AUSTRIAN JOURNAL OF STATISTICS
Volume 37 (2008), Number 3&4, 335–347

Variablenselektion bei gebundener Hochrechnung

Melanie Knobelspies[1] und Ralf Münnich[2]
[1]Interbrand Zintzmeyer & Lux AG, Schweiz
[2]Universität Trier, Deutschland

Zusammenfassung: Zur Verbesserung der Schätzqualität werden in Stichprobenerhebungen häufig vorhandene Zusatzinformationen in die Hochrechnung eingebunden, insbesondere bei der Verwendung von verallgemeinerten Regressionsschätzern. In einigen Erhebungen, wie in der Einkommens- und Verbrauchsstichprobe oder im deutschen Survey on Income and Living Conditions (Leben in Europa; D-SILC) stehen hierfür sehr viele Variablen zur Verfügung. In der Praxis muss man sich jedoch aus methodisch-technischen Gründen auf eine geeignete Auswahl von Hilfsvariablen beschränken.

In diesem Aufsatz wird der Frage nachgegangen, inwieweit statistisch-ökonometrische Variablenselektionsverfahren bei der Auswahl geeigneter Hilfsvariablen herangezogen werden können. Es wird insbesondere untersucht, ob die Effizienz eines Regressionsmodells, die das Ziel solcher Selektionsverfahren ist, auch gleichzeitig die Genauigkeit der eigentlich interessierenden Stichprobenschätzung optimiert. Darüber hinaus wird untersucht, inwiefern spezielle geschichtete Stichprobenziehungen die Ergebnisse beeinflussen.

Abstract: In survey sampling, the adequate use of auxiliary variables may considerably improve the efficiency of survey estimates gained from generalized regression techniques. In some surveys like the German income and expenditure survey or the German survey of income and living conditions (D-SILC), a huge amount of potential candidates of auxiliary information is available. Due to methodological and numerical limitations, efficient variable selection need to be applied for gaining efficient estimates.

Within this paper, classical statistical variable selection procedures are studied in order to elaborate their efficiency for survey estimation problems. Special emphasis is put on optimizing the model for regression estimation techniques. Additionally, the influence of stratification and allocation on the results will be considered.

Keywords: Verallgemeinerter Regressionsschätzer, Hilfsinformationen, Hauptkomponentenregression, Schichtung.

1 Einleitung

In modernen Stichprobenverfahren steht vielfach die Einbindung geeigneter Hilfsinformationen in die Modelle im Vordergrund, mit dem Ziel eine Effizienzsteigerung der

Schätzung zu erreichen. Solche Informationen können auf verschiedene Weise eingebunden werden, etwa in das Stichprobendesign oder aber auch in Modelle, die im Rahmen der Schätzung verwendet werden. Weit verbreitet sind in diesem Zusammenhang die verallgemeinerten Regressionsschätzer oder Kalibrierungsmethoden.

Bei der Umsetzung sogenannter modellunterstützender Verfahren entsteht jedoch das Problem, welche Information in Form von Hilfsvariablen herangezogen werden soll. Vielfach stehen in der Praxis nur sehr wenige Variablen zur Verfügung, womit eine Auswahl höchstens eingeschränkt möglich ist. Im Allgemeinen gehören hierzu Variablen aus Personen- und Haushaltserhebungen, bei denen vielfach lediglich verschiedene demographische Merkmale wie Altersgruppe, Geschlecht und Nationalität verwendet werden können. In einigen neueren Erhebungen, wie etwa der deutschen Erhebung des europäischen Survey on Income and Living Conditions stehen darüber hinaus jedoch wesentlich mehr Hilfsinformationen zur Verfügung, so dass die Frage nach einer geeigneten Auswahl der zu verwendenden Variablen entsteht. Da die Erhebung D-SILC via deutschem Mikrozensus aus der Stichprobe befragungsbereiter Haushalte gezogen wird (vgl. Körner, Nimmergut, and Nökel, 2006, S. 452, oder Münnich, Huergo, Magg, and Ohly, 2005, S. 2f.), steht prinzipiell sogar eine dreistellige Anzahl möglicher Hilfsvariablen zur Verfügung.

Häufig erweist sich jedoch die Verwendung einer Vielzahl von Variablen als ungeeignet, da die Schätzungen aufgrund von Kollinearitäten in der Regressionsmatrix möglicherweise numerisch sehr instabil werden. Hinzu kommt, dass in Erhebungen eher kategoriale Variablen zur Verfügung stehen, bei denen, insbesondere im Zusammenhang mit kleinen Stichprobenumfängen, diese Probleme noch häufiger auftreten. Zur Vermeidung solcher Schwierigkeiten bzw. allgemeiner Fehlspezifikationen der Modelle sollten daher geeignete Hilfsvariablensets ermittelt werden, die stabile, aber auch effiziente Schätzungen unterstützen. Im Rahmen dieser Arbeit werden ausgewählte Variablenselektionsverfahren in Bezug auf ihren möglichen Effizienzgewinn beim verallgemeinerten Regressionsschätzer (GREG) hin untersucht. Dabei stehen klassische Stichprobenschätzungen, also die Schätzung von Mittel- und Totalwerten, im Vordergrund.

Im zweiten Kapitel werden zunächst die theoretischen Grundlagen aufgezeigt. Nach der Darstellung des verallgemeinerten Regressionsschätzers wird eine kurze Übersicht über die später verwendeten Variablenselektionsverfahren gegeben. Schließlich wird der Einsatz von Regressionen auf Basis von Hauptkomponenten aus Hilfsvariablen, der sogenannten Hauptkomponentenregression, als Grundlage für den verallgemeinerten Regressionsschätzer aufgezeigt.

Anschließend werden im dritten Kapitel die Ergebnisse einer Monte-Carlo-Simulationsstudie vorgestellt. Im Vordergrund stehen empirische Untersuchungen, inwieweit Variablenselektionsverfahren bei der Optimierung von Regressionsschätzverfahren helfen können. Hierbei wird überprüft, inwiefern sich ein Zusammenhang zwischen den Bewertungskriterien der Regressionsmodelle und der Schätzqualität bei der Regressionsschätzung erkennen lässt. Dabei interessiert, ob statistisch-ökonometrische Überlegungen für die Modellbildung automatisch auch einen Effizienzgewinn der Stichprobenschätzung verursachen.

Ebenso wird im Rahmen der Simulationsstudie der Einsatz der Hauptkomponentenregression auf die Regressionsschätzung untersucht. Mittels einer Hauptkomponenten-

regression können die ursprünglichen Hilfsvariablen zu neuen synthetischen Faktoren transformiert werden, welche daraufhin in der Regressionsschätzung eingesetzt werden. Im Gegensatz zur klassischen statistischen Vorgehensweise, bei der die Aussagekraft eines Modells untersucht wird und somit Hauptkomponentenregressionen weniger geeignet sind, spielen bei Stichprobenschätzungen lediglich Effizienzgewinne der eigentlich interessierenden Schätzung eine Rolle.

Da die statistisch-ökonometrischen Verfahren vielfach auf der uneingeschränkten Zufallsstichprobe basieren, soll zusätzlich die Robustheit der Ergebnisse in Bezug auf verschiedene geschichtete Stichprobenziehungen überprüft werden. Es folgt schließlich eine Zusammenfassung der Ergebnisse nebst Ausblick.

2 Methodische Grundlagen

2.1 Der verallgemeinerte Regressionsschätzer

Zur Schätzung eines unbekannten Totalwertes τ_Y einer interessierenden Variablen Y wird eine Stichprobe vom Stichprobenumfang n aus einer endlichen Population vom Umfang N gezogen. Dann ist der Horvitz-Thompson-Schätzwert (HT) durch

$$\hat{\tau}_{Y,\text{HT}} = \sum_{i=1}^{n} \frac{1}{\pi_i} y_i = \sum_{i=1}^{n} d_i y_i$$

definiert. Dabei bezeichnen π_i die Inklusionswahrscheinlichkeiten 1. Ordnung und $d_i := 1/\pi_i$ die zugehörigen Designgewichte (vgl. etwa Lohr, 1999, Kapitel 6).

Darüberhinaus werden mögliche Hilfsvariablen mit X bezeichnet. Unterstellt man in der Grundgesamtheit ein lineares Modell $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, wobei $\mathbf{X}$ die Matrix der Hilfsvariablen darstellt, dann kann man den verallgemeinerten Regressionsschätzer durch

$$\hat{\tau}_{Y,\text{GREG}} = \hat{\tau}_{Y,\text{HT}} + \underbrace{(\boldsymbol{\tau}_X - \hat{\boldsymbol{\tau}}_{X,\text{HT}})^T \hat{\boldsymbol{\beta}}}_{\text{Korrekturterm}}$$

angeben. Der verallgemeinerte Regressionsschätzer (GREG) lässt sich also als HT-Schätzer darstellen, der um die gewichtete Abweichung zwischen den bekannten Totalwerten $\boldsymbol{\tau}_X$ der Hilfsvariablen und den korrespondierenden HT-Schätzungen $\hat{\boldsymbol{\tau}}_X$ korrigiert wird. Im Falle mehrdimensionaler Hilfsinformationen sind die Totalwerte im Korrekturterm Vektoren. Von den Hilfsvariablen wird lediglich gefordert, dass deren Totalwerte aus der Grundgesamtheit, zum Beispiel aus vorangegangenen Untersuchungen oder aus externen Quellen, als bekannt vorliegen. Bei der Ermittlung der Regressionskoeffizienten $\hat{\boldsymbol{\beta}}$

$$\hat{\boldsymbol{\beta}} = \left(\sum_{i=1}^{n} \frac{\mathbf{x}_i \mathbf{x}_i^T}{\pi_i} \right)^{-1} \sum_{i=1}^{n} \frac{\mathbf{x}_i y_i}{\pi_i}$$

sollten ebenso die Designfaktoren berücksichtigt werden (vgl. etwa Särndal et al., 1992, S. 226). Die Betrachtungen lassen sich ebenso auf die allgemeinere Klasse der Kalibrierungsschätzer erweitern. Auf eine erweiterte Darstellung des GREG-Schätzers durch Kalibrierungsverfahren wird an dieser Stelle jedoch verzichtet, da die Untersuchungen bisher keine erwähnenswerte Unterschiede aufwiesen. Der interessierte Leser sei hier auf Deville and Särndal (1992) verwiesen.

Für eine uneingeschränkte Zufallsstichprobe ergibt sich für die Design-Gewichte speziell $d_i = N/n$ und für die geschichtete Zufallsstichprobe $d_i = N_h/n_h$, wobei das i-te Element gerade aus der h-ten Schicht entstammt. Eine explizite Darstellung von Varianzschätzungen wird hier nicht aufgeführt, da bei den klassischen Designs, die dieser Studie zugrundeliegen, geeigneterweise direkte Varianzschätzmethoden bzw. Residualvarianzschätzer verwendet werden (vgl. Lohr, 1999, S. 226, oder Münnich, 2008). Ebenso sei angemerkt, dass sich die Untersuchungen auf Grund der allgemeinen Darstellung problemlos auf komplexere Stichprobendesigns erweitern lassen. Ferner wird ohne weitere Einschränkungen das Ziehungsmodell ohne Zurücklegen unterstellt.

Der Effizienzgewinn durch Anwendung der Regressionsschätzverfahren hängt im wesentlichen von der Anzahl und Struktur der Hilfsvariablen ab. Um für eine verfügbare Datenmenge jene Variablen zu selektieren, die im multivariaten Kontext einer Regression einen besonders positiven Einfluss auf die Zielgröße haben, existieren in der Literatur verschiedene Verfahren. Im nächsten Abschnitt wird hierbei im Rahmen der Regressionsanalyse eine Übersicht über die Verfahren der *vollständigen Modellselektion* sowie der *sequentiellen Prozeduren*, wie die *Vorwärtsselektion* oder die *Rückwärtselimination*, gegeben. Tiefer gehende Darstellungen über diese Verfahren findet der Leser beispielsweise in Draper and Smith (1981), Harrell (2001), Weisberg (2005) sowie in zahlreichen weiteren Werken.

2.2 Vollständige Modellselektion

Im Folgenden wird von k zu untersuchenden Hilfsvariablen im Regressionsmodell ausgegangen. Bei k zu untersuchenden Variablen müssen bei der vollständigen Modellselektion insgesamt $2^k - 1$ mögliche Modelle betrachtet werden. Zur Bewertung der Modelle werden im allgemeinen zahlreiche unterschiedliche Gütekriterien herangezogen. Das wohl bekannteste Gütekriterium ist das adjustierte Bestimmtheitsmaß R^2_{adj}. Es setzt die Summe der quadrierten Residuen (SSR) ins Verhältnis zur Gesamtstreuung (SST) und berücksichtigt zusätzlich die Anzahl der Freiheitsgrade

$$R^2_{adj} = 1 - \frac{\text{SSR}/(n-k-1)}{\text{SST}/(n-1)}\,.$$

Eine weitere Möglichkeit zur Modellbewertung stellt die Betrachtung des geschätzten quadrierten Vorhersagefehlers Γ_p dar. Bei einem Modell mit $p < k$ Variablen kann dieser über

$$\hat{\Gamma}_p = \hat{\sigma}^2(2(p+1) - n) + \text{SSR}_p$$

abgeschätzt werden, wobei $\hat{\sigma}^2$ für die Schätzung der Fehlervarianz der Störterme σ^2 steht und SSR_p für die Summe der quadrierten Residuen im betrachteten Modell. Mittels Division durch $\hat{\sigma}^2$ ergibt sich hieraus das Informationskriterium C_p von Mallows (vgl. Mallows, 1973):

$$C_p = (2(p+1) - n) + \frac{\text{SSR}_p}{\hat{\sigma}^2}\,.$$

Nach Draper and Smith (1981, S. 300) gilt $\text{E}(\text{SSR}_p) = (n-p-1)\sigma^2$ bei einem geeigneten Modell, woraus $\text{E}(C_p) \approx p+1$ folgt. Hieraus ist zu erkennen, dass bei einem nach diesem

Kriterium geeignet zu wählendem Modell die Anzahl der zu schätzenden Parameter dem Wert des Informationskriteriums C_p entspricht. Große Differenzen zwischen C_p und $p+1$ können demzufolge zur Messung der Verzerrung herangezogen werden.

Die Informationskriterien nach Akaike und Bayes kombinieren hingegen die Devianz mit einem Strafterm. Die Devianz ergibt sich als (-2)-facher maximierter Wert der logarithmierten Likelihood-Funktion $\mathcal{L}$. Mit zunehmender Komplexität des Modells wird auch der Wert des Strafterms größer. Das Modell soll somit möglichst einfach sein und gleichzeitig einen hohen Likelihood-Wert erzielen. Als allgemeine Formel erhält man:

$$\text{Kriterium} = \underbrace{-2\mathcal{L}_p}_{\text{Devianz}} + \eta \times \text{Strafterm}\,,$$

wobei η für das relative Gewicht der Modellkomplexität zur Modellanpassung steht (vgl. Reineking and Schröder, 2004, S. 40).

Für das Akaike Informationskriterium (AIC), welches auf Akaike (1974) zurückgeht, setzt man $\eta = 2$ und betrachtet die Anzahl p der verwendeten Regressoren als Modellkomplexität. Verwendet man die Log-Likelihood-Funktion des linearen Regressionsmodells unter der Annahme normalverteilter Fehler

$$\mathcal{L}(\boldsymbol{\beta}, \sigma^2; \mathbf{y}) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2}\frac{(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})}{\sigma^2}\,,$$

so erhält man nach Maximierung dieser Funktion bezüglich der Maximum-Likelihood-Schätzer von $\boldsymbol{\beta}$ und σ^2 das Informationskriterium nach Akaike durch

$$\text{AIC} = -2\mathcal{L}\Big(\hat{\boldsymbol{\beta}}, \frac{n-p}{n}\hat{\sigma}^2; \mathbf{y}\Big) + 2p\,.$$

Das Akaike Informationskriterium tendiert dazu, Modelle mit vielen Parametern zu bevorzugen (Fahrmeir et al., 2007, S. 163). Alternativ dazu eignet sich die Verwendung des Bayesianischen Informationskriteriums (BIC) von Schwarz (1978), welches $\eta = \log(n)$ als Gewicht für die Modellkomplexität verwendet:

$$\text{BIC} = -2 \cdot \mathcal{L} + p \cdot \log(n)\,.$$

Ab einem Stichprobenumfang von $n = 8$ fällt damit der Strafterm des BIC größer als der Strafterm des AIC aus, womit unter Verwendung des BIC im Allgemeinen weniger komplexe Modelle als beim AIC resultieren.

2.3 Schrittweise Verfahren

Neben der Betrachtung aller $2^k - 1$ Modelle bieten die schrittweisen Verfahren eine hilfreiche Alternative zur vollständigen Variablenselektion (vgl. Weisberg, 2005, S. 211ff.). Unter diesem Begriff verbirgt sich ein diskreter Prozess, in welchem jeweils eine Variable dem Modell hinzugefügt bzw. aus dem Modell entfernt wird, bis ein vorgegebenes Bewertungskriterium sich nicht mehr verbessern lässt.

Bei der Vorwärtsselektion startet man mit dem Nullmodell $y_i = \beta_0 + \varepsilon_i$, welches keine erklärenden Variablen enthält. Die Schätzung des Niveauparameters β_0 ergibt sich

aus dem arithmetischen Mittel aller Beobachtungen $\overline{y}$. In jeder Iteration dieses Verfahrens wird jeweils diejenige Variable aufgenommen, welche die größte Verbesserung eines Modellwahlkriteriums, also AIC, BIC, R^2 oder C_p, liefert. Sollte die Verbesserung durch die Aufnahme dieser neuen Variable nicht mehr signifikant im Sinne der t-Statistik sein, wird diese aus dem Modell entfernt und das Verfahren bricht ab. In jedem Schritt wird bei diesem Verfahren die Anzahl der Regressoren folglich um eins erhöht.

Bei der Rückwärtselimination startet man, im Gegensatz zur Vorwärtsselektion, mit dem maximalen Modell, das alle verfügbaren erklärenden Variablen enthält. Der Reihe nach werden nun diejenigen Variablen aus dem Modell eliminiert, die keinen signifikanten Beitrag zur Verbesserung des Modells liefern. Auch hier ergibt sich die Stoppregel analog zur Vorwärtsselektion.

2.4 Hauptkomponentenregression

Bei der Verwendung sehr vieler Hilfsvariablen lassen sich gewisse lineare Abhängigkeiten, so genannte Multikollinearitäten, unter den Variablen kaum vermeiden. Die obigen Verfahren der Variablenselektion liefern dann im Allgemeinen keine zufriedenstellenden Ergebnisse mehr. Ebenso verlieren die Maße zur Modellselektion an Aussagekraft und es resultieren unpräzise Schätzungen für den Parametervektor $\hat{\boldsymbol{\beta}}$. Im Extremfall der perfekten Multikollinearität ist der Parametervektor $\hat{\boldsymbol{\beta}}$ im linearen Regressionsmodell sogar nicht mehr schätzbar. Um diesem Problem entgegenzuwirken, stellt die Methode der Hauptkomponentenregression eine hilfreiche Alternative zur klassischen KQ-Schätzung dar. Die Regressoren, die zur Modellierung der Zielvariablen in das Regressionsmodell mit eingehen, werden hierbei in neue synthetische Variablen mittels Hauptkomponentenanalyse transformiert. Diese Variablen werden anschließend als Hilfsvariablen bei der Regressionsschätzung verwendet.

Die Hauptkomponenten werden hierbei so bestimmt, dass ihre jeweilige Varianz maximal ist und sie paarweise orthogonal, und damit empirisch unkorreliert, sind. Die Regression findet daraufhin im neuen Faktorenraum statt, wobei in der Regel nicht alle Faktoren in das Modell mit aufgenommen werden (vgl. Fahrmeir et al., 2007, S. 172). Da die Eigenwerte den Erklärungsanteil eines Faktors in Hinblick auf die Varianz aller erklärenden Variablen beschreiben, wird häufig gefordert, dass alle Faktoren mit einem Eigenwert größer eins im Regressionsmodell berücksichtigt werden. Da beim GREG-Schätzer die Schätzung eines Mittel- oder Totalwertes im Vordergrund steht und damit das Regressionsmodell, bzw. dessen Erklärungsgehalt, eher zweitrangig ist, können die synthetischen Variablen, die mittels Hauptkomponentenmethoden gewonnen werden, zumindest bei stetigen Hilfsvariablen problemlos als Grundlage für den GREG-Schätzer herangezogen werden.

3 Simulationsstudie

Gegenstand der Monte-Carlo-Simulationsstudie ist die Analyse einer geeigneten Wahl von Hilfsvariablen bei der Schätzung von Totalwerten mit Hilfe des verallgemeinerten Regressionsschätzers. Dabei wurden nach verschiedenen Stichprobendesigns jeweils 1.000

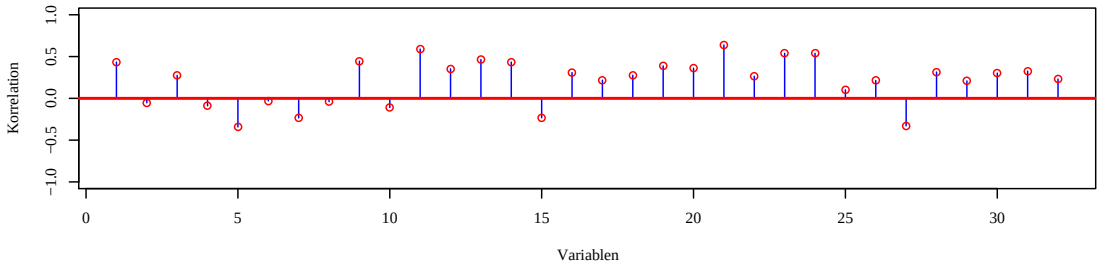

Abbildung 1: Korrelation der 32 Hilfsvariablen zur Untersuchungsvariablen in der Grundgesamtheit.

Stichproben vom Umfang $n = 500$ gezogen und anschließend die Monte-Carlo-Schätzverteilungen geeignet ausgewertet. Eine eingehende Darstellung dieser Vorgehensweise findet sich beispielsweise in Münnich and Rässler (2005) bzw. Hulliger and Münnich (2006) und der dort zitierten Literatur.

Die Grundgesamtheit vom Umfang $N = 42.733$ basiert auf der Stichprobe der deutschen Einkommens- und Verbrauchsstichprobe (EVS) von 2003. Der interessierende Populationsparameter ist der Totalwert des monatlichen Nettoeinkommens. Insgesamt wurden 32 Hilfsvariablen für die Untersuchung aus statistischen Gründen ausgewählt, um recht unterschiedliche Korrelationen zur Zielvariablen zu erhalten (siehe Abbildung 1). Überwiegend beinhaltet dieses Variablenset metrische Variablen, wie zum Beispiel monatliche Ausgaben für diverse Güter und Dienstleistungen oder monatliche Ersparnisse.

Für die Wahl der *optimalen* Hilfsvariablen, welche in der Simulationsstudie bei der gebundenen Hochrechnung berücksichtigt werden, erfolgt zunächst eine Selektion anhand der Grundgesamheit. Diese sollen zum Vergleich mit den aus den Stichproben gewonnenen Modellen dienen. Für das Akaike-, sowie auch für das Bayesianische Informationskriterium liefert die Vorwärtsselektion (FW) dieselben Ergebnisse wie die Rückwärtselimination (BW; siehe Tabelle 1). Zusätzlich wird noch ein Variablenset aus rein metrischen Variablen, sowie das *volle* Modell mit allen 32 zur Verfügung stehenden Variablen betrachtet, welche auch später bei der Hochrechnung in den GREG Schätzer eingehen.

Für das Verfahren der vollständigen Modellselektion wird die Anzahl der eingehenden Variablen auf 2 bis 8 beschränkt. Als Gütekriterium wird das adjustierte Bestimmtheitsmaß R^2_{adj} verwendet, welches im Gegensatz zum klassischen Bestimmtheitsmaß die

Tabelle 1: Werte der Maße AIC, BIC und R^2_{adj} für das volle Modell, das Modell mit nur metrischen Variablen und die nach Vorwärtsselektion und Rückwärtselimination bezüglich AIC und BIC in der Grundgesamtheit optimierten Modelle.

Modell	Volles Modell	FW/BW (AIC)	FW/BW (BIC)	Metrisch
# Variablen	32	28	23	23
AIC	695.824	695.817	695.878	699.063
BIC	696.517	696.476	696.363	699.263
R^2_{adj}	0.7287	0.7287	0.7282	0.7069

Anzahl der in das Modell eingehenden Variablen berücksichtigt. Bei der Verwendung der *zwei besten* Variablen resultiert ein R^2_{adj} von 0.55, welches bei Erweiterung um die *drittbeste* Variable auf 0.62 ansteigt. Ab der Hinzunahme der sechsten Variable (bei einem R^2_{adj} von 0.689) kann jedoch nur noch eine marginale Verbesserung des Gütekriteriums R^2_{adj} erreicht werden. Auch die Betrachtung des BIC-Kriteriums führt zu denselben Resultaten.

Neben der uneingeschränkten Zufallsstichprobe (SRS) wurden auch vier verschiedene geschichtete Zufallsstichprobenverfahren (StrRS) mit $H = 6$ Schichten verwendet. Zum einen wurde eine Stratifikationsvariable gewählt, die eine Korrelation von 0.9 zur Zielvariablen aufweist und welche nicht im Hilfsvariablenset enthalten ist. Zum anderen wurde die Stratifikation anhand der Variablen Alter (bis unter und über 18) und Geschlecht durchgeführt. Im Folgenden wurden diese Ziehungsmodelle mit StrRS 1 und StrRS 2 bezeichnet. In beiden Fällen wurde jeweils eine proportionale Allokation (prop) und eine optimale Allokation (opt) verwendet.

Um zunächst die Auswirkung des Stichprobeneffekts auf die Variablenselektion zu beurteilen, wird in jeder einzelnen Ziehung eine schrittweise Selektion durchgeführt. Abbildung 2 veranschaulicht die relativen Häufigkeiten, mit der jede der 32 Variablen in Abhängigkeit des Maßes AIC bzw. BIC, Vorwärtsselektion bzw. Rückwärtselemination sowie des Stichprobendesigns selektiert wurde. Zunächst erkennt man, dass sich die jeweiligen Häufigkeiten, mit denen die Variablen für die Regressionen ausgewählt wurden, über die Stichprobendesigns hinweg sehr ähnlich sind. Angesichts der Verwendung disproportionaler Auswahlsätze hätte man das nicht unbedingt erwartet.

Im Vergleich zur Variablenselektion in der Grundgesamtheit tendieren die Selektionsverfahren in der Stichprobe zu sparsameren Modellen. Während die Verwendung des AIC-Kriteriums bei sehr vielen Variablen zu weniger stabilen Ergebnissen führt (die relative Häufigkeit der gewählten Variablen liegt überwiegend zwischen 0.3 und 0.7), besitzen die Verfahren unter Verwendung des BIC-Kriteriums eine viel höhere Trennschärfe. In Analogie zum Ergebnis aus der Grundgesamtheit liefert die Rückwärtselimination anhand des BIC-Kriteriums dieselben Ergebnisse wie die Vorwärtsselektion mit BIC. Auch unter Verwendung des AIC-Kriteriums lassen sich nur geringe Unterschiede zwischen den zwei Verfahren erkennen.

Bei Verwendung der Hauptkomponentenregression interessiert zunächst die Stabilität der Eigenwerte in den einzelnen Stichprobenziehungen. Bei allen Stichprobendesigns resultieren in jedem Simulationsdurchlauf sieben Eigenwerte, die größer eins sind. In 80 Prozent der Ziehungen war auch der achtgrößte Eigenwert noch größer eins. Neun Eigenwerte größer eins resultieren hingegen nur noch in 10 Prozent der Simulationsdurchläufe. Weiterhin kann festgehalten werden, dass das Verhalten der Faktoren in Bezug auf die Eigenwerte über die Stichprobendesigns hinweg wiederum sehr stabil ist.

Um zum eigentlichen Ziel der Simulationsstudie zu kommen, werden für den verallgemeinerten Regressionsschätzer bei der Hochrechnung nun jeweils die Hilfsvariablensets verwendet, welche sich bei der Variablenselektion anhand der Grundgesamtheit als *optimal* erwiesen haben. Es bezeichnen GREG 01 das volle Modell, GREG 02 und 03 das AIC- bzw. BIC-optimale Modell sowie GREG 04 das Modell mit nur stetigen Variablen gemäß Tabelle 1. Mit GREG 05 bis GREG 11 werden die Hilfsvariablensets bezeichnet, welche auf Basis der vollständigen Modellselektion mit 2 bis 8 Variablen resultieren. Bei

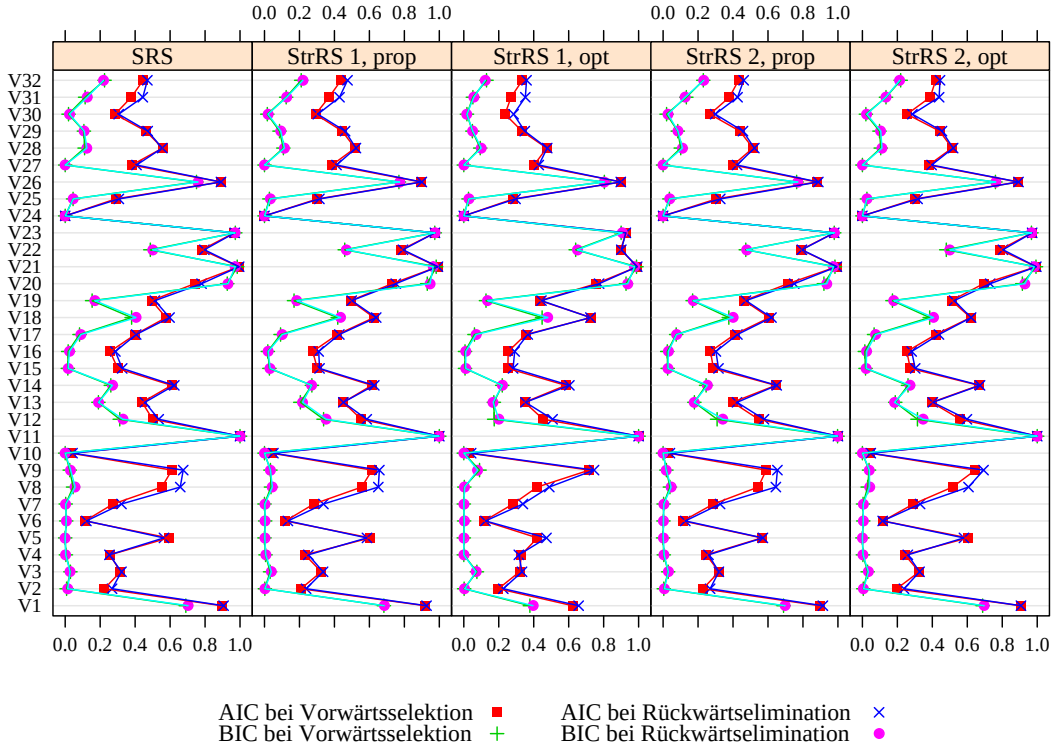

Abbildung 2: Relative Häufigkeiten der Selektion der 32 Variablen bzgl. der Maße AIC und BIC bei Vorwärts- und Rückwärtsselektion und verschiedene Stichprobendesigns.

der Hauptkomponentenregression, mit GREG PC2 bis PC10 bezeichnet, werden jeweils 2, 4, 6, 8 oder 10 Hauptkomponenten berücksichtigt. Bei GREG PCopt wurden in jeder Stichprobe nur Faktoren mit Eigenwert größer eins berücksichtigt. Der Abbildung 3 kann man in der letzten Spalte die Anzahl der Hilfsvariablen (bzw. Hauptkomponenten) aus den diversen Verfahren entnehmen.

Betrachtet man in Abbildung 3 zunächst die Verteilung der Gütekriterien AIC und BIC unter den verschiedenen Modellen, so tendiert erwartungsgemäß das AIC zu umfangreicheren Modellen. Die besten Ergebnisse erzielen demnach GREG 01 bis GREG 04, welche alle über 20 Variablen berücksichtigen. Die selben Resultate erhält man hier auch beim adjustierten Bestimmtheitsmaß. Beim BIC hingegen schneiden die Modelle aus der Hauptkomponentenregression sehr gut ab. Auch die 2 Modelle GREG 09 (mit 6 Variablen) und GREG 11 (mit 8 Variablen) scheinen nach dem BIC-Kriterium für die gegebene Datensituation angemessen zu sein.

In Abbildung 4 sind die Verteilungen der Totalwertschätzer der verschiedenen Regressionsschätzer und des Horvitz-Thompson-Schätzers veranschaulicht. Für eine übersichtlichere Darstellung wird lediglich der Bereich zwischen 0.4 Mrd. Euro und 0.5 Mrd. Euro angezeigt (der wahre Wert liegt bei 0.454 Mrd. Euro). Es sei angemerkt, dass bei den sehr umfangreichen Modellen in Einzelfällen sehr starke Ausreißer in den Schätzungen auf Grund von numerischen Instabilitäten möglich sind und daher die Mittelwerte sehr vom Median abweichen.

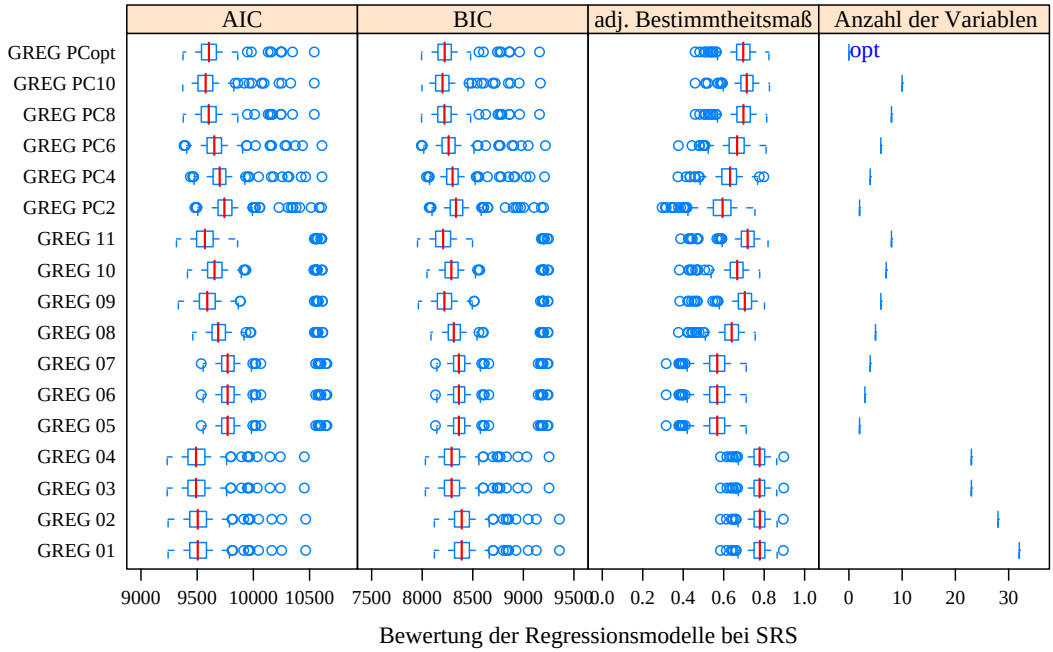

Abbildung 3: Verteilung der Informationskriterien und die Anzahl der verwendeten Variablen in Abhängigkeit der verschiedenen Regressionsmodelle.

Die vertikal durchgezogene Linie in der Grafik kennzeichnet den wahren Totalwert. Die für jedes Regressionsmodell separat eingezeichneten Linien hingegen stehen für die jeweiligen Mittelwerte der Schätzverteilungen. Die Mediane der Schätzverteilungen sind durch ein Kreuz gekennzeichnet.

Man erkennt, dass entgegen der Erwartungen und der zuvor aufgezeigten Ergebnisse die sehr umfangreichen Modelle nicht notwendigerweise am besten abschneiden. Vielmehr neigen diese in Einzelfällen zu sehr schlechten oder gar irrationalen Schätzungen. Der Median der Schätzergebnisse erweist sich über die Modelle hinweg eher konstant. Die Instabilität der umfangreichen Modelle macht sich auch in der Varianz der Schätzungen, die hier nicht explizit dargestellt ist, bemerkbar. Weiterhin ist festzuhalten, dass die Modelle der Hauptkomponentenregression sowohl für die Punkt- als auch für die Varianzschätzung sehr zufriedenstellende Ergebnisse liefern.

Neben der eigentlichen Modellierung und Schätzung kann ebenso das Argument einer stabilen Varianzschätzung eine Rolle spielen. Die enormen Ausreißer bei den Schätzverteilungen wird man sicher in der Praxis erkennen und daraufhin ein besser geeignetes Regressionsmodell auswählen. Allerdings fällt auch auf, dass der Effekt des Stichprobendesigns geringer als erwartet ausfällt. Während beim frei hochgerechneten HT-Schätzer über das optimierte Stratifikationsdesign eine erhebliche Varianzreduktion erreicht werden kann, verhalten sich die Ergebnisse der GREG-Schätzer über das Stichprobendesign hinweg fast identisch. Lediglich bei den optimalen Allokationen können geringe Unterschiede erkannt werden.

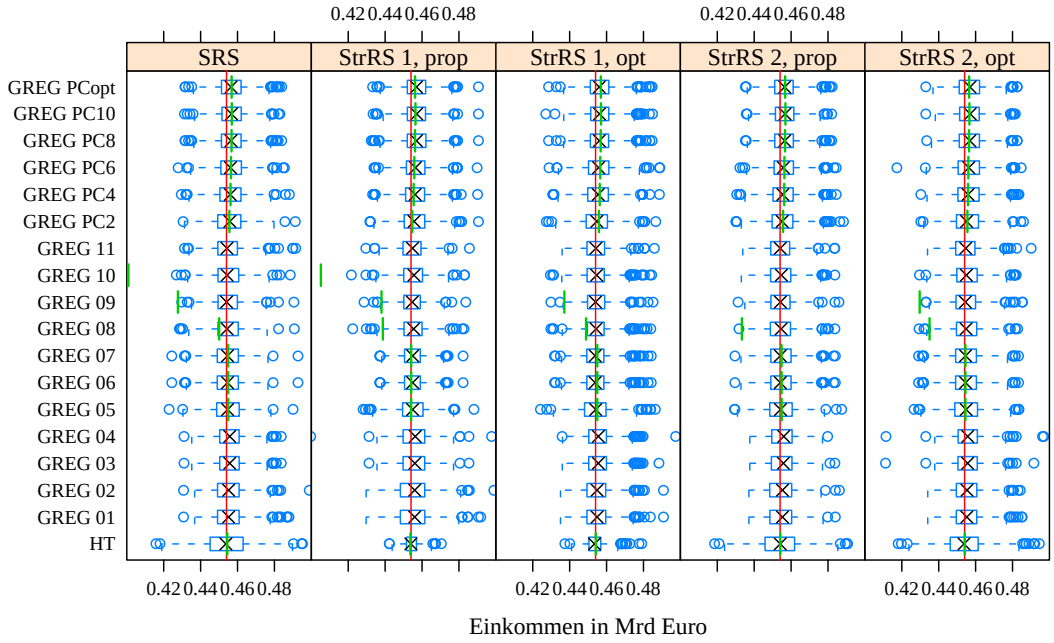

Abbildung 4: Ergebnisse der Punktschätzung.

4 Zusammenfassung und Ausblick

Sofern durch theoretische oder praktisch-inhaltliche Überlegungen keine konkrete Auswahl an Hilfsvariablen getroffen werden kann, bieten die Verfahren der Variablenselektion eine hilfreiche Unterstützung bei der Suche nach einem geeigneten Modell. Diese Verfahren helfen jedoch nur bei der Auswahl von Hilfsvariablen für eine geeignete Schätzung eines Regressionsmodells (im Sinne eines Optimalitätskriteriums und bei Vermeidung von sachlich ungeeigneten Korrelationen), jedoch weniger bei der eigentlich interessierenden Stichprobenschätzung.

Insgesamt lässt sich feststellen, dass Variablenselektionsverfahren und Informationskriterien nicht die gewünschte Unterstützung bei den Stichprobenschätzungen liefern. Das liegt zum einen daran, dass bei den vorliegenden Stichprobenschätzungen zwar Regressionsmodelle unterstützend verwendet werden, die Effizienz der Modelle selber aber keine wesentliche Rolle spielt. Zum anderen muss aber auch der Trade-off zwischen Effizienz der Modelle, die im Allgemeinen bei größeren Modellen eher erreicht wird, und Empfindlichkeit der Modellschätzungen, die bei ungeeigneter Spezifikation in zumeist zu großen Modellen und bei kleinen Stichprobenumfängen entsteht, betrachtet werden. Hinzu kommt, dass eine Beurteilung anhand der Varianzschätzungen möglicherweise noch zu differenzierteren Beurteilungen führt.

Erstaunlicherweise spielten die verwendeten Designs nur eine untergeordnete Rolle bei der Beurteilung der Verfahren. Lediglich bei den Schätzungen, insbesondere bei der Varianzschätzung, konnten kleinere Abweichungen gefunden werden. Größere Abweichungen im Sinne eines *sample selection bias* (vgl. Chambers and Skinner, 2005, Kapitel

1.1) kann man wohl erst bei komplexeren Designs, wie etwa Klumpenstichproben, erwarten. Eine Erhöhung des Stichprobenumfangs auf $n = 2.000$ zeigt kaum andere Beurteilungen der Verfahren. Lediglich bei den Variablenselektionsverfahren in der Stichprobe anhand des AIC-Kriteriums wurde eine höhere Trennschärfe erreicht.

Besonders erfreuliche Resultate konnten bei der Verwendung der Hauptkomponentenregression beobachtet werden. Hier muss allerdings einschränkend erwähnt werden, dass nur metrisch skalierte Variablen verwendet werden können. In der Praxis, sofern man nicht Haushalts- und Budgeterhebungen hat, wird man jedoch eher nicht metrisch skalierte Daten vorfinden. Weitere Studien werden sich daher sicher noch eingehender mit der Skalierungsproblematik befassen.

Danksagung

Die Autoren danken Herrn Akad. Direktor Dr. Rolf Wiegert sowie zwei anonymen Referees für zahlreiche wertvolle Hinweise, die zu einer Verbesserung dieses Artikels beigetragen haben.

Literatur

Akaike, H. (1974). A new look at statistical-model identification. *IEEE Transactions on Automatic Control*, *19*, 716-723.

Chambers, R. L., and Skinner, C. J. (2005). *Analysis of Survey Data*. Chichester: John Wiley & Sons.

Deville, J. C., and Särndal, C. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, *87*, 376-382.

Draper, N. R., and Smith, H. (1981). *Applied Regression Analysis* (2nd ed.). Chichester: John Wiley & Sons.

Fahrmeir, L., Kneib, T., and Lang, S. (2007). *Regression: Modelle, Methoden und Anwendungen*. Berlin: Springer.

Harrell, F. E. (2001). *Regression Modeling Strategies*. New York: Springer.

Hulliger, B., and Münnich, R. (2006). Variance estimation for complex surveys in the presence of outliers. In *Proceedings of the Section on Survey Research Methods* (p. 3153-3161). American Statistical Association.

Körner, T., Nimmergut, A., and Nökel, J. (2006). Die Dauerstichprobe befragungsbereiter Haushalte. *Wirtschaft und Statistik*, *5*, 451-468.

Lohr, S. (1999). *Sampling: Design and Analysis*. CA: Pacific Grove: Duxbury Press.

Mallows, C. (1973). Some comments on C_p. *Technometrics*, *15*, 661-676.

Münnich, R. (2008). Varianschätzung in komplexen Erhebungen. *Austrian Journal of Statistics*, *37*. (in press)

Münnich, R., Huergo, L., Magg, K., and Ohly, D. (2005). *Konzeption und Test von Varianzschätzung für Erhebungen auf Basis der Dauerstichprobe befragungsbereiter Haushalte.* Universität Tübingen. (Abschlussbericht des Forschungsprojekts B 1.22–5118/04/StBA)

Münnich, R., and Rässler, S. (2005). PRIMA: A new multiple imputation procedure for binary variables. *Journal of Official Statistics*, *21*, 325-341.

Reineking, B., and Schröder, B. (2004). Variablenselektion. UFZ-Bericht: Habitatmodelle - Methodik, Anwendung, Nutzen. In *Tagungsband zum Workshop 8.-10. Oktober 2003, UFZ Leipzig.*

Schwarz, G. (1978). Estimating dimension of a model. *Annals of Statistics*, *6*, 461-464.

Särndal, C. E., Swensson, B., and Wretman, J. (1992). *Model Assisted Survey Sampling*. New York: Springer.

Weisberg, S. (2005). *Applied Linear Regression* (3rd ed.). Chichester: Wiley.

Adressen der Autoren:

Melanie Knobelspies
Interbrand Zintzmeyer & Lux AG
Kirchenweg 5
CH-8008 Zürich

melanie.knobelspies@interbrand.ch
http://www.interbrand.ch

Ralf Münnich
Universität Trier, FB IV, VWL
Wirtschafts- und Sozialstatistik
Universitätsring 15
D-54286 Trier

muennich@uni-trier.de
http://www.statistik.uni-trier.de