Thema:

Bootstrap-Methoden für die Regressionsanalyse

Bachelorarbeit

im Fachgebiet Wirtschaftsinformatik
am Lehrstuhl für Quantitative Methoden

Themensteller: Prof. Dr. Ulrich Müller-Funk

vorgelegt von: Michael Höing
Bismarckallee 49, C101
48151 Münster
0251/83851770

Abgabetermin: 2004-04-26

Inhaltsverzeichnis

1 Einleitung

Statistischen Fragestellungen liegt es in der Natur, dass ihre Beantwortung oft aufwendige theoretische Ansätze erfordert, die umständlich oder aufgrund ihrer Komplexität in hohem Maße fehlerbehaftet sind. *Simulation* ist ein Werkzeug, das Wege vorbei an komplexen Berechnungen eröffnet und oft Resultate ermöglicht, die weniger Variabilität aufweisen. War lange Zeit die Durchführung von Simulation aufgrund nicht vorhandener Rechenkapazität schlecht realisierbar oder unmöglich, so stellt sie vor dem Hintergrund moderner Rechnergenerationen einen generischen Ansatz dar, der schnell zu Ergebnissen führt, zur Validierung anderer Methoden herangezogen werden kann oder vielfach Möglichkeiten eröffnet, wenn andere Verfahren wegen ihrer zu strengen Voraussetzungen scheitern.

Die grundlegende Idee der Simulationsmethoden, die in dieser Arbeit vorgestellt werden, ist das Generieren neuer Datensätze auf Basis *eines* gegebenen Ausgangdatensatzes. der als Stichprobe einer unbekannten Population angesehen wird. Dieses *resampling* geschieht entweder unter Zugrundelegung eines gefitteten Modells[1] oder direkt von der empirischen Verteilungsfunktion der Daten. Auf die künstlich erweiterte Datenbasis wird dann der so genannte *bootstrap* angewendet, der auf folgendem Prinzip beruht:

> *"The population is to the sample as the sample is to the bootstrap samples."*[2]

Die Tatsache, dass der Zusammenhang von Ausgangsdatensatz und simulierten Daten „berechnet" werden kann, ermöglicht nach dem bootstrap-Prinzip also Aussagen über die unbekannte Population bzw. die „wahren Eigenschaften" des Ausgangsdatensatzes.

Der Begriff *bootstrap* bzw. *bootstrapping* für dieses Verfahren leitet sich von einer Sage über den Baron von Münchhausen ab, nach der es diesem gelang, einmal in einen Sumpf eingesunken, sich selber an den Stiefelschlaufen (engl.: bootstraps) herauszuziehen. Die Analogie ist hier, dass die Datenbasis für den bootstrap „aus sich selbst heraus" wächst.[3]

Der bootstrap sei an dieser Stelle als ein generisches Verfahren verstanden, dass Simulation, genauer resampling, verwendet und sich für die verschiedenen Domänen der statistischen Analyse und für verschiedenste Fragestellungen operationalisieren

---

1 „fit" bedeutet „passend". Als „gefittete Werte" werden die vom Modell unterstellten responses bezeichnet. Sie enthalten keinen Fehlerterm.

2 Vgl. Fox (2002), S. 2.

3 Vgl. Davison, Hinkley (1997), S. 2 f.

lässt. In dieser Arbeit ist die Anwendung des bootstrappings auf die *Regressionsanalyse* thematisiert. Verschiedene Fragestellungen werden zunächst unter Zugrundelegung des *klassischen linearen Modells* für die lineare Regression behandelt. Im Anschluss werden Ansätze für *generalisierte lineare Modelle* vorgestellt, um einen ersten Ausblick auf die Möglichkeiten der Erweiterung der bootstrap-Technik auf mächtigere praxisrelevantere Modellklassen zu gewähren.

Nachdem auf die Annahmen des klassischen linearen Modells und die lineare Regression eingegangen wurde, werden zunächst das modellbasierte resampling-Schema *resampling errors* und das nicht modellbasierte Schema *resampling cases* erläutert, die zur Erzeugung neuer Datensätze, den bootstrap samples, eingesetzt werden. Sie sind bei der linearen Regression die Basis für jeden bootstrap. Zur Prüfung der Relevanz einzelner Kovariate werden anschließend Permutationstests und verschiedene *bootstrap Tests* als resampling-Varianten des Permutationstests erläutert. Im Anschluss an diese Verfahren für den deskriptiven Part der Regressionsanalyse werden im Folgenden bootstrap-Verfahren erläutert, die es ermöglichen, auf Basis eines Modellfits *Prognoseintervalle* für prognostizierte Werte anzugeben. Werden hier lediglich Punktschätzungen um ein Intervall erweitert, so dient die nachfolgende Betrachtung *aggregierter Fehlermaße* und deren bootstrap-Varianten zur Bewertung der Güte des Modellfits und dessen Eignung zur Prognose. Ein letzter Aspekt der Betrachtung der linearen Regression sind Methoden zur *Identifikation von „Ausreißern“*. Die Verallgemeinerung der resampling-Schemata für generalisierte lineare Modelle soll im Folgenden einen Eindruck vermitteln, wie die vorgestellten bootstrap-Verfahren für komplexere Modelle modifiziert werden können. Eingegangen wird hier insbesondere auf das Problem, dass die erzeugten bootstrap samples bei generalisierten linearen Modellen nicht zwangsläufig notwendige Eigenschaften des Ausgangsdatensatzes aufweisen, wie z.B. Nichtnegativität von Zähldaten.

Um die vorgestellten Verfahren zu verdeutlichen und diskutierte Aspekte zu untermauern, sind die theoretischen Kapitel dieser Arbeit um neun Beispiele ergänzt, in denen bootstrapping auf ausgewählte oder synthetische Datensätze angewendet wird. Für die Implementierung der bootstraps und die Erzeugung der Abbildungen dieser Arbeit, die die Analyseergebnisse jeweils illustrieren, wurde die Sprache *S* verwendet. Die Beispiele dienen als „link“ zwischen Theorie und Praxis und veranschaulichen eine mögliche Umsetzung der „computer-intensive bootstrap methods“.[4]

[4] Vgl. Anhang B.

Literaturverzeichnis

Davison, A. C.; Hinkley, D. V.: Bootstrap Methods and their Application. Cambridge et al. 1997.

Fahrmeir, L.; Hamerle, A.; Tutz, G.: Multivariate statistische Verfahren. 2. Auflage. Berlin, New York 1996.

Hampel, F. R. et al.: Robust statistics: An approach based on influence functions. New York et al. 1986.

Internet-Adressen

Fox, J.; Bootstrapping Regression Models. 2002. http://cran.r-project.org/doc/contrib/Fox-Companion/appendix-bootstrapping.pdf. Abrufdatum: 2004-04-25.

Olive, D. J.; ROBUST ESTIMATORS FOR TRANSFORMED LOCATION-SCALE FAMILIES. 2004. http://www.math.siu.edu/olive/pprloc2.pdf. Abrufdatum: 2004-04-25.

Schmidt, C.; Verallgemeinerte Lineare Modelle. 2002. http://www.uni-ulm.de/~cschmid/oldstat/se4_2.htm. Abrufdatum: 2004-04-25.