Krause, Joscha

Article

Robuste Schätzung regionaler Indikatoren auf Basis unsicherer Daten durch regularisierte Regression

WISTA - Wirtschaft und Statistik

Provided in Cooperation with:
Statistisches Bundesamt (Destatis), Wiesbaden

Suggested Citation: Krause, Joscha (2022) : Robuste Schätzung regionaler Indikatoren auf Basis unsicherer Daten durch regularisierte Regression, WISTA - Wirtschaft und Statistik, ISSN 1619-2907, Statistisches Bundesamt (Destatis), Wiesbaden, Vol. 74, Iss. 1, pp. 25-33

This Version is available at:
https://hdl.handle.net/10419/250083

ROBUSTE SCHÄTZUNG REGIONALER INDIKATOREN AUF BASIS UNSICHERER DATEN DURCH REGULARISIERTE REGRESSION

Joscha Krause

↘ **Schlüsselwörter:** Internetdaten – Messfehler – regularisierte Regression – robuste Statistik – Small Area Estimation

ZUSAMMENFASSUNG

Small Area Estimation erlaubt die Schätzung regionaler Indikatoren anhand kleiner Stichproben. Hierfür werden Survey-Daten mehrerer Regionen in statistischen Modellen kombiniert. Aufgrund der Digitalisierung der Gesellschaft entstehen immer mehr zusätzliche Datenquellen, wie etwa Website- und Social-Media-Daten, welche die Modelle weiter verbessern könnten. Bei der Verwendung dieser Daten ist jedoch zu beachten, dass ihre Eigenschaften sich deutlich von Stichprobendaten unterscheiden. Sie sind unter anderem mit einer generellen Unsicherheit assoziiert, welche sich nicht quantifizieren lässt. Klassische Schätzverfahren können dies nicht antizipieren, was zu ungenauen Ergebnissen führt. Dieser Artikel zeigt, dass die Schätzung regionaler Indikatoren gegen Unsicherheit mithilfe von Regularisierung robustifiziert werden kann.

↘ ***Keywords:*** *internet data – measurement errors – regularised regression – robust statistics – small area estimation*

ABSTRACT

Small area estimation facilitates the estimation of regional indicators from small samples. This requires the combination of survey data from multiple regions in statistical models. Due to the rapid digitalisation of public life, more and more additional data sources arise, such as website and social media data. These data sources can further improve the models. However, researchers must be aware that these data have fundamentally different properties than sample data. They are often associated with a general uncertainty that cannot be quantified. Classical estimators cannot account for this, which leads to inaccurate estimates. This paper demonstrates that regularisation is a useful tool to robustify small area estimates against uncertainty.

Dr. Joscha Krause

war wissenschaftlicher Mitarbeiter am Lehrstuhl für Wirtschafts- und Sozialstatistik an der Universität Trier. In dieser Zeit veröffentlichte er zahlreiche Artikel und war in mehreren Forschungsprojekten tätig, unter anderem für das Statistische Bundesamt sowie für das Wissenschaftliche Institut der AOK. Aktuell arbeitet er als Senior Data Scientist bei CURE Intelligence sowie als externer Lehrbeauftragter für Statistik und Data Science an der Universität Trier. Seine Doktorarbeit mit dem Titel „Regularization methods for statistical modelling in small area estimation“ wurde von Prof. Dr. Ralf Münnich (Universität Trier) und Prof. Dr. Domingo Morales (Universität Miguel Hernández Elche) betreut. Für seine Dissertation, aus der er Auszüge im vorliegenden Artikel vorstellt, wurde er 2021 mit dem Gerhard-Fürst-Preis des Statistischen Bundesamtes ausgezeichnet.

1

Einleitung

Forschung und Politik benötigen belastbare Zahlen zu Bevölkerungsindikatoren auf regionaler Ebene. Solche Informationen liegen häufig nicht vollerhoben in amtlichen Registern vor, sie müssen stattdessen von Forschenden statistisch geschätzt werden. Hierfür werden in der Regel Stichprobendaten verwendet, welche aus Surveys wie dem Mikrozensus stammen. Aufgrund hoher Erhebungskosten und des großen personellen Aufwands enthalten Stichproben häufig nur wenige Beobachtungen in Relation zu Vollerhebungen. Werden auf dieser Basis nun regionale Indikatoren anhand direkter Schätzer ermittelt, so sind die Schätzwerte mit sehr großen Varianzen assoziiert. Die abgeleiteten Ergebnisse sind in der Regel nicht verlässlich.

Small Area Estimation ist ein Sammelbegriff für statistische Methoden, welche diese Problematik adressieren. Sie kombinieren die Stichprobendaten mehrerer Regionen (Areas) mit geeigneten Hilfsinformationen in statistischen Modellen. Area-Indikatoren werden anschließend anhand von Prädiktionen auf Basis der Modelle geschätzt. Dies erlaubt, den effektiven Stichprobenumfang für die Schätzung eines regionalen Indikators zu vergrößern, und reduziert somit die Varianz relativ zu klassischen direkten Schätzern. Die Ergebnisse sind folglich genauer und verlässlicher. Eine umfassende Einführung in Small Area Estimation bieten Rao und Molina (2015) sowie Morales und andere (2021a).

Eine wesentliche Bedingung für den Effizienzgewinn gegenüber direkten Schätzern ist die Verfügbarkeit von geeigneten Hilfsvariablen. Diese müssen einen möglichst starken statistischen Zusammenhang zu jener Zielvariablen aufweisen, anhand dessen der Area-Indikator berechnet wird (Molina und andere, 2015). Je stärker der Zusammenhang ist, desto genauer können die Indikatoren mit Small Area Estimation geschätzt werden. Hier bietet die fortschreitende Digitalisierung des öffentlichen Lebens neue Möglichkeiten für die sozioökonomische Forschung. Es entstehen immer mehr alternative Datenquellen, etwa Website- und Social-Media-Daten, mit neuen Hilfsvariablen. Diese Daten sind in großer Zahl verfügbar und gewähren soziale Einblicke, die sich mit klassischen Stichprobendaten nur schwer erfassen lassen. Sie können statistische Modelle entscheidend bereichern und die Schätzung von Area-Indikatoren deutlich verbessern.

Doch bei der Nutzung solcher modernen Datenquellen für Small Area Estimation ist zu beachten, dass ihre Eigenschaften sich sehr deutlich von Stichprobendaten unterscheiden. Internetdaten weisen häufig eine generelle Unsicherheit auf. Sie unterliegen Selektionsverzerrungen und sind durch fehlerhafte Angaben kontaminiert. Diese Probleme werden nachfolgend unter dem Begriff Messfehler zusammengefasst.

Messfehler sind keine Eigenheit von Internetdaten, sondern treten auch bei Stichprobendaten auf. Deshalb ist die Schätzung von Area-Indikatoren mittels fehlerhafter Daten eine bereits bekannte methodische Herausforderung. Sie wurde in der Vergangenheit auf Basis von Messfehlermodellen gelöst. Das Grundprinzip dieser Verfahren ist es, Verteilungsannahmen bezüglich der Messfehler einzuführen und anschließend robuste Schätzwerte unter dieser Prämisse herzuleiten. Methodische Beiträge hierzu wurden von unter anderem von Ybarra/Lohr (2008), Torabi und anderen (2009) sowie Burgard und anderen (2020a, 2020b, 2021b) und Morales und anderen (2021b) veröffentlicht.

Internetdaten unterscheiden sich von Stichprobendaten jedoch darin, dass bei Surveys der Erhebungsprozess kontrollierbar ist. Dies erlaubt eine approximative Quantifizierung der Messfehler, oder zumindest das Ableiten von Verteilungsannahmen bezüglich der Fehler. Bei Internetdaten ist dies nicht möglich. Sollen solche Daten also für Small Area Estimation (SAE) verwendet werden, braucht es Schätzverfahren, welche robuste Ergebnisse in der Gegenwart nicht quantifizierbarer Messfehler ohne Verteilungsannahmen liefern. Hierfür bietet die statistische Literatur derzeit kaum verwendbare Ansätze.

Dieser Artikel präsentiert einen mathematischen Rahmen, der es erlaubt, bekannte SAE-Modelle weiterzuentwickeln, um eine solche Robustheit zu erreichen. Hierzu wird eine Verbindung von robuster Optimierung (El Ghaoui/Lebret, 1997; Ben-Tal und andere, 2009) und regularisierter Regression (Hoerl/Kennard, 1970; Tibshirani, 1996) verallgemeinert, die Bertsimas/Copenhaver (2018) erstmals charakterisierten. Auf dieser Basis können robuste Schätzverfahren hergeleitet werden, welche Area-Indikatoren mithilfe unsicherer Daten verlässlich quantifizieren. Darauf aufbauend ist es sogar

möglich, eine konservative Abschätzung des mittleren quadratischen Vorhersagefehlers (Mean Squared Error; MSE) vorzunehmen. Dies ist für die praktische Anwendung von zentraler Bedeutung, da die Schwankung der Schätzwerte ein wichtiger Indikator der Schätzqualität ist.

Das folgende Kapitel 2 charakterisiert das Robustheitskonzept, welches den neuen SAE-Verfahren zugrunde liegt. Auf dieser Basis wird anschließend eine robuste Variante des Fay-Herriot-Modells (Fay/Herriot, 1979) vorgestellt. Kapitel 3 präsentiert eine Simulationsstudie, welche die vorgeschlagene Methodik testet. Der Aufsatz schließt in Kapitel 4 mit einem Fazit und einem Ausblick.

Der Artikel behandelt ausgewählte Aspekte der Dissertation des Autors (Krause, 2019). Tiefergehende mathematische Details sowie weitere Eigenschaften regularisierter Regression in Small Area Estimation enthalten die Dissertation sowie Burgard und andere (2021a).

2

Robuste Small Area Estimation

2.1 Regularisierung und Robustheit

Der Begriff Robustheit wird innerhalb der Statistik in verschiedenen Kontexten verwendet. Deswegen wird zunächst die im Artikel verwendete Robustheit charakterisiert. Auf dieser Basis ist es möglich, anschließend in Kapitel 3 einen robusten SAE-Ansatz herzuleiten.

Ein Regressionsproblem ist im Allgemeinen wie folgt definiert: Es gibt eine Zielvariable Y, deren beobachtete Werte in einem Vektor $\boldsymbol{y} \in \mathbb{R}^D$ enthalten sind. Diese sollen anhand einer Menge an Hilfsvariablen $X = \{X_1, \dots, X_p\}$ beschrieben werden, deren Werte mit einer Matrix $\boldsymbol{X} \in \mathbb{R}^{D \times p}$ notiert werden. Im Falle von parametrischer Inferenz wird angenommen, dass es einen Regressionsparametervektor $\boldsymbol{\beta} \in \mathbb{R}^p$ gibt, welcher die Beziehung von Y und X beschreibt. Dieser soll auf Basis der verfügbaren Daten geschätzt werden. Nun wird eine Verlustfunktion $g: \mathbb{R}^D \to \mathbb{R}_+$ definiert, welche umso näher an Null liegt, je besser $\boldsymbol{y}$ auf Basis von $\boldsymbol{X}$ mithilfe von β beschrieben werden kann. Im Falle von linearer Regression könnte dies etwa die Summe der Residualquadrate sein. Der Vektor β wird so gewählt, dass die Verlustfunktion möglichst nahe an Null liegt:

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^p} g(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) \tag{1}$$

Gehen wir nun davon aus, dass die Hilfsvariablenwerte fehlerhaft gemessen werden. Anstelle von $\boldsymbol{X}$ beobachtet man $\widetilde{\boldsymbol{X}} = \boldsymbol{X} + \boldsymbol{E}$, wobei $\boldsymbol{E} \in \mathbb{R}^{D \times p}$ eine Matrix mit unbekannten Messfehlern ist. Es nicht möglich, die individuellen Komponenten zu ermitteln. Ferner sind keine Aussagen über die Verteilung der Messfehler möglich. Im Hinblick auf Kapitel 1 spiegelt dies die Problematik vieler moderner Datenquellen wider. Man erhält das Minimierungsproblem

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^p} g(\boldsymbol{y} - \widetilde{\boldsymbol{X}}\boldsymbol{\beta}). \tag{2}$$

Die unbekannte Kontaminierung in $\widetilde{\boldsymbol{X}}$ muss bei der Schätzung der Regressionsparameter berücksichtigt werden, um valide Ergebnisse zu erhalten. An dieser Stelle bietet das Feld der robusten Optimierung einen hilfreichen Ansatz. Es wird eine Perturbationsmatrix $\boldsymbol{\Delta} \in \mathbb{R}^{D \times p}$ eingeführt, um die Unsicherheit in $\widetilde{\boldsymbol{X}}$ zu repräsentieren. Diese Perturbationsmatrix stammt aus einer Unsicherheitsmenge $\mathbb{U} \subseteq \mathbb{R}^{D \times p}$, welche Vermutungen widerspiegelt, wie die Messfehler strukturiert sein könnten. Da jedoch keine Verteilungsannahmen bezüglich der Messfehler gerechtfertigt werden können, ist es sinnvoll, diese möglichst konservativ zu betrachten. Das Minimierungsproblem sieht dann wie folgt aus:

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^p} \max_{\boldsymbol{\Delta} \in \mathbb{U}} g\big(\boldsymbol{y} - (\widetilde{\boldsymbol{X}} + \boldsymbol{\Delta})\boldsymbol{\beta}\big) \tag{3}$$

Die Verlustfunktion wird wie zuvor durch die Wahl der Regressionsparameter β minimiert. Zeitgleich wird jedoch auch die Perturbationsmatrix $\boldsymbol{\Delta}$ maximal im Hinblick auf die Unsicherheitsmenge $\mathbb{U}$ gewählt. Vereinfacht ausgedrückt wird von einem Worst-Case-Szenario bezüglich der Messfehler ausgegangen. Der Vektor β wird unter der Annahme des maximal möglichen Messfehlers gegeben $\mathbb{U}$ optimal gewählt. Somit sind die geschätzten Regressionsparameter robust gegen Messfehler einer bestimmten Größenordnung, jedoch unabhängig von deren Verteilung. Diese Perspektive auf Robustheit wurde unter anderem von El Ghaoui/Lebret (1997) sowie Ben-Tal und anderen (2009) studiert. Im Folgenden wird dieser Ansatz Min-Max-Robustifizierung genannt.

Eine wichtige Frage ist nun, wie sich dieser Ansatz praktisch anwenden lässt, also wie das skizzierte Regressionsproblem gelöst werden kann. Hier kann auf ein Ergebnis von Bertsimas/Copenhaver (2018) aufgebaut werden. Es zeigt, dass unter bestimmten Umständen Min-Max-Robustifizierung eine unerwartete Verbindung zu regularisierter Regression aufweist. Regularisierte Regression hat die Form

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^p} g(\boldsymbol{y}-\widetilde{\boldsymbol{X}}\boldsymbol{\beta})+\lambda h(\boldsymbol{\beta}). \tag{4}$$

Hierbei ist $\lambda > 0$ ein Tuning-Parameter und $h\colon \mathbb{R}^p \to \mathbb{R}_+$ eine Regularisierung. Diese Problemart ist in der Literatur seit einigen Jahren bekannt. Regularisierte Regression lässt sich schnell und effizient lösen, was insbesondere bei großen Datensätzen ein wichtiger Aspekt ist. In Krause (2019) sowie Burgard und anderen (2021a) wird diese Verbindung verallgemeinert. Es kann gezeigt werden, dass

$$\begin{aligned}\operatorname*{argmin}_{\boldsymbol{\beta}\in\mathbb{R}^p} \max_{\boldsymbol{\Delta}\in\mathbb{U}} g\left(\boldsymbol{y}-\left(\widetilde{\boldsymbol{X}}+\boldsymbol{\Delta}\right)\boldsymbol{\beta}\right) \\ = \operatorname*{argmin}_{\boldsymbol{\beta}\in\mathbb{R}^p} f\left(g\left(\boldsymbol{y}-\widetilde{\boldsymbol{X}}\boldsymbol{\beta}\right)\right)+\sum_{l=1}^{L}\lambda_l f_l\left(h_l(\boldsymbol{\beta})\right)\end{aligned} \tag{5}$$

für

$$\mathbb{U} = \{\boldsymbol{\Delta}\in\mathbb{R}^{D\times p}\colon g(\boldsymbol{\Delta\gamma}) \le \sum_{l=1}^{L}\varphi_l h_l(\boldsymbol{\gamma})\ \forall\,\boldsymbol{\gamma}\in\mathbb{R}^p\} \tag{6}$$

gilt, wenn $f, f_1, \ldots, f_L\colon \mathbb{R}_+ \to \mathbb{R}_+$ monoton steigende konvexe Funktionen und $h_1, \ldots, h_L\colon \mathbb{R}^P \to \mathbb{R}_+$ Normen sind. Es wird deutlich, dass Min-Max-Robustifizierung äquivalent zu regularisierter Regression ist. Die Äquivalenz gilt für viele Verfahren, welche bereits seit Jahren in der Statistik verwendet werden. Ein Beispiel hierfür ist die Ridge-Regression (Hoerl/Kennard, 1970). Sie ergibt sich für $L=1$ durch $g(\boldsymbol{z}) = \|\boldsymbol{z}\|_2, h_1(\boldsymbol{z}) = \|\boldsymbol{z}\|_2, f(z) = z^2, f_1(z) = z^2$. Im Hinblick auf Robustheit erhält man

$$\begin{aligned}\widehat{\boldsymbol{\beta}}_{ridge} &= \operatorname*{argmin}_{\boldsymbol{\beta}\in\mathbb{R}^p} \max_{\boldsymbol{\Delta}\in\mathbb{U}_{ridge}} \|\boldsymbol{y}-\left(\widetilde{\boldsymbol{X}}+\boldsymbol{\Delta}\right)\boldsymbol{\beta}\|_2 \\ &= \operatorname*{argmin}_{\boldsymbol{\beta}\in\mathbb{R}^p} \left\|\boldsymbol{y}-\widetilde{\boldsymbol{X}}\beta\right\|_2^2+\lambda\|\boldsymbol{\beta}\|_2^2\end{aligned} \tag{7}$$

mit

$$\mathbb{U}_{ridge} = \{\boldsymbol{\Delta}\in\mathbb{R}^{D\times p}\colon \|\boldsymbol{\Delta\gamma}\|_2 \le \varphi\|\boldsymbol{\gamma}\|_2 \forall\,\boldsymbol{\gamma}\in\mathbb{R}^p\}. \tag{8}$$

Für den Least Absolute Shrinkage and Selection Operator (LASSO; Tibshirani, 1996) gilt $L=1$ mit $g(\boldsymbol{z}) = \|\boldsymbol{z}\|_2, h_1(\boldsymbol{z}) = \|\boldsymbol{z}\|_1, f(z) = z^2, f_1(z) = z$.
Es folgt daraus

$$\begin{aligned}\widehat{\boldsymbol{\beta}}_{lasso} &= \operatorname*{argmin}_{\boldsymbol{\beta}\in\mathbb{R}^p} \max_{\boldsymbol{\Delta}\in\mathbb{U}_{lasso}} \|\boldsymbol{y}-\left(\widetilde{\boldsymbol{X}}+\boldsymbol{\Delta}\right)\boldsymbol{\beta}\|_2 \\ &= \operatorname*{argmin}_{\boldsymbol{\beta}\in\mathbb{R}^p} \left\|\boldsymbol{y}-\widetilde{\boldsymbol{X}}\beta\right\|_2^2+\lambda\|\boldsymbol{\beta}\|_1\end{aligned} \tag{9}$$

mit

$$\mathbb{U}_{lasso} = \{\boldsymbol{\Delta}\in\mathbb{R}^{D\times p}\colon \|\boldsymbol{\Delta\gamma}\|_2 \le \varphi\|\boldsymbol{\gamma}\|_1 \forall\,\boldsymbol{\gamma}\in\mathbb{R}^p\}. \tag{10}$$

Es ist also möglich, regularisierte Regression einzusetzen, um Schätzwerte zu erzeugen, welche robust gegen Messfehler unbekannter Verteilung sind. Damit eignen sich diese Verfahren für unsichere Daten, wie etwa Website- und Social-Media-Daten. Die Unsicherheitsmenge $\mathbb{U}$ determiniert dabei, wie groß die Messfehler sein dürfen, ohne die Optimalität der geschätzten Regressionsparameter zu beeinträchtigen. Dies geschieht durch den Unsicherheitsparameter $\varphi > 0$, welcher eine nicht triviale positive Verbindung zu dem Tuning-Parameter $\lambda > 0$ hat. Diese Verbindung wird in Krause (2019) sowie Burgard und andere (2021a) ausführlich charakterisiert. Die Intuition ist jedoch, dass je höher der Tuning-Parameter gewählt wird, desto größer sind die zulässigen Messfehler, und desto konservativer sind die Schätzergebnisse.

Ein interessanter Effekt ist, dass die Natur der Robustheit sich je nach Art der Regularisierung unterscheidet. Ridge-Regression verwendet die quadrierte ℓ_2-Norm als Regularisierungsterm. Dadurch wird in der Unsicherheitsmenge $\mathbb{U}_{ridge}$ der maximale Singulärwert in $\boldsymbol{\Delta}$ restringiert. Es wird also eine allgemeine Schranke für die Messfehler in $\widetilde{\boldsymbol{X}}$ eingeführt. Der LASSO verwendet die ℓ_1-Norm als Regularisierungsterm. Dadurch wird in der Unsicherheitsmenge $\mathbb{U}_{lasso}$ die maximale ℓ_2-Norm jeder einzelnen Spalte in $\boldsymbol{\Delta}$ restringiert. Somit wird eine Schranke für jede einzelne Hilfsvariable in $\widetilde{\boldsymbol{X}}$ eingeführt. Die Eigenschaft der Robustheit als solches ist jedoch in beiden Fällen gegeben.

2.2 Robustes Fay-Herriot-Modell

Im Folgenden wird eine robuste Variante des Fay-Harriot-Modells präsentiert, welche auf dem in Abschnitt 2.1 beschriebenen Robustheitskonzept basiert. Es sei U eine Population der Größe N, welche in D disjunkte

Areas U_d, $d = 1, \ldots, D$, der Größe N_d unterteilt ist. Wir nehmen an, dass eine Zufallsstichprobe $S \subset U$ aus der Population gezogen wird. Vereinfachend gehen wir hierbei davon aus, dass die Stichprobe in jeder Area eine Teilstichprobe $S_d \subset U_d$ der Größe $n_d > 0$ aufweist. Es sei nun Y eine Zielvariable, anhand welcher der relevante Area-Indikator berechnet wird. Dies könnte etwa der area-spezifische Mittelwert von Y sein:

$$\mu_d = \frac{1}{N_d} \sum_{i \in U_d} y_{id}\,, \qquad d = 1, \ldots, D, \tag{11}$$

wobei y_{id} den Wert von Y der i-ten Person in Area U_d symbolisiert. Das Ziel ist $\bar{Y}_d$ für alle Areas auf Basis von S zu schätzen. Nehmen wir an, dass uns hierfür in jeder Area ein Datenpaar $(y_d, \boldsymbol{x}_d), d = 1, \ldots, D$, zur Verfügung steht. Der Wert y_d ist ein direkter design-basierter Schätzer von μ_d, welcher auf Basis der Teilstichprobendaten in S_d berechnet wurde. Dies könnte ein Horvitz-Thompson-Schätzer sein (Horvitz/Thompson, 1952). Der direkte Schätzer ist design-unverzerrt, es gilt also

$$y_d = \mu_d + \varepsilon_d, \qquad d = 1, \ldots, D, \tag{12}$$

wobei $\varepsilon_d \sim N(0, \sigma_e^2)$ ein normalverteilter Stichprobenfehler mit Erwartungswert 0 und Varianz $\sigma_{\varepsilon d}^2 = \sigma_{\varepsilon d}^2(n_d) > 0$, $d = 1, \ldots, D$ ist. Der Einfachheit halber wird angenommen, dass $\sigma_{\varepsilon d}^2$ in jeder Area bekannt ist. Wenn der Teilstichprobenumfang n_d groß genug ist, so ist σ_e^2 klein und y_d ist ein verlässlicher Schätzer von μ_d. Vor dem Hintergrund des SAE-Settings in Kapitel 1 ist dies jedoch in der Praxis oft nicht der Fall. Deshalb wird im klassischen Fay-Harriot-Modell angenommen, dass der Area-Indikator eine lineare Beziehung zu geeigneten Hilfsvariablen hat:

$$\mu_d = \boldsymbol{x}_d \boldsymbol{\beta} + v_d, \qquad d = 1, \ldots, D, \tag{13}$$

wobei $\boldsymbol{x}_d \in \mathbb{R}^{1 \times p}$ ein Vektor mit korrekt gemessenen Hilfsvariablenwerten in Area U_d ist. Der Term $v_d \sim N(0, \sigma_v^2)$ ist ein Random Intercept mit Varianz $\sigma_v^2 > 0$ auf Ebene der Areas. Es wird hier vereinfachend angenommen, dass v_d unabhängig von $\varepsilon_1, \ldots, \varepsilon_D$ ist. Das komplette Fay-Harriot-Modell kann dann als ein Random-Intercept-Modell dargestellt werden:

$$y_d = \boldsymbol{x}_d \boldsymbol{\beta} + v_d + \varepsilon_d, \qquad d = 1, \ldots, D. \tag{14}$$

Das klassische Fay-Harriot-Modell darf in unserem Fall jedoch nicht verwendet werden, da die Hilfsvariablenwerte mit unbekannten Messfehlern kontaminiert sind. Deswegen wird die Modellgleichung verändert, um das in Abschnitt 2.1 eingeführte Robustheitskonzept zu nutzen:

$$y_d = (\boldsymbol{x}_d + \mathbf{e}_d)\boldsymbol{\beta} + v_d + \varepsilon_d, \qquad d = 1, \ldots, D, \tag{15}$$

wobei $\mathbf{e}_d \in \mathbb{R}^{1 \times p}$ ein Messfehlervektor ist. Über alle Areas formuliert erhalten wir dann

$$\boldsymbol{y} = (\boldsymbol{X} + \mathbf{E})\boldsymbol{\beta} + \boldsymbol{v} + \boldsymbol{\varepsilon}, \tag{16}$$

wobei $\boldsymbol{v} = (v_1, \ldots, v_D)'$ und $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_D)'$. Nachdem das Modell aufgestellt ist, müssen nun seine Modellparameter $(\boldsymbol{\beta}', \sigma_v^2)'$ robust geschätzt werden. Hierfür definieren wir die Varianz-Matrix
$\boldsymbol{V}(\sigma_v^2) = \operatorname{diag}(\sigma_{\varepsilon 1}^2 + \sigma_v^2, \ldots, \sigma_{\varepsilon D}^2 + \sigma_v^2) \in \mathbb{R}_+^{D \times D}$.
Die Modellparameterschätzung wird anschließend in zwei Teile unterteilt: die Schätzung der Regressionsparameter β und die Schätzung der Varianz σ_v^2. Im Folgenden wird von einer Robustifizierung durch Ridge-Regression ausgegangen.

1. Schritt: Schätzung der Regressionsparameter β

Es wird zunächst ein Startwert $\hat{\sigma}_v^{2[0]}$ definiert und die Varianz-Matrix $\boldsymbol{V}\left(\hat{\sigma}_v^{2[0]}\right) = \widehat{\boldsymbol{V}}$ als fix betrachtet. Somit erhält man das gewichtete regularisierte Regressionsproblem

$$\begin{aligned} \widehat{\boldsymbol{\beta}} &= \operatorname*{argmin}_{\boldsymbol{\beta} \in \mathbb{R}^p} \max_{\boldsymbol{\Delta} \in \mathbb{U}} \left\| \widehat{\boldsymbol{V}}^{-\frac{1}{2}} \left[\boldsymbol{y} - \left(\widetilde{\boldsymbol{X}} + \boldsymbol{\Delta} \right) \boldsymbol{\beta} \right] \right\|_2 \\ &= \operatorname*{argmin}_{\boldsymbol{\beta} \in \mathbb{R}^p} \left\| \widehat{\boldsymbol{V}}^{-\frac{1}{2}} \left[\boldsymbol{y} - \left(\widetilde{\boldsymbol{X}} + \boldsymbol{\Delta} \right) \boldsymbol{\beta} \right] \right\|_2^2 + \|\boldsymbol{\beta}\|_2^2. \end{aligned} \tag{17}$$

Unter Ridge-Regression hat $\widehat{\boldsymbol{\beta}}$ eine geschlossene Lösung. Wird jedoch der LASSO oder ein anderes Verfahren mit ℓ_1-Norm verwendet, so muss die Lösung iterativ bestimmt werden, etwa durch Gradientenabstiegsverfahren.

2. Schritt: Schätzung der Varianz σ_v^2

Die Varianz wird bedingt auf die robuste Lösung $\widehat{\boldsymbol{\beta}}$ durch eine Maximum-Likelihood-Schätzung bestimmt. Hierfür wird die logarithmierte Likelihood-Funktion minimiert:

$$\hat{\sigma}_v^2 = \operatorname*{argmin}_{\sigma_v^2 \in \mathbb{R}_+} -\frac{1}{2} \left[\log(2\pi) + \log(|\boldsymbol{V}(\sigma_v^2)|) + \hat{\boldsymbol{r}}' \boldsymbol{V}^{-1} \hat{\boldsymbol{r}} \right] \tag{18}$$

Dabei ist $\hat{\boldsymbol{r}} = \boldsymbol{y} - \widetilde{\boldsymbol{X}}\beta$. Die Lösung kann iterativ durch einen Newton-Raphson-Algorithmus gefunden werden.

Es wird deutlich, dass Schritt 1 und Schritt 2 jeweils voneinander abhängig sind. Das Ergebnis von Schritt 1 geht in Schritt 2 ein, dessen Ergebnis wiederum in Schritt 1 eingeht. Folglich werden bei der Modellparameterschätzung beide Schritte immer wieder bedingt aufeinander wiederholt, bis sich Konvergenz einstellt. In Krause (2019) sowie Burgard und andere (2021a) wird gezeigt, dass dieser Ansatz konsistente Modellparameterschätzung erlaubt.

Sind die geschätzten Modellparameter $\widehat{\boldsymbol{\beta}}$ und $\hat{\sigma}_v^2$ einmal bestimmt, kann auf ihrer Basis nun der Area-Indikator geschätzt werden. Hierfür wird schließlich folgender Prädiktor verwendet:

$$\hat{\mu}_d = \hat{\gamma}_d y_d + (1-\hat{\gamma}_d)(\boldsymbol{x}_d + \mathbf{e}_d)\widehat{\boldsymbol{\beta}}, \tag{19}$$

$$\hat{\gamma}_d = \frac{\hat{\sigma}_v^2}{\hat{\sigma}_v^2 + \sigma_{\varepsilon d}^2}, \qquad d = 1, \dots, D.$$

3

Simulation

3.1 Simulationsaufbau

Das im Folgenden präsentierte Simulationsexperiment verdeutlicht die Effektivität der beschriebenen Methodik. Hierfür wird eine Monte-Carlo-Simulation mit $R = 500$ Iterationen durchgeführt, welche mit $r = 1, \dots, R$ indiziert sind. Für die Simulation wird eine synthetische Population unter dem klassischen Fay-Harriot-Modell ohne Messfehler erzeugt. Anschließend wird eine Datenbasis generiert, welche Stichproben- und Messfehler wie in Abschnitt 2.2 beschrieben enthält. Auf dieser Basis sollen die Werte der Area-Indikatoren in der synthetischen Population geschätzt werden. Hierbei werden die Performance des klassischen Fay-Harriot-Modells (FH) unter Maximum-Likelihood-Schätzung sowie seiner robusten Variante unter Ridge-Regression (ℓ_2-FH) verglichen. Die Simulation erfolgt nach dem folgenden Algorithmus:

1. Definiere $D \in \{50, 100\}$, $\boldsymbol{\beta} = (2,2,2)'$, $\boldsymbol{\mu}_{\boldsymbol{X}} = (2,2,2), \sigma_v^2 = 1$.
2. Ziehe $\boldsymbol{X} \sim N(\boldsymbol{\mu}_{\boldsymbol{X}}, \boldsymbol{I}_3)$ und $\boldsymbol{E} \sim \mathrm{F}(\mathbf{0}_3, \boldsymbol{\Sigma})$, wobei F eine Messfehlerverteilung ist.
3. Ziehe $\sigma_{\varepsilon d}^2 \sim U(30, 40), d = 1, \dots, D$.
4. Kontaminiere die Hilfsvariablenwerte $\widetilde{\boldsymbol{X}} = \boldsymbol{X} + \boldsymbol{E}$.
5. Für $r = 1, \dots, R$:
   a. Ziehe $v_d^{[r]} \sim N(0, \sigma_v^2)$ und
   $$\varepsilon_d^{[r]} \sim N(0, \sigma_{\varepsilon d}^2), d = 1, \dots, D.$$
   b. Erzeuge $\mu_d^{[r]} = \boldsymbol{x}_d\boldsymbol{\beta} + v_d^{[r]}$ und
   $$y_d^{[r]} = \mu_d^{[r]} + \varepsilon_d^{[r]}, d = 1, \dots, D.$$
   c. Berechne $\widehat{\boldsymbol{\beta}}^{[r]}$ und $\hat{\sigma}_v^{2[r]}$ anhand der Datenpaare $\left(y_d^{[r]}, \widetilde{\boldsymbol{x}}_d\right), d = 1, \dots, D$.
   d. Berechne $\mu_d^{[r]}$, $d = 1, \dots, D$, anhand von $\widehat{\boldsymbol{\beta}}^{[r]}$ und $\hat{\sigma}_v^{2[r]}$.
6. Berechne die Genauigkeitsmaße.

Für die Messfehlerverteilung F werden vier Szenarien angenommen:

- Szenario 1: keine Messfehler,
- Szenario 2: symmetrische schwach-korrelierte Fehler,
- Szenario 3: symmetrische stark-korrelierte Fehler,
- Szenario 4: asymmetrische Fehler.

Die Genauigkeitsmaße für die Schätzung der Area-Indikatoren sind der Relative Bias (RBias) und der Relative Root Mean Squared Error (RRMSE):

$$\mathrm{RBias}(\hat{\mu}) = \frac{1}{D \cdot R} \sum_{d=1}^{D} \sum_{r=1}^{R} \frac{\left(\hat{\mu}_d^{[r]} - \mu_d^{[r]}\right)}{\bar{\mu}_d}, \tag{20}$$

$$\mathrm{RRMSE}(\hat{\mu}) = \frac{1}{D \cdot R} \sum_{d=1}^{D} \sum_{r=1}^{R} \frac{\sqrt{\left(\hat{\mu}_d^{[r]} - \mu_d^{[r]}\right)^2}}{\bar{\mu}_d}, \tag{21}$$

wobei $\bar{\mu}_d$ der Mittelwert des Area-Indikators für U_d über alle Iterationen ist. Der RBias gibt an, inwiefern die Schätzwerte das generelle Niveau des wahren Parameters treffen. Er ist ein Maß für das generelle Verhalten eines Schätzers. Der RRMSE misst die Effizienz der Schätzung unter Berücksichtigung von Bias und Varianz. Je kleiner dieser Wert ist, desto genauer sind die Schätzergebnisse des jeweiligen Verfahrens. Für die SAE-Praxis ist dieses Maß von größerer Bedeutung als der RBias, da er die Schätzgenauigkeit quantifiziert.

3.2 Simulationsergebnisse

↘ Tabelle 1 präsentiert die Ergebnisse der Area-Indikatorschätzung. Anhand des RBias wird deutlich, dass der klassische Fay-Harriot-Ansatz insgesamt weniger verzerrte Ergebnisse produziert. Schaut man jedoch auf den RRMSE, wird deutlich, dass ℓ_2-FH in jedem Szenario die effizienteren Schätzungen liefert. Dieses Resultat ist vor dem Hintergrund der Literatur über Ridge-Regression erwartbar. In der Abwesenheit von Messfehlern haben Hoerl/Kennard (1970) gezeigt, dass die quadrierte ℓ_2-Norm als Regularisierungsterm zwar die Verzerrung der Schätzung erhöht, jedoch ihre Effizienz insgesamt verbessert. Entscheidend hierfür ist die Wahl des Regularisierungsparameters $\lambda > 0$. Er balanciert Verzerrung und Varianz bei der Schätzung, weswegen seine Wahl ein wichtiger Aspekt der Methodik ist. Dies wird in Kapitel 4 adressiert.

Tabelle 1
Ergebnisse der Area-Indikatorschätzung

	D	FH RBias($\hat{\mu}$)	ℓ_2-FH RBias($\hat{\mu}$)	FH RRMSE($\hat{\mu}$)	ℓ_2-FH RRMSE($\hat{\mu}$)
Szenario 1	50	– 0.0006	– 0.0437	0.1240	0.1139
Szenario 1	100	– 0.0016	– 0.0365	0.1003	0.0955
Szenario 2	50	– 0.0394	– 0.0841	0.2116	0.1984
Szenario 2	100	– 0.0299	– 0.0588	0.1771	0.1719
Szenario 3	50	– 0.0604	0.0917	0.2235	0.2189
Szenario 3	100	0.0355	– 0.0780	0.2097	0.1983
Szenario 4	50	– 0.1425	– 0.1528	0.3324	0.3310
Szenario 4	100	– 0.1439	– 0.1638	0.3288	0.3266

D: Testvariable; FH: Fay-Harriot; RBias: Relative Bias; RRMSE: Relative Root Mean Squared Error.

Grundsätzlich lässt sich also festhalten, dass die Regularisierung den Bias erhöht, um die Schätzvarianz zu reduzieren. Im Kontext von fehlerhaften Daten führt dies zu einer Verbesserung der Schätzergebnisse gemessen am RRMSE. Die Schätzungen sind insgesamt genauer, was vor dem Hintergrund der Ausgangssituation gewünscht war.

4 Fazit und Ausblick

Der Artikel thematisierte den Nutzen von regularisierter Regression für die Schätzung von regionalen Indikatoren auf Basis unsicherer Daten. Es wurde gezeigt, dass Regularisierung eine allgemeine Verbindung zu robuster Optimierung hat. Diese ermöglicht es aus der Literatur bereits bekannte SAE-Verfahren mittels Regularisierung so zu erweitern, dass sie robuste Ergebnisse in der Gegenwart unbekannter Messfehler produzieren. Vor diesem Hintergrund wurde eine robuste Variante des Fay-Harriot-Modells vorgestellt. Dessen Effektivität wurde anschließend in einer Simulationsstudie demonstriert.

Die dargelegte Verwendung von Regularisierung ist im Hinblick auf die statistische Literatur untypisch. Derartige Verfahren werden meist im Kontext hochdimensionaler Inferenz, Multikollinearität sowie Variablenselektionsproblemen eingesetzt. Doch die demonstrierte allgemeine Verbindung zu robuster Optimierung legt weitere Anwendungsfelder nahe. Die Studie verdeutlicht, dass regularisierte Regression robuste und konsistente Schätzergebnisse in der Gegenwart von Messfehlern unbekannter Verteilung liefert. Dies erlaubt die Einbeziehung moderner Datenquellen, wie etwa Website- und Social-Media-Daten, zur Verbesserung statistischer Modelle für die Schätzung regionaler Indikatoren. Somit kann perspektivisch die sozioökonomische Forschung erweitert werden, da nun eine größere Menge und Vielfalt von Daten nutzbar ist.

Ein wichtiger Aspekt für den Einsatz regularisierter Regressionsverfahren ist die Wahl des Tuning-Parameters. Im Kontext von Robustheit determiniert er durch seine Verbindung zu den Parametern der Unsicherheitsmenge die maximal zulässigen Messfehler, unter welchen die geschätzten Modellparameter optimal sind. In dem präsentierten Artikel wurde der Tuning-Parameter durch Kreuzvalidierung gewählt, wie es in der Literatur vielfach vorgeschlagen wird. Die Natur des Robustifizierungseffekts hängt von dem gewählten Regularisierungsterm ab. Daher liegt es nahe, dass sowohl Tuning-Parameter als auch Regularisierung optimal gegeben der Datenkontaminierung gewählt werden können. Dies ist jedoch Gegenstand künftiger Forschung.

LITERATURVERZEICHNIS

Ben-Tal, Aharon/El Ghaoui, Laurent/Nemirovski, Arkadi. *Robust optimization*. In: Princeton University Press. Band 28 der Reihe Princeton Series in Applied Mathematics.2009. DOI: 10.1515/9781400831050

Bertsimas, Dimitres/Copenhaver, Martin S. *Characterization of the equivalence of robustification and regularization in linear and matrix regression*. In: European Journal of Operational Research. Band 270. 2018, Seite 931 ff.

Burgard, Jan Pablo/Esteban, María Dolores/Morales, Domingo/Pérez, Agustín. *A Fay-Herriot model when auxiliary variables are measured with error*. In: TEST. Jahrgang 29. 2020a. Seite 166 ff. DOI: 10.1007/s11749-019-00649-3

Burgard, Jan Pablo/Esteban, María Dolores/Morales, Domingo/Pérez, Agustín. *Small area estimation under a measurement error bivariate Fay-Herriot model*. In: Statistical Methods and Applications. Online-first. 2020b. DOI: 10.1007/s10260-020-0051519.

Burgard, Jan Pablo/Krause, Joscha/Kreber, Dennis/Morales, Domingo. *The generalized equivalence of regularization and min-max robustification in linear mixed models*. In: Statistical Papers. Jahrgang 62. 2021a. Seite 2857 ff. DOI: 10.1007/s00362-020-01214-z

Burgard, Jan Pablo/Krause, Joscha/Morales, Domingo. *A measurement error Rao-Yu model for regional prevalence estimation over time using uncertain data obtained from dependent survey estimates*. In: TEST. Online-first. 2021b. DOI: 10.1007/s11749-021-00776-w.

El Ghaoui, Laurent/Lebret, Hervé. *Robust solutions to least-squared problems with uncertain data*. In: SIAM Journal on Matrix Analysis and Applications. Jahrgang 18. Ausgabe 4/1997, Seite 1035 ff. DOI: 10.1137/S0895479896298130

Krause, Joscha. *Regularization methods for statistical modelling in small area estimation.* Dissertation, Universität Trier. 2019. DOI: 10.25353/ubtr-xxxx-de9f-02c8.

Hoerl, Arthur E./Kennard, Robert W. *Ridge regression: Biased estimation for non-orthogonal problems*. In: Techometrics. Jahrgang 12. Ausgabe 1/1970, Seite 55 ff. Online publiziert: 2012. DOI: 10.1080/00401706.1970.10488634

Horvitz, Daniel G./Thompson, Donovan J. *A generalization of sampling without replacement from a finite universe*. In: Journal of the American Statistical Association. Ausgabe 47/1952, Seite 663 ff. Online publiziert: 2012. DOI: 10.1080/01621459.1952.10483446

Molina, Isabel/Rao, J. N. K./Datta, Gauri Sankar. *Small area estimation under a Fay-Herriot model with preliminary testing for the presence of random effects*. In: Survey Methodology. Jahrgang 41. Ausgabe 1/2015, Seite 1 ff.

Morales, Domingo/Esteban, María Dolores/Pérez, Agustín/Hobza, Tomáš. *A course on small area estimation and mixed models. Methods, theory and applications in R*. In: Statistics for Social and Behavioral Sciences. 2021a.

LITERATURVERZEICHNIS

Morales, Domingo/Krause, Joscha/Burgard, Jan Pablo. *On the use of aggregate survey data for estimation regional major depressive disorder prevalence*. In: Psychometrika. 2021b. Online-first. DOI: 10.1007/s11336-021-09808-8.

Rao, J. N. K./Molina, Isabel. *Small area estimation*. 2. Auflage. Wiley Series in Survey Methodology. 2015.

Tibshirani, Robert. *Regression shrinkage and selection via the lasso*. In: Journal of the Royal Statistical Society. Series B (Methodological). Jahrgang 58. Ausgabe 1/1996, Seite 267 ff. DOI: 10.1111/j.2517-6161.1996.tb02080.x

Torabi, Mahmoud/Datta, Gauri Sankar/Rao, J. N. K. *Empirical Bayes estimation of small area means under a nested error linear regression model with measurement errors in the covariates*. In: Scandinavian Journal of Statistics. Jahrgang 36. Ausgabe 2/2009, Seite 355 ff. DOI: 10.1111/j.1467-9469.2008.00623.x

Ybarra, Lynn M. R./Lohr, Sharon L. *Small area estimation when auxiliary information is measured with error*. In: Biometrika. Band 95. 2008. Seite 919 ff.

Herausgeber
Statistisches Bundesamt (Destatis), Wiesbaden

Schriftleitung
Dr. Daniel Vorgrimler
Redaktion: Ellen Römer

Ihr Kontakt zu uns
www.destatis.de/kontakt

Erscheinungsfolge
zweimonatlich, erschienen im Februar 2022
Ältere Ausgaben finden Sie unter www.destatis.de sowie in der Statistischen Bibliothek.

Artikelnummer: 1010200-22001-4, ISSN 1619-2907