxis $A = 3, 4$ oder 5 Beobachtungen zu einer aggregierten Beobachtung zusammengefasst.

Da von einer Zufallsstichprobe unabhängig und identisch verteilter Beobachtungen ausgegangen wird, kann zur Vereinfachung der Notation ohne Verlust der Allgemeingültigkeit angenommen werden, dass die Mikroaggregation gemäß der Reihenfolge der Beobachtungen im Datensatz erfolgt. Hierzu sei die $N \times N-$blockdiagonale Matrix

$$D = I_M \otimes \frac{1}{A} \iota\iota' \tag{2.2}$$

definiert, wobei ι ein A-dimensionaler Vektor von Einsen ist. Das lineare Regressionsmodell auf der Grundlage der mikroaggregierten Daten ergibt sich durch Prämultiplikation von D mit dem auf den Originaldaten basierenden Modell (2.1):

$$Y^* = X^* \beta + \varepsilon^* \tag{2.3}$$

mit $Y^* = DX, \quad X^* = DX$ und $\varepsilon^* = D\varepsilon.$

Der gewöhnliche KQ-Schätzer für das mikroaggregierte Modell $\hat{\beta}_A$ hat die Form

$$\hat{\beta}_A = \left(X^{*\prime} X^*\right)^{-1} X^{*\prime} Y^* = \left(XDX\right)^{-1} X'DY, \tag{2.4}$$

wobei zur Berechnung des rechten Terms in (2.4) die Symmetrie und Idempotenz von D verwendet wurde. Offensichtlich bleibt durch die listenweise Mikroaggregation die Unverzerrtheit des KQ-Schätzers erhalten. Mit dem exogenen Datenfilter D verwenden wir hier das denkbar einfachste Aggregationsschema mit gleicher Gewichtung über alle Beobachtungen und gleichem Aggregationsniveau für alle Gruppen. Ein Aggregationsschema, basierend auf den exogenen Variablen $D = D(X)$, stellt nur eine unwesentliche Erweiterung dar. Wenn jedoch das Gewichtungsschema der Aggregation von der abhängigen Variablen abhängt, $D = D(Y, X)$, ist der KQ-Schätzer für die aggregierten Daten nichtlinear mit unbekannten Verteilungseigenschaften.

Schon die einfache exogene Aggregation führt jedoch zu einem Informationsverlust, so dass der KQ-Schätzer auf Grundlage der anonymisierten Daten gegenüber dem gewöhnlichen KQ-Schätzer $\hat{\beta}$ an Effizienz verliert (Beweis siehe Anhang):

$$V\left[\hat{\beta}_A\right] - V\left[\hat{\beta}\right] > 0,$$

wobei das Ungleichheitszeichen für die positive Definitheit der Differenz der beiden Varianz-Kovarianzmatrizen steht. Der durch Aggregation bedingte Effizienzverlust kann für den Fall $k = 1$ leicht verdeutlicht werden. Es sei

$$X = \left(X_1, X_2, \ldots, X_N\right)' = \left(X_{11}, X_{21}, \ldots, X_{A1}, X_{12}, X_{22}, \ldots, X_{A2}, \ldots, X_{AM}\right)'$$

der Vektor der erklärenden Variablen, wobei die Doppelindizierung a,m die Beobachtung a in Gruppe m bezeichnet. Der Vergleich der Präzisionen beider Schätzer

$$V\left[\hat{\beta}\right]^{-1} - V\left[\hat{\beta}_A\right]^{-1} = \frac{1}{\sigma^2}\sum_{a=1}^{A}\sum_{m=1}^{M}\left(X_{am} - \overline{X}_m\right)^2$$

zeigt, dass der Effizienzverlust durch Aggregation besonders klein ist, wenn die Aggregation über möglichst homogene Gruppenmitglieder erfolgt, d.h. wenn die Variation innerhalb der Gruppen (within group variation) gegen null geht.

Die Mikroaggregation führt zu einer Verzerrung des herkömmlichen Schätzers für die Varianz des Fehlerterms und somit der Standardfehler des KQ-Schätzers von β. Sofern das Aggregationsniveau bekannt ist, ergibt sich ein unverzerrter Schätzer für σ^2 wie folgt:

$$\sigma^2 = \frac{1}{M-K} e^{*'} e^{*},$$

wobei $e^{*'} e^{*}$ die Summe der quadrierten Fehler des KQ-Schätzers auf der Grundlage der aggregierten Beobachtungen ist (Beweis siehe Anhang). Da $M < N$, führt eine Ignorierung der wahren Freiheitsgrade des Modells zu einer Unterschätzung der Standardfehler, so dass die auf der aggregierten Datenbasis erzielten t-Werte des KQ-Schätzers überhöht ausgewiesen werden.

Bootstrap-Aggregation

Bei der einfachen Mikoaggregation erscheint jede anonymisierte Beobachtungseinheit $A-$mal im Datensatz. Alternativ kann jedoch auch für jede Beobachtung i des Originaldatensatzes per Zufallsziehung mit Zurücklegen eine (möglichst homogene) Gruppe i zusammengestellt werden und die Mittelwerte der Kovariate dieser Gruppe i als anonymisierte Beobachtungseinheit verwendet werden. Die Idee für diese Art von Mikroaggregation hat gewisse Ähnlichkeiten mit dem Bootstrap-Verfahren, da durch Ziehung aus der Stichprobe künstlich neue Datensätze gezogen werden, über die dann aggregiert wird. Die Struktur des gewöhnlichen KQ-Schätzers auf Grundlage einer Bootstrap-Aggregation ist äquivalent zum Schätzer $\hat{\beta}_A$. Jedoch ist die $N \times N-$Aggregationsmatrix D bei der Bootstrap-Aggregation eine Zufallsmatrix der Form

$$D = \frac{1}{B}\left(I_n + S_1 + S_2 + \ldots + S_{B-1}\right), \qquad (2.5)$$

S_b stellt hierbei eine $N \times N$ Selektionsmatrix dar, die jeweils in einer Zeile an einer zufällig ausgewählten Position eine Eins und sonst Nullen enthält. Prämultiplikation des Originalmo-

dells (2.1) mit D liefert das lineare Modell auf der Grundlage von N verschiedenen Gruppenmittelwerten. Da D nun eine Zufallsmatrix ist, haben wir es trotz fester $X-$Variablen mit einem Modell mit stochastischen Regressoren zu tun.

$$V\left[\hat{\beta}_B\right] = E\,V\left[\hat{\beta}_B|D\right] = \sigma^2 E\left[(X'DX)^{-1}\right] \tag{2.6}$$

Gegenüber der einfachen Mikroaggregation wird das Reidentifikationssrisiko durch die Bootstrap-Aggregation weiter verringert, da aus der zufälligen Aggregation die Wahrscheinlichkeit, eine korrekte Schlussfolgerung über die Originaldaten zu ziehen, weiter reduziert wird. Die Wahrscheinlichkeit, dass eine aggregierte Beobachtung i genau der Originalbeobachtung entspricht, beträgt N^{-B}.

Für die Standardfehler sollte bei einer Bootstrap-Aggregation der heteroskedastie-robuste Varianz-Kovarianzmatrix-Schätzer verwendet werden, denn Regressoren bei einer Bootstrap-Aggregation können als gewogenes Mittel aus Originalbeobachtung und arithmetischem Mittel über alle Beobachtungen mit Gewichtungsfaktor 1/B und 1-1/B formuliert werden, wobei die Stichprobenvariation über einen heteroskedastischen Fehlerterm aufgefangen wird. Ersetzt man nämlich den stochastischen Aggregationsfilter D durch seinen Erwartungswert und einer zufälligen Abweichung ζ mit Erwartung $E[\zeta] = 0$,

$$D = E[D] + \zeta,$$

ergibt sich aus dem bootstrap-aggregierten Regressionsmodell

$$Y^* = E[D]X\beta + \omega,$$

wobei $\omega = \zeta X\beta + \varepsilon$ ein heteroskedastischer Fehlerterm ist. Die Regressormatrix $E[D]X$ ist das gewogene Mittel aus Originalbeobachtung und arithmetischem Mittel über alle Beobachtungen.

Stochastische Überlagerung

Als Alternative zur Mikroaggregation wird oftmals die stochastische Überlagerung verwendet. Dieses Verfahren ist besonders bei Paneldatensätzen attraktiv, wenn die stochastische Überlagerung multiplikativ und zeitlich konstant ist. Das loglineare Modell ist in diesem Fall nur durch einen stochastischen Individualeffekt vom loglinearen Modell auf Basis der Originaldaten verschieden. Differenzbildung oder Within-Transformation des loglinearen Modells beseitigen den Einfluss der multiplikativen stochastischen Überlagerung. Allerdings setzt dieses Verfahren voraus, dass der wahre datengenerierende Prozess tatsächlich loglinear ist. Die Überprüfung der funktionalen Form der Erwartungswertfunktion mit Hilfe von Tests auf funktionale Form ist nicht mehr trivial, weil entsprechende Annahmen über Art der stochastischen Überlagerung als beizubehaltende Hypothese berücksichtigt werden müssen (z.B. Null-

hypothese: wahres Modell ist linear, beizubehaltende Hypothese: stochastische Überlagerung ist multiplikativ).

Im Folgenden sei von einer additiven stochastischen Überlagerung oder einem log-linearen Modell mit multiplikativer stochastischer Überlagerung ausgegangen. Zur abhängigen Variablen und zum Vektor der erklärenden Variablen werden unabhängig identisch verteilte Störgrößen hinzu addiert

$$\begin{aligned} Y_i^* &= Y_i + v_i, \\ X_i^* &= X_i + u_i, \end{aligned} \tag{2.7}$$

so dass das verfügbare Modell auf der Grundlage der stochastisch überlagerten Beobachtungen die Form

$$Y^* = X^*\beta + \omega \tag{2.8}$$

mit $\omega = \varepsilon + v - u\beta$ annimmt. Stochastische Überlagerung führt zu einem klassischen Fehler-in-den-Variablen-Modell. Aufgrund der stochastischen Überlagerung sind Fehlerterm und Regressoren miteinander korreliert, so dass der gewöhnliche KQ-Schätzer inkonsistent ist:

$$\operatorname{plim}\hat{\beta}_{EIV} = \left(I_K - \left(Q + \Sigma_{uu}\right)^{-1}\Sigma_{uu}\right)\beta \neq \beta, \tag{2.9}$$

wobei $\Sigma_{uu} = E\left[u_i u_i'\right]$ die Varianz-Kovarianz-Matrix des Fehlertermvektors u_i bezeichnet. Definieren wir $\kappa_{XX} = \left(\operatorname{plim}\frac{1}{N}X^{*'}X^*\right)^{-1}\operatorname{plim}\frac{1}{N}X'X = \left(Q + \Sigma_{uu}\right)^{-1}Q$ als die Zuverlässigkeitsmatrix im Sinne einer multivariaten Erweiterung des Zuverlässigkeitskoeffizienten (reliability ratio) von Fuller (1987, S. 3), erhalten wir

$$\operatorname{plim}\hat{\beta}_{EIV} = \kappa_{XX}\beta. \tag{2.10}$$

Anders als beim Fehler-in-den-Variablen-Modell ist jedoch hier der datengenerierende Prozess bekannt, so dass die asymptotische Verzerrung des KQ-Schätzers $\hat{\beta}_{EIV}$ leicht korrigiert werden kann, sofern Q und Σ_{uu} bzw. κ_{XX} bekannt sind. Der korrigierte unverzerrte KQ-Schätzer $\hat{\beta}_{CEIV}$ weist die Form

$$\hat{\beta}_{CEIV} = \left(I_K - \left(Q + \Sigma_{uu}\right)^{-1}\Sigma_{uu}\right)^{-1}\hat{\beta}_{EIV} = \kappa_{XX}^{-1}\hat{\beta}_{EIV} \tag{2.11}$$

auf. In der Praxis könnte dieser korrigierte Fehler-in-der-Variablen-Schätzer ohne großen Aufwand für die datenerhebende Institution und ohne Erhöhung des Reidentifikationsrisikos implementiert werden. Als einzige zusätzliche Information müsste dem Datennutzer die Kovarianzmatrix $\sum_{uu}$ bereitgestellt werden. Da das Reidentifikationsrisiko nicht unbedingt mit der Annahme unkorrelierter Anonymisierungsstörgrößen steigt, kann Unkorreliertheit vorausgesetzt werden, so dass die Information über die Varianzen der Störgrößen ausreicht. Ein konsistenter Schätzer des Terms $Q + \sum_{uu}$ ist die empirische Momentenmatrix der anonymisierten Regressoren

$$\text{plim} \frac{1}{N} X^{*'} X^{*} = Q + \sum_{uu},$$

so dass ein verfügbarer korrigierter Fehler-in-dem-Variablen-Schätzer $\widetilde{\beta}_{CEIV}$ die Form

$$\widetilde{\beta}_{CEIV} = \left(I_k - \left(\tfrac{1}{N} X^{*'} \; X^{*} \right)^{-1} \sum_{uu} \right)^{-1} \hat{\beta}_{EIV} \tag{2.12}$$

aufweist.

Als Alternative zum korrigierten Fehler-in-dem-Variablen-Schätzer wäre auch die Bereitstellung von Instrumentvariablen für die anonymisierten Variablen denkbar, in dem die wahren Variablen mit anderen Anonymisierungsstörgrößen stochastisch überlagert werden. Dieser zweite Satz von stochastisch überlagerten Variablen weist alle Eigenschaften von validen Instrumenten eines Instrumentvariablenschätzers auf. Da somit sowohl die Unkorreliertheit zwischen Instrumenten und Fehlerterm als auch die Korrelation zwischen Instrumenten und anonymisierten Regressoren garantiert ist, werden die notwendigen Verteilungsannahmen der IV-Schätzers per Datenkonstruktion erfüllt. Durch Bereitstellung von Instrumentvariablen steigt allerdings das Reidentifikationsrisiko, da anonymisierte Regressoren und Instrumente gemeinsam mehr Information über die wahren Merkmalsausprägungen liefern.

Monte Carlo-Evidenz

Mit Hilfe einer einfachen Monte-Carlo-Studie soll im folgenden die quantitative Auswirkung von Mikroaggregation und stochastischer Überlagerung illustriert werden. Hierzu soll das lineare Modell

$$Y = \beta_0 + \beta_1 \, X_1 + \beta_2 \, X_2 + \varepsilon$$

geschätzt werden. Der Fehlerterm ε wird als unabhängig, identisch $t-$verteilt mit 4 Freiheitsgraden unterstellt, so dass $V\,[\,\varepsilon\,] = 2$. Die beiden erklärenden Variablen werden aus einer bivariaten Normalverteilung der Form

$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & .4 \\ .4 & 1 \end{pmatrix}\right)$$

gezogen. Die Verwendung zweier, relativ stark korrelierter Regressoren ist vor allem im Kontext des in Abschnitt 3 untersuchten semiparametrischen Selektionsschätzers von Interesse, da Multikollinearität zwischen Regressoren und Kontrollfunktion in kleinen Stichproben von Bedeutung sein kann. Für alle Simulationen werden als wahre Parameterwerte für die Regressionskoeffizienten die Werte $\beta_0 = .5,\ \beta_1 = 1,\ \beta_2 = -1$ verwendet.

Basierend auf einem Monte-Carlo Design für stochastische Regressoren werden Schätzungen auf Datensätzen mit Beobachtungsumfang $N = 120$, 1200 und 3600 durchgeführt.[3)] Die Auswertung der Monte-Carlo Schätzungen beruht auf $R = 1000$ Replikationen. Mit dem idealtypischen KQ-Schätzer auf Grundlage der Originaldaten $\hat{\beta}$ wird der KQ-Schätzer $\hat{\beta}_A$ unter Verwendung mikroaggregierter Daten des Aggregationsniveaus $A = 3$, 4 und 5 verglichen. Desweiteren wird der bootstrap-aggregierte Schätzer $\hat{\beta}_B$ untersucht, wobei wir über 3 Beobachtungen $(B = 3)$ aggregieren.

Für die mit stochastischer Überlagerung anonymisierten Daten verwenden wir als Anonymisierungsstörgröße für Y, X_1 und X_2 jeweils unabhängig und identisch normalverteilte Zufallsvariablen mit Erwartungswert 0 und Varianz 0.25. Untersucht werden im Kontext der stochastischen Überlagerung der inkonsistente KQ-Schätzer auf Grundlage der anonymisierten Daten $\hat{\beta}_{EIV}$, der Instrumentvariablenschätzer $\hat{\beta}_{IV}$ mit Instrumenten, die analog zu den anonymisierten Regressoren erzeugt werden, sowie der korrigierte Fehler-in-den-Variablen Schätzer $\hat{\beta}_C$. Zur Berechnung des Korrekturterms verwenden wir die bekannte Varianz-Kovarianz-Matrix der Anonymisierungsstörgrößen Σ_{uu} sowie die empirische Momentenmatrix von X als Schätzung für Q.

Tabelle 1 fasst die Monte-Carlo-Ergebnisse für die Aggregationsschätzer zusammen. Aus Platzgründen werden nur die Resultate für den Koeffizienten vor der X_1 – Variablen wieder gegeben, zumal die Ergebnisse für β_0 und β_2 sich nicht substantiell von den Ergebnissen für β_1 unterscheiden. Neben der mittleren Schätzung und der mittleren Verzerrung berechnen wir die Wurzel des mittleren quadratischen Fehler, RMSE, als Maß für die Schätzunsicherheit in endlichen Stichproben. Der relative Standardfehler, RELSE, ist definiert als das Verhältnis von mittlerem Standardfehler der R Schätzungen zur Standardabweichung der Schätzungen. Da für $R \to \infty$ die Standardabweichungen der Schätzungen gegen den wahren Standardfehler der Schätzung für endliches N konvergieren, geben Abweichungen des relativen Standardfehlers von 1 Auskunft über die Genauigkeit der Schätzung der Standardabweichung des Schätzers aufgrund der asymptotischen Verteilung. Dieses Maß ist vor allem für Schätzer von Interesse, deren Standardfehler nicht für endliche Stichproben berechnet werden können, und

3) Die etwas ungewöhnlichen Werte für den Beobachtungsumfang wurden gewählt, so dass N ein Vielfaches des Aggregationsniveaus A = 3, 4 und 5 ist.

bei denen deshalb asymptotische Approximationen verwendet werden müssen. Der relative Standardfehler wird in zwei Varianten ausgewiesen. Der unkorrigierte relative Standardfehler gibt Auskunft über das Ausmaß der fehlerhaften Inferenz, wenn das Aggregationsniveau bei der Inferenz durch eine entsprechende Korrektur der Freiheitsgrade unberücksichtigt bleibt. Der korrigierte relative Standardfehler verwendet im Zähler die korrekten Standardfehler basierend auf M statt auf N Beobachtungen.

Tabelle 1: Monte-Carlo Ergebnisse: Mikroaggregation im linearen Modell*

$\beta = 1$	Mittelwert	Verzerrung	RMSE	RELSE korrigiert	RELSE unkorrigiert
		$N = 120$			
$\hat{\beta}$	.995	-.005	.145	-	.993
$\hat{\beta}_{A=3}$	.987	-.013	.277	.912	.513
$\hat{\beta}_{A=4}$	.991	-.009	.308	.957	.460
$\hat{\beta}_{A=5}$	.989	-.011	.365	.927	.393
$\hat{\beta}_{B=3}$	.992	-.008	.200	1.277	.718
		$N = 1200$			
$\hat{\beta}$	1.002	.002	.045	-	1.028
$\hat{\beta}_{A=3}$	1.001	.001	.077	1.000	.576
$\hat{\beta}_{A=4}$	.999	-.001	.089	.989	.493
$\hat{\beta}_{A=5}$	1.000	.000	.095	1.040	.463
$\hat{\beta}_{B=3}$	1.001	.001	.063	1.270	.732
		$N = 3600$			
$\hat{\beta}$	1.000	.000	.032	-	1.016
$\hat{\beta}_{A=3}$	1.000	.000	.045	.978	.564
$\hat{\beta}_{A=4}$	1.003	.003	.056	.978	.493
$\hat{\beta}_{A=5}$	1.001	.001	.056	1.021	.456
$\beta_{B=3}$	.999	-.001	.032	1.263	.729

* Schätzung des Koeffizienten vor der ersten erklärenden Variablen, Anzahl der Replikationen = 1000

Wie bereits theoretisch gezeigt, führt die einfache listenweise Mikroaggregation zu keinerlei Verzerrung. Selbst für den kleinen Stichprobenumfang von $N = 120$ liegen die Schätzungen im Mittel für alle betrachteten Aggregationsniveaus recht nahe beim wahren Parameterwert, was offensichtlich eine Folge der gewählten Größe von $V[\varepsilon]$ ist. Allerdings führt die Aggregation zu Effizienzverlusten. Die Streuung der Schätzungen gemessen in termini des mittleren quadratischen Fehlers (RMSE) steigt erheblich mit dem Aggregationsniveau an. Nur für die größte Stichprobe mit 3600 Beobachtungen sind Unterschiede im mittleren quadratischen Fehler nicht mehr auszumachen. Recht erfolgreich schneidet die Boostrap-Aggregation ab. Bei gleichem Aggregationsniveau ist der mittlere quadratische Fehler für $\hat{\beta}_{B=3}$ deutlich geringer als der mittlere quadratische Fehler von $\hat{\beta}_{A=3}$.

Die unkorrigierten relativen Standardfehler der Aggregationsschätzer liegen deutlich unter 1. Wie zu erwarten war, steigt die Verzerrung der Standardfehler mit dem Aggregationsniveau. Die Ignorierung der wahren Freiheitsgrade des Modells führt zu einer fehlerhaften Inferenz, die dem empirischen Wirtschaftsforscher niedrigere $p-$Werte (höhere $t-$Statistiken) vorgaukelt als tatsächlich vorhanden sind. Die Korrektur um die wahre Anzahl von Freiheitsgraden führt dagegen zu Standardabweichungen, die den empirischen Standardabweichungen recht nahe kommen. Eine Ausnahme bildet der Schätzer auf Grundlage der bootstrap-aggregierten Daten. Die unkorrigierten relativen Standardfehler liegen deutlich unter 1, während die korrigierten relativen Standardfehler auf eine Überkorrektur hinweisen. Aufgrund der heteroskedastischen Struktur des bootstrap-aggregierten Modells ist deshalb zu überprüfen, ob eine heteroskedastie-robuster Schätzer der Standardfehler genauere Schätzungen liefert.

Tabelle 2: Monte-Carlo Ergebnis: Stochastische Überlagerung im linearen Modell*

$\beta = 1$	Mittelwert	Verzerrung	RMSE	RELSE
		$N = 120$		
$\hat{\beta}$	.995	-.005	.145	.993
$\hat{\beta}_{EIV}$	.707	-.293	.324	1.018
$\hat{\beta}_{IV}$	1.001	.001	.195	.999
$\hat{\beta}_{CEIV}$	1.008	.008	.184	1.040
		$N = 1200$		
$\hat{\beta}$	1.002	.002	.045	1.028
$\hat{\beta}_{EIV}$	.707	-.293	.297	1.016
$\hat{\beta}_{IV}$	.998	-.002	.063	.985
$\hat{\beta}_{CEIV}$	1.002	-.002	.055	1.028
		$N = 3600$		
$\hat{\beta}$	1.000	.000	.032	1.016
$\hat{\beta}_{EIV}$	.706	-.294	.295	1.020
$\hat{\beta}_{IV}$	.999	-.001	.032	1.018
$\hat{\beta}_{CEIV}$	1.000	.000	.032	1.040

* Schätzung des Koeffizienten vor der ersten erklärenden Variablen, Anzahl der Replikationen = 1000

Tabelle 2 enthält die Ergebnisse der Monte-Carlo-Studie für die Anonymisierung durch stochastische Überlagerung. Die Verzerrung des gewöhnlichen KQ-Schätzers erweist sich als erheblich und wird aufgrund der Inkonsistenz dieses Schätzers auch nicht mit zunehmendem Stichprobenumfang reduziert. Von einer naiven Verwendung von Datensätzen, die durch stochastische Überlagerung anonymisiert werden, ist deshalb abzuraten. Der Instrumentvariablenschätzer erweist sich als recht leistungsstark, obwohl hier sogar auch der Fehlerterm der abhängigen Variablen aufgrund der Anonymisierungsstörgröße von Y eine größere Varianz aufweist als im Originalmodell. Anders als bei der konventionellen IV-Schätzung auf Grundlage nicht experimenteller Instrumente, sind hier die Instrumente per Konstruktion des Anonymisierungsverfahrens stark mit den anonymisierten erklärenden Variablen korreliert. Der mittlere quadratische Fehler, der von der Korrelation zwischen (anonymisierten) Regressoren und den Instrumenten abhängt, ist für die gewählte Parameterkonstellation auch im Vergleich zum KQ-Schätzer auf Grundlage der Originaldaten recht klein. Ähnlich erfolgreich ist der korrigierte Fehler-in-den-Variablen Schätzer. Hier hängt die Präzision des Schätzers von der Schätzgenauigkeit von Q ab. Mit steigendem Beobachtungsumfang konvergiert die Schätzung für Q gegen den wahren Wert. Schon bei einem Stichprobenumfang von $N = 1200$ lassen sich keine wesentlichen Unterschiede zwischen $\hat{\beta}$ und $\hat{\beta}_{CEIV}$ ausmachen. Aus der Sicht des Ökonometrikers, für den bei gegebener faktischer Anonymisierung die Qualität der Schätzung im Vordergrund steht, stellt die stochastische Überlagerung im Kontext des linearen Regressionsmodells eine echte Alternative zur Mikroaggregation dar.

3. Anonymisierung und nichtlineare ökonometrische Modelle

Da sowohl Mikroaggregation als auch die stochastische Überlagerung, wie sie im vorherigen Abschnitt eingeführt wurden, lineare Transformationen der Originaldaten darstellen, sind die Auswirkungen dieser Anonymisierungsmethoden auf Schätzungen linearer Regressionsmodelle sehr viel einfacher zu analysieren als im Falle nichtlinearer Modelle. So führt die stochastische Überlagerung bei nichtlinearen Modellen zu einem nichtlinearen Fehler-in-den-Variablen-Modell. Der Umfang der Literatur zu Messfehlern in nichtlinearen Modellen muss als vergleichsweise gering bezeichnet werden. Spezielle Aspekte werden in den Arbeiten von Amemiya (1985), Hausman, Newey und Powell (1995), Lee und Sepanski (1995) sowie Hong und Tamer (2002) behandelt. Eine Übersicht über neuere Verfahren bietet die Monographie von Carroll, Ruppert und Stefanski (1995). Die Mikroaggegation nichtlinearer Modelle scheint nur unter in Kaufnahme von Approximationsfehlern ein gangbarer Weg zu sein (Lechner u. Pohlmeier, 2003). Als Alternative bieten sich teilweise nichtlineare Modelle für gruppierte Daten an, die jedoch nicht unbedingt Rückschlüsse auf den datengenerierenden Prozess der Mikroebene zulassen.

Im Folgenden schlagen wir deshalb ein alternatives Anonymisierungsverfahren vor, das auch auf den Fall nichtlinearer Regressionsmodelle erweiterbar ist. Die Idee beruht darauf, dass Beobachtungen, die ein hohes Risiko der Reidentifikation aufweisen, zensiert werden, bzw. aus dem Datensatz gelöscht werden (Blanking). Geschätzt werden soll das nichtlineare Regressionsmodell mit additivem Fehlerterm

$$Y_i = f\left(X_i, \beta\right) + \varepsilon_i \tag{3.1}$$

Wir unterstellen, dass das Reidentifikationsrisiko nur bei den Beobachtungseinheiten des Datensatzes groß ist, die extreme Werte für irgendeine der Variablen aufweisen. Es sei W_i der Vektor von insgesamt L Variablen für Beobachtung i. Dieser Vektor enthält die erklärenden Variablen, die zu erklärende Variable Y_i sowie andere sicherheitsrelevante Variablen des Datensatzes, die nicht zwingend Regressoren in (3.1) sein müssen. Eine Beobachtung wird nicht anonymisiert übernommen, wenn alle Variablen von W_i innerhalb der Quantile θ_l und θ_u liegen. Der binäre Indikator für die nichtanonymisierte Übernahme der Variablen von i in den Datensatz ist demnach definiert als

$$S_i = \begin{cases} 1 \text{ wenn} & q_{\theta_l}\left(W_{1j},\ldots,W_{nj}\right) < W_{ij} < q_{\theta_u}\left(W_{1j},\ldots,W_{nj}\right), \quad \forall \; j = 1,\ldots,L \\ 0 \text{ sonst}, & \end{cases} \tag{3.2}$$

wobei $q_\theta(\cdot)$ das $\theta-$Quantil der Variablen W_j bezeichnet mit $\theta_l < \theta_u$. In der Regel sollte das Reidentifikationsrisiko für besonders große Werte von W_{ij} hoch sein, so dass eine Selektion über das untere Quantil zu vernachlässigen ist. Alternativ können auch andere Anonymisierungsregeln unterstellt werden, die beispielsweise von einer hohen Reidentifikationswahrscheinlichkeit aufgrund von Kombinationen der Variablen ausgehen. Es sei nun unterstellt, dass die gewählte Selektionsregel durch eine semiparametrische "Single-Index"-Form approximiert werden kann.

$$S_i = 1\left(\varphi\left(Z_i'\gamma\right) > \tau\right) \tag{3.3}$$

Hierbei bezeichnet $\varphi(\cdot)$ eine zweifach differenzierbare bekannte Funktion bezüglich der Indexfunktion $I_i = Z_i'\gamma$ und τ einen unbekannten zusätzlichen Schwellenparameter. Gleichung (3.3) stellt eine Verallgemeinerung der üblichen linearen Selektionsregel $S_i = 1\left(Z_i'\gamma + u_i > 0\right)$ dar. Kontrollvariablen der Selektionsgleichung (3.3) sind die erklärenden Variablen der Strukturgleichung, da sie via Strukturgleichung die Größe der abhängigen Variablen bestimmen sowie andere Variablen, die aus dem Datensatz für die Beobachtung i zu löschen sind und S_i über Kreuzkorrelationen beeinflussen. Da die erklärenden Variablen in der Selektionsgleichung ebenfalls der Anonymisierung unterliegen können, müssen sie anonymisierter in die Selektionsgleichung eingehen. Für eine zu anonymisierende Variable Z_{ij} der Selektionsgleichung schlagen wir folgende Transformation vor, wenn nur eine Anonymisierung großer Werte erfolgen soll:

$$Z_{ij}^* = 1\left(Z_{ij} < q_{\theta u}\left(Z_j\right)\right) Z_{ij} + \left[1 - 1\left(Z_{ij} < q_{\theta u}\left(Z_j\right)\right)\right] E\left[Z_{ij} \middle| Z_{ij} \geq q_{\theta u}\left(Z_j\right)\right] \tag{3.4}$$

Bei dieser Transformation bleiben die nicht zu anonymisierenden Werte in Originalform erhalten, während die zu anonymisierenden Beobachtungen durch den konditionalen Erwartungswert besetzt werden, der durch das bedingte arithmetische Mittel aus den Originaldaten geschätzt werden kann.

Letztlich kann die Selektionsgleichung auch Variablen enthalten, die über exogene, nicht auf der Quantilsregel (3.2) beruhende Selektionskriterien beruhen. Beispielsweise ist es üblich, Informationen über Branchen, die weniger als eine vorgegebene Anzahl von Unternehmen aufweisen, zu löschen. Die Anzahl der Firmen in einer Branche könnte in diesem Fall ein derartiger Regressor sein.

Im Falle einer linearen Strukturgleichung, $f(X_i, \beta) = X_i'\beta$, können die von einer hohen Reanonymisierungswahrscheinlichkeit betroffenen Beobachtungen zusätzlich auch als Mikroaggregate verwenden werden. Die abhängige Variable $\widetilde{Y}_i$ ist in diesem Fall eine Originalbeobachtung oder ein anonymisierter Wert Y_i^*:

$$\widetilde{Y}_i = S_i \, Y_i + (1 - S_i) \, Y_i^* \tag{3.5}$$

Damit wird der Informationsverlust durch Anonymisierung im Vergleich zur Mikroaggregation über sämtliche Beobachtungen reduziert.

Es sei $n < N$ die Anzahl der Beobachtungen, die nicht von der Anonymisierung (Blanking) betroffen sind. Für diese Bebachtungen gilt die konditionale Populationsregressionsfunktion:

$$\begin{aligned} E\left[Y_i | \varphi(Z_i'\gamma), S_i = 1\right] &= f(X_i, \beta) + E\left[\varepsilon_i | \varphi(Z_i', \gamma), S_i = 1\right] \\ &= f(X_i, \beta) + \lambda(Z_i'\gamma) + \zeta_i, \end{aligned} \tag{3.6}$$

wobei $\lambda(\cdot)$ eine allgemeine Selektionskontrollfunktion bezeichnet und $\zeta_i = \varepsilon_i - \lambda_i$ ein heteroskedastischer Fehlerterm mit $E\left[\zeta_i | \varphi(Z_i, \gamma), S_i = 1\right] = 0$ ist. Die Identifikationsbedingungen für dieses semiparametrische Modell mit linearer Regressionsfunktion werden ausführlich in Newey (1999) diskutiert. Das Problem adäquater Ausschlussrestriktionen im Fall der Anonymisierung ist deutlich geringer als bei typischen Anwendungen von Selektionskorrekturverfahren, die auf dem Prinzip der Selektion über unbeobachtbare Faktoren (selection on unobservables) beruhen. In unserem Fall beruht die Selektion in aller Regel auch auf Variablen, die nicht in der Strukturgleichung als erklärende Variablen enthalten sind. Diese Variablen liefern die notwendigen überidentifizierenden Restriktionen. Es ist wichtig darauf hinzuweisen, dass eine eventuelle Konstante in diesem Modell über den Korrekturterm λ_i aufgefangen wird und nicht ohne weitere Annahmen (vgl. Andrews u. Schafgans, 1998) identifizierbar ist.

Das nichtlineare Modell mit semiparametrischer Selektionskontrollfunktion wird im Folgenden über ein zweistufiges Verfahren ähnlich dem Zwei-Stufen-Schätzer von Heckman geschätzt. In der ersten Stufe werden die Parameter der Selektionsgleichung mit Hilfe eines semiparametrischen Schätzers für binäre Auswahlmodelle geschätzt. Hierfür verwenden wir den von Klein und Spady (1993) vorgeschlagenen semiparametrisch effizienten Schätzer. Als Alternative sind andere semiparametrische $\sqrt{N-}$ konsistente Schätzer denkbar, wie z.B. der semi-nichtparametrische Likelihood-Ansatz für binäre Auswahlmodelle von Gabler, Laisney, Lechner (1993) oder der semiparametrische Momentenschätzer von Ichimura (1993). Für die zweite Schätzstufe verwenden wir Neweys (1999) semiparametrischen Schätzer, bei dem die Selektionskontrollfunktion durch eine allgemeine Reihenapproximation ersetzt wird.[4)]

Das Klein-Spady Verfahren beruht auf einem parametrischen Likelihood-Ansatz, bei dem die binäre Auswahlwahrscheinlichkeit $P\left(S_i = 1 \middle| Z_i'\gamma\right)$ unspezifiziert bleibt:

$$ln\ L(\gamma) = \sum_{i=1}^{n} S_i\ ln\ P\left[S_i = 1 \middle| Z_i'\gamma\right] + \left(1 - S_i\right) ln\left[1 - P\left[S_i = 1 \middle| Z_i'\gamma\right]\right] \qquad (3.7)$$

Klein und Spady formulieren diese Wahrscheinlichkeit mittels des Bayes Theorems um als

$$P\left(S_i = 1 \middle| Z_i'\gamma\right) = \frac{P\left(S_i = 1\right) g_{I|S=1}\left(Z_i'\gamma \middle| S_i = 1\right)}{g_I\left(Z_i'\gamma\right)}, \qquad (3.8)$$

wobei g_I die Dichte der Indexfunktion $I_i = Z_i'\gamma$ ist und $g_{I|S=1}$ die konditionale Dichte, gegeben $S_i = 1$. Die Auswahlwahrscheinlichkeit (3.8) wird geschätzt, indem sämtliche Terme dieser Wahrscheinlichkeit unabhängig voneinander nichtparametrisch geschätzt werden.[5)] Durch Ersetzen der Auswahlwahrscheinlichkeit durch den Schätzer ergibt sich die Quasi-Likelihood-Funktion:[6)]

$$\max_{\gamma} \ln Q\left(\gamma\right) = \sum_{i=1}^{n} S_i \ln\left(\left[\hat{P}\left[S_i = 1 \middle| Z_i'\gamma\right]\right]^2\right) + \left(1 - S_i\right) \ln\left(\left(1 - \hat{P}\left[S_i = 1 \middle| Z_i'\gamma\right]\right)^2\right) \qquad (3.9)$$

Für die zweite Schätzstufe schlägt Newey vor, die unbekannte Kontrollfunktion $\lambda(\cdot)$ mit einer linearen Kombination von J Grundfunktionen ρ_j zu approximieren:

---

4) Für den linearen Fall bietet sich auch an, das Verfahren von Powell (1987) zu verwenden, das auf einer Kernschätzung der Kontrollfunktion beruht. Siehe Newey, Powell und Walker (1990) für eine vergleichende Studie.

5) Die beiden Dichten lassen sich mit univariatem Kernschätzer schätzen. $P\left[S_i = 1\right]$ kann durch das arithmetische Mittel geschätzt werden.

6) Die geschätzte Wahrscheinlichkeit wird quadriert, weil deren Schätzung u. U. auch negativ sein kann.

$$\lambda(\cdot) \approx \sum_{j=1}^{J} \eta_j \cdot \rho_j \,, \tag{3.10}$$

wobei für $J \to \infty$ der Approximationsfehler verschwindet und η_j ein unbekannter zu schätzender Koeffizient ist. Diese Grundfunktionen hängen nur von der Indexfunktion ab. Ersetzen wir λ durch die Approximation (3.10) erhalten wir:

$$Y_i = f(X_i, \beta) + \sum_{j=1}^{J} \eta_j \rho_j(\tau - Z_i' \hat{\gamma}) + \hat{\xi}_i \qquad \hat{\xi}_i = \sum \eta_j \, (\rho_j - \hat{\rho}_j) + \xi_i \tag{3.11}$$

Die Koeffizienten β und η_j können nun mit der nichtlinearen KQ-Methode geschätzt werden. Die optimale Ordnung von J wird durch ein Optimierungsverfahren bestimmt (siehe Appendix A II). Newey schlägt vor, die folgende polynomiale Approximation zu verwenden:

$$\rho_j(\tau - Z_i' \gamma) = [\Psi \ (\tau - Z_i' \gamma)]^j \,,$$

wobei Ψ eine monotone auf das Intervall $[-1;1]$ beschränkte Funktion ist. Weitere Details zum Newey-Verfahren findet der interessierte Leser in Appendix A II.

Monte Carlo Evidenz

Mit Hilfe einer einfachen Monte-Carlo Studie soll im folgenden überprüft werden, ob die wahren Modellparameter möglichst akkurat mit Hilfe des vorgestellten zweistufigen semiparametrischen Selektionsverfahrens geschätzt werden können, wenn die Anonymisierung durch Blanking gemäß der Quantilsregel (3.2) erfolgt. Die zu anonymisierenden Regressoren der Selektionsgleichung werden gemäß (3.4) transformiert. Das gewählte Design der Simulationen ist im Wesentlichen das gleiche wie im vorherigen Abschnitt. Geschätzt werden soll wiederum ein lineares Modell mit den selben wahren Parameterwerten und einem $t-$ verteilten Fehlertermprozess.

Die erklärenden Variablen und die weiteren Instrumente der Selektionsgleichung werden als multivariate normalverteilte Zufallsvariablen der Form

$$\begin{pmatrix} Z_1 \\ Z_2 \\ Z_3 \\ Z_4 \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \begin{bmatrix} 1 & .4 & .2 & .1 \\ .4 & 1 & .3 & .2 \\ .2 & .3 & 1.2 & .1 \\ .1 & .2 & .1 & .8 \end{bmatrix} \right)$$

gezogen, wobei $X_1 = Z_1$ und $X_2 = Z_2$ die erklärenden Variablen des Regressionsmodells bilden. Eine Beobachtung i wird aus dem Datensatz gelöscht $(S_i = 0)$, wenn irgendeine der Variablen von $(W_i = Y_{1i}, Z_{1i}, \ldots, Z_{4i})$ größer ist als das 90-Prozent Quantil dieser Variablen.

$$S_i = \begin{cases} 1 \text{ wenn } 1(Y_i < q_{.90}(Y)) \cdot \prod_{j=1}^{4} 1(Z_{ij} < q_{.90}(Z_j)) = 1 \\ 0 \text{ sonst} \end{cases}$$

Tabelle 3 gibt die Ergebnisse der Monte-Carlo Simulationen für die Stichprobenumfänge $N =$ 120, 1200 und 3600 wieder. Durch das Löschen von Beobachtungen beruhen jedoch die Regressionen der zweiten Stufe auf geringeren Stichprobenumfängen. Für die Stichprobenverzerrung relevant ist ausschließlich die Unterdrückung von Beobachtungen mit Ausprägung der abhängigen Variablen oberhalb des 90%-Quantils, während die Unterdrückung von Beobachtungen aufgrund extremer Werte anderer Variablen zu einem Verlust an Effizienz führt. Der Effizienzverlust durch die hier vorgegebene Form des Data Blanking speist sich aus zwei Quellen. Zum einen werden in der zweiten Stufe weniger Beobachtungen verwendet, zum anderen wird die Stichprobenvariation der erklärenden Variablen reduziert. In Tabelle 3 bezeichnet $\bar{n}$ die durchschnittliche Anzahl von Beobachtungen, die aufgrund der Anonymisierung in der zweiten Stufe verwendet wurde, während $\bar{n}_Z$ den durchschnittlichen Beobachtungsumfang bezeichnet, wenn die Selektion ausschließlich über die Z-Variablen erfolgt. Für die drei Experimente reduziert sich der Stichprobenumfang in der zweiten Stufe um 37-40%.

Bei einem kleinen Stichprobenumfang von $N = 120$ (bzw. $\bar{n} = 72.81$) weist der zweistufige semiparametrische Schätzer eine mittlere Verzerrung auf, die über der Verzerrung der Aggregationsschätzer liegt. Allerdings reduziert sich diese Verzerrung deutlich mit steigendem Stichprobenumfang. Selbst bei kleinen Stichprobenumfängen ist die Schätzunsicherheit in termini des RMSE auf einem vergleichbaren Niveau wie die der IV-Schätzer und liegt deutlich niedriger als bei den Aggregationsschätzern.

Für kleine Stichprobenumfänge wird der Standardfehler des Selektionsschätzers deutlich zu hoch ausgewiesen. Aber schon bei einer Stichprobengröße von $N = 1200$ $(\bar{n} = 754.09)$ scheint die asymptotische Approximation zu greifen, so dass sich Standardabweichung der Schätzungen und Mittelwert der geschätzten Standardabweichungen angleichen.

Tabelle 3: Monte-Carlo Ergebnis: Semiparametrisches Selektionskorrektur-Modell*

$\beta = 1$	Mittelwert	Verzerrung	RMSE	RELSE
	$N = 120\ (\overline{n} = 72.81, \overline{n}_Z = 80.51)$			
$\hat{\beta}$	1.006	.006	.133	1.016
$\hat{\beta}_{NS}$	1.019	.019	.194	1.646
	$N = 1200\ (\overline{n} = 754.09, \overline{n}_Z = 827.22)$			
$\hat{\beta}$	1.001	.001	.042	1.003
$\hat{\beta}_{NS}$	1.004	-.004	.054	1.048
	$N = 3600\ (\overline{n} = 2269.54, \overline{n}_Z = 2487.20)$			
$\hat{\beta}$	1.000	.000	.024	.995
$\hat{\beta}_{NS}$	1.007	-.007	.034	.989

* Schätzung des Koeffizienten vor der ersten erklärenden Variablen, Anzahl der Replikationen = 1000

4. Ein illustratives Beispiel

Da die praktische Relevanz von Monte-Carlo Ergebnissen von den Annahmen über den zugrunde gelegten stochastischen Prozess bzw. der Realitätsnähe dieser Annahmen abhängen, sollen anhand einer empirischen Anwendung die Auswirkungen von Aggregationsmethoden untersucht werden. Hierfür verwenden wir einen Querschnitt von 3600 Firmen des verarbeitenden Gewerbes der Kostenstrukturerhebung (KSE) des Jahres 1999. Erklärt werden soll der Anteil der gesetzlichen Sozialkosten einer Firma in Abhängigkeit von der Anzahl der vollzeitbeschäftigten Arbeitnehmer und der Anzahl der teilzeitbeschäftigten Arbeitnehmer. Das gewählte Anwendungsbeispiel soll eine mögliche, wenn auch stark vereinfachte Anwendung für anonymisierte Daten der KSE sein. In diesem Beispiel geben die Regressionskoeffizienten einen Hinweis darauf, inwieweit die gesetzliche Sozialkostenbelastung auf Unternehmensebene von der Beschäftigungsstruktur abhängt. Nicht uninteressant ist die Fragestellung, ob die Beschäftigung von Teilzeitbeschäftigten im Vergleich zu Vollzeitbeschäftigten kostenneutral erfolgt. Die beiden erklärenden Variablen werden in standardisierter Form als Regressoren verwendet. Eine Standardisierung ist sinnvoll, um Regressoren von unterschiedlicher Dimension oder unterschiedlicher Skalierung mit Störgrößen mit gleicher Varianz zu überlagern. Wie in der Monte-Carlo-Studie zuvor, wählen wir normal verteilte Überlagerungsfehler mit einer Varianz von .25.

Tabelle 4 gibt die Schätzergebnisse für die gewöhnliche KQ-Schätzung auf der Grundlage der Originaldaten sowie die Ergebnisse für anonymisierten Datensätze wieder. Deutlicher als in den beiden Monte-Carlo Experimenten zuvor zeigen sich erhebliche Unterschiede zwischen der "Originalschätzung" und den Schätzungen, die auf den weniger informativen anonymisierten Datensätzen beruhen.

Unsere Schätzergebnisse verdeutlichen recht anschaulich, dass die Wahl der Anonymisierungsmethode sowie die Wahl der entsprechenden Anonymisierungsparameter (z. B. Höhe des Aggregationsniveaus, Größenordnung der Überlagerung) die Schätzergebnisse substantiell beeinflussen. Die auf Grundlage der Originaldaten geschätzten Koeffizienten sind statistisch auf dem 1% Signifikanzniveau abgesichert. Die Aggregationsschätzer und der Bootstrap-Aggregationsschätzer liefern ähnliche Parameterschätzungen. Allerdings ist der Koeffizient vor der Variablen Teilzeitbeschäftigte für den Bootstrap-Schätzer und den Aggregationsschätzer mit A=5 nicht mehr statistisch abgesichert. Die Ergebnisse sind aber möglicherweise für die einfachen Aggregationsschätzer beschönigend, da durch die spezielle Sortierung des Originaldatensatzes homogene Firmen der gleichen Bereiche aggregiert wurden. Der Bootstrap-Aggregationsschätzer beruht auf einer einzigen Bootstrap-Aggregation für $B = 3$. Eine Schätzung auf einer anderen zufälligen Aggregation, die hier nicht wiedergegeben wird, führt zu einem positiven Koeffizienten vor der Teilzeitbeschäftigungsvariablen. Der Instrumentvariablen-Schätzer und der korrigierte Fehler-in-den-Variablen-Schätzer liefern ähnliche Ergebnisse wie der OLS-Schätzer, jedoch ist auch hier der letzte Regressionskoeffizient statistisch nicht abgesichert.

Tabelle 4: Auswirkungen der Anonymisierung: Ein Anwendungsbeispiel*

	Konstante	Vollzeitbeschäftigte	Teilzeitbeschäftigte
OLS	.120	.598	-.165
	(5.39)	(9.98)	(-2.747)
$B = 3$	.120	.485	-.029
	(9.15)	(2.14)	(-.16)
$A = 3$	.120	.627	-.212
	(4.98)	(6.44)	(-2.00)
$A = 4$	.120	.755	-.404
	(4.83)	(6.48)	(-3.12)
$A = 5$	.120	.608	-.157
	(4.71)	(4.73)	(-1.09)
EIV	.130	.786	.083
	(5.41)	(8.96)	(2.59)
IV	.131	.412	.003
	(5.45)	(3.33)	(.02)
CEIV	.134	.660	.0243
	(5.58)	(4.95)	(-1.82)

* Abhängige Variable: log Gesetzliche Sozialkosten, t-Werte in Klammern.

5. Schlussfolgerung

In dieser Arbeit werden verschiedene Anonymisierungsmethoden hinsichtlich ihrer Auswirkung auf die Qualität von ökonometrischen Schätzungen untersucht. Es wird gezeigt, dass standardmäßige Anonymisierungsverfahren wie Mikroaggregation und stochastische Überlagerung, sofern ihre Auswirkungen auf den generierenden Prozess für den Anwender bekannt sind, nicht unbedingt zu einer gravierenden Reduktion der Qualität der Schätzungen führen müssen. Hierzu muss jedoch die Struktur des Anonymisierungsverfahrens (z.B. Verlässlichkeitsquoten im Falle der stochastischen Überlagerung) dem Empiriker bekannt sein. Bei kleinen Stichproben kann Mikroaggregation zu einer deutlichen Reduktion der Schätzgenauigkeit führen. Wir zeigen, dass die stochastische Überlagerung als Anonymisierungsverfahren eine attraktive Alternative zur Mikroaggregation darstellt, sofern die datenerhebende Institution Informationen über die Kovarianzstruktur der Überlagerung dem Empiriker zu Händen gibt.

Die schöne heile Welt der Anonymisierung kann aber nur für einfache Anonymisierungsverfahren und Anwendungen des linearen Regressionsmodells aufrecht erhalten werden. Sobald die Aggregation gewichtet erfolgt und die Gewichtung auf einer potentiellen endogenen Variablen beruht, haben wir es mit komplexen Selektionsmechanismen zu tun, die sich nur schwerlich modellieren lassen.

Die Analyse von Mikrodaten erfordert fast zwangsläufig die Verwendung von nichtlinearen Regressionsmodellen (qualitative Auswahlmodelle, Regressionsmodelle für begrenzt abhängige Variablen, Zähldatenmodelle etc.). Stochastische Überlagerung führt in diesem Fall zu komplexen nichtlinearen Fehler-in-den-Variablen-Modellen. Diese Modelle für eine allgemeine Struktur der Überlagerungsfehler (Zählvariablen-Fehler, Fehler für nominal skalierte Variablen, Fehler für stetige intervallskalierte Variablen etc.) und eine allgemeine nichtlineare Form zu schätzen, ist nicht unbedingt als trivial zu bezeichnen. In dieser Arbeit zeigen wir, wie ein allgemeines, möglicherweise nichtlineares Modell über einen semiparametrischen, zweistufigen Selektionskontrollschätzer geschätzt werden kann. Der Schätzer unterscheidet sich von Heckmans Zwei-Stufen-Schätzer für Selektionsmodelle dadurch, dass keine Verteilungsannahmen bezüglich der Fehlerterme der Selektionsgleichung und der Strukturgleichung getroffen werden und die Selektionswahrscheinlichkeit nur auf der Single-Index-Struktur beruht. Anhand von Monte-Carlo-Simulationen und eines empirischen Beispiels zeigen wir, dass dieser Ansatz zumindest bei größeren Stichproben ein gangbarer Weg ist, eine Selektionskorrektur infolge von “Data Blanking“ in nichtlinearen Modellen durchzuführen. Obwohl der hier verwendete Blanking-Mechanismus nicht die Form eines schwellenüberschreitenden binären Auswahlmodells aufweist, scheint die semiparametrische Single-Index-Struktur durchaus geeignet zu sein, den Selektionsmechanismus abzubilden.

Die zukünftige Forschung sollte sich weiter darauf konzentrieren, adäquate nichtlineare Schätzer für anonymisierte Mikrodaten zu entwickeln, da anderenfalls der Wert wissenschaftlich ergiebiger, aber anonymisierter Individualdaten erheblich eingeschränkt wird. Mehrere Wege bieten sich für die zukünftige Forschung an. Im Kontext der Selektionsmodelle scheint der Versuch sinnvoll zu sein, die Anonymisierungswahrscheinlichkeit genauer abzubilden, um in der zweiten Stufe eine präziser geschätzte Kontrollfunktion zu erhalten. Für lineare Strukturgleichungen sollten andere Verfahren (z. B. der Schätzer von Powell (1987)) mit den hier verwendeten Schätzern verglichen werden.

Das “Blanking“ von Daten ist nur ein grobes Anonymisierungsverfahren. Selektionsmodelle könnten analog zu Lanot und Walker (1998) um eine weitere Gleichung für anonymisierte Beobachtungen erweitert werden, um sämtliche Beobachtungen des Originaldatensatzes für die Regressionsanalyse zu verwenden und somit den Informationsverlust zu reduzieren.

Literaturhinweise

Amemiya, T. (1985): Instrumental Variable Estimator for the Non-linear Errors in Variable Model, in: Journal of Econometrics, 28, S. 273-289.

Andrews, D. and M. Schafgans (1998): Semiparametric Estimation of the Intercept of a Sample Selection Model, in: Review of Economic Studies, 65, S. 497-517.

Brand, R. (2000): Anonymität von Betriebsdaten, Beiträge zur Arbeitsmarkt- und Berufsforschung, BeitrAB 237, IAB, Nürnberg.

Carroll, R., D. Ruppert and L.F. Stefanski (1995): Measurement Error in Nonlinear Models, Chapman and Hall.

Fuller, W.A. (1987): Measurement Error Models, Wiley.

Gabler, S., F. Laisney und M. Lechner (1993): Semiparametric Estimation of Binary Choice Models with an Application to Labor Force Participation, in: Journal of Business and Economic Statistics, 11, S. 61-8.

Gottschalk, S. (2002): Anonymisierung von Unternehmensdaten: Ein Überblick und beispielhafte Darstellung anhand des Mannheimer Innovationspanels, Discussion Paper No. 02-23, Mannheim, ZEW.

Hausman, J., W. Newey and J. Powell (1995): Nonlinear Errors in Variables Models, in: Journal of Econometrics, 41, S. 159-185.

Hong, H. and E. Tamer (2002): A Simple Estimator for Nonlinear Error in Variable Models, Princeton University, unpublished.

Ichimura, H. (1993): Semiparametric Least Squares (SLS) and weighted SLS Estimation of Single-Index Models, in: Journal of Econometrics, 58, S. 71-12.

Klein, R.W. und R.S. Spady (1993): An Efficient Semiparametric Estimator of the Binary Response Model, in: Econometrica, 61, S. 387-421.

Lanot, G. and I. Walker (1998): The Union/Non Union Wage Differential: An Application of Semi-Parametric Methods, in: Journal of Econometrics, 84, S. 327-349.

Lee, L.F. und J.H. Sepanski (1995): Estimation of Linear and Nonlinear Error in Variables Models Using Validation Data, in: Journal of the American Statistical Association, 90, S.130-14.

Lechner, S. and W. Pohlmeier (2003): Microaggregation in Nonlinear Models: A Note, Center of Finance and Econometrics, University of Konstanz, unpublished working paper.

Newey, W.K., Powell, J.L. und J.R. Walker (1990): Semiparametric Estimation of Selection Models: Some Empirical Results, in: American Economic Review, Paper and Proceedings, 80, S. 324-328.

Newey, W.K. (1999): Two step Series Estimation of Sample selection Models, Department of Economics, Working Papers No-99-04, Massachusetts, Institute of Technology.

Paaß, G., und U. Wauschkuhn (1984): Datenzugang, Datenschutz, und Anonymisierung, Analysepotential und Identifizierbarkeit von anonymisierten Individualdaten, in: Berichte der Gesellschaft für Mathematik und Datenverarbeitung, Bericht 148, Oldenbourg Verlag.

Powell, J.L. (1987): Semiparametric Estimation of Bivariate Latent Variable Models, Working Paper No. 8704, SSRI, University of Wisconsin.

Appendix A I

Proposition 1:

$V\left[\hat{\beta}_A\right] - V\left[\hat{\beta}\right]$ ist positiv definit.

Beweis:

$V\left[\hat{\beta}_A\right] - V\left[\hat{\beta}\right]$ ist nur positiv definit, wenn und nur wenn die Differenz der Inversen der Varianz-Kovarianzmatrizen $V\left[\hat{\beta}\right]^{-1} - V\left[\hat{\beta}_A\right]^{-1}$, positiv definit ist.

Unter Vernachlässigung von σ^2 gilt hierfür

$$\begin{aligned} X'X' - X^{*'}X^* &= X'\left[I - D'D\right]X \\ &= X'\left[I - D\right]X \\ &= X'W\,X\,, \end{aligned}$$

wobei D und $W = I - D$ symmetrische, idempotente Formen sind. Da X und D vollen Spaltenrang besitzen, gilt für jeden Vektor $q \neq 0$:

$q'\,X'W\,X\,q = q'\,\widetilde{X}'\,\widetilde{X}\,q = v'v > 0$,

wobei $\widetilde{X} = W\,X$ und $v = \widetilde{X}\,q$.

Proposition 2:

$$E\left[\frac{e^{*'}\,e^*}{M-K}\right] = \sigma^2$$

Beweis:

$$\begin{aligned} E\left[e^{*'}\,e^*\right] &= E\left[\varepsilon^{*'}\,M^*\varepsilon^*\right] \\ &= E\left[tr\varepsilon^{*}{}'M^*\varepsilon^*\right] \\ &= trM^*E\left[\varepsilon^*\varepsilon^{*}{}'\right] \\ &= \sigma^2 trM^*D \\ &= \sigma^2\left(trD - trX^{*'}\left(X^{*'}X^*\right)^{-1}X^{*'}\right) \\ &= \sigma^2\left(M-K\right), \end{aligned}$$

so dass für die um die Freiheitsgrade $M-K$ korrigierte Fehlerquadratsumme die Proposition 2 hält.

Appendix A II

Varianz Matrix des Newey-Series Schätzer

Es sei $\hat{W}_i = \left(X_i', \hat{p}_{i1}, \hat{p}_{i2}, \ldots, \hat{p}_{Ji}\right)'$,

$\hat{W} = \left(\hat{W}_1, \hat{W}_2 \ldots, \hat{W}_n\right)'$,

$\hat{\theta} = \left(\hat{\beta}_{NS}', \hat{\eta}_1, \hat{\eta}_2, \ldots, \hat{\eta}_J\right)'$,

Die optimale Anzahl der Grundfunktionen minimiert die folgende Funktion

$$J_{OPT} = arg\,min\; CV\left(J\right) = \sum_{i=1}^{n}\left[\frac{\left(1-2\,\hat{\delta}_i\right)\hat{e}_i}{1-\hat{\delta}_i}\right]^2,$$

wobei $\hat{\delta}_i = \hat{W}_i'\left(\hat{W}'\hat{W}\right)^{-1}\hat{W}_i$ und $\hat{e}_i = Y_i - \hat{W}_i'\hat{\theta}$.

$$\hat{V}_{NS} = \left[I_k, 0\right]\left(\frac{\hat{W}'\hat{W}}{n}\right)^{-1} \quad A\left(\frac{\hat{W}'W}{n}\right)^{-1}\left[I_k\;0\right]'$$

$$A = \frac{1}{n}\sum_{i=1}^{n}\hat{W}_i\hat{W}_i' \quad \left(Y_i - \hat{W}_i\,\hat{\theta}\right)^2 + \hat{H}\,\hat{V}\left(\hat{\gamma}\right)\hat{H}'$$

$$\hat{H} = \frac{1}{n}\sum_{i=1}^{n}\hat{W}_i\left[\frac{\partial\left(\sum_{j=1}^{J}\hat{\eta}_j\cdot\rho_j\left(Z_i'\,\hat{\gamma}\right)\right)}{\partial\left(Z_i'\,\hat{\gamma}\right)}\right]\cdot\left[\frac{\partial\left(Z_i'\,\hat{\gamma}\right)}{\partial\,\gamma'}\right],$$

wobei I_k die Einheitsmatrix, deren Dimension gleich der Anzahl der erklärenden Variablen in der Strukturellgleichung ist. $\hat{V}\left(\hat{\gamma}\right)$ ist eine konsistente Schätzung der Varianz des Schätzers der ersten Stufe.