

name=Datamining , description=Systematische Anwendung statistischer Methoden auf einen grossen Datenbestand mit dem Ziel, neue Muster zu erkennen

Automatische Verfahren zur Prädiktorauswahl in Regressionsmodellen

Literaturarbeit vorgelegt von
Markus Graf (markus.graf@uzh.ch)
am Psychologisches Institut der Universität Zürich
Betreut durch Dr. Christina Werner

Zusammenfassung

Eigener Abstract: -

Themenvorgabe: In vielen psychologischen Bereichen geht es darum, Kriteriumsvariablen durch Prädiktorvariablen möglichst gut vorherzusagen. Wenn viele potentielle Prädiktorvariablen in Frage kommen und es keine theoretischen Gründe gibt, die nur ganz bestimmte Prädiktorvariablen nahelegen, werden in Anwendungssituationen oft automatische Verfahren der Auswahl von Prädiktorvariablen verwendet, um mit möglichst wenigen Prädiktoren eine möglichst gute Vorhersage des Kriteriums zu erreichen, beispielsweise die sog. “Stepwise”-Methode in multiplen Regressionsmodellen. Die Literaturarbeit soll einen Überblick über verschiedene existierende Möglichkeiten zur Selektion von Prädiktorvariablen in Regressionsmodellen geben, und deren Eignung für psychologische Anwendungen kritisch diskutieren.

Inhaltsverzeichnis

Einführung	3
Automatische Modellwahlverfahren	4
Exhaustiv Regression	4
Schrittweise Regression	5
Multikollinearität	8
Überanpassung - Der einfachste Ausdruck verwickelter Probleme ist immer der beste	9
Kriterienbasierte Strategien	10
Kreuzvalidierung	11
Software zu den vorgestellten Verfahren	12
Modellselektion	13
Kreuzvalidierung	13
Zusammenfassung / Diskussion	13
Literatur	16
Anhang	17

Einführung

Das Standardverfahren um den quantitative Zusammenhang zwischen einer abhängigen und einer unabhängigen Variablen zu beschreiben stellt die Regressionsanalyse dar. Begründet wurde dieses Verfahren durch Carl Friedrich Gauss in seiner Schrift, in der er, mit Hilfe der Methode der kleinsten Quadrate, die Bewegung der Himmelskörper um die Sonne im Kegelschnitt beschrieb (?, ?).

Im Unterschied zur einfachen linearen Regression, werden in einem multiplen Regressionsmodell mehrere unabhängige Variablen mit einbezogen. Es resultiert eine Regressionsgleichung welche zur Vorhersage einer Kriteriumsvariable aufgrund mehrerer Prädiktorvariablen genutzt wird (Bortz & Schuster, 2011, S. 448). Als Gretchenfrage stellt sich nun welche Prädiktoren nun ein Modell am besten erklären. Zu Beginn der Psychologischen Forschung mussten Modelle von Hand berechnet werden. Zwangsläufig wurden wenige Prädiktoren erhoben und einfache Modelle gerechnet. Friedman analysierte beispielsweise 1944 die Lebensdauer von Turbinenschaufeln in Abhängigkeit von Stress, Temperatur und einigen Legierungsparametern. Zwar wurde die Berechnung nicht mehr von Hand durchgeführt, doch benötigte eine Regressionsschätzung inklusive Berechnung der Teststatistiken rund 40 Stunden (?, ?, p.2). Jeder durchschnittliche Computer erledigt dies heutzutage Sekundenbruchteile.

Mit dem technischen Fortschritt einhergehend wurden Verfahren entwickelt, welche möglichen Kombinationen von Prädiktoren berücksichtigen und gegeneinander testen. In einem ersten Teil dieser Arbeit werden die wichtigsten Verfahren dargestellt.

Es können beliebig viele potentiell erklärende Variablen erhoben werden um sich komplexe Modelle generieren zu lassen. Menschen tendieren zu glauben, dass komplexe Probleme komplexe Lösungen benötigen. Die Forschung zeigt jedoch, dass gerade das Umgekehrte der Fall ist (?, ?, p.3). Insbesondere Gigerenzer demonstrierte eindrucksvoll wie mit einfachen Rekonitionsheuristiken bessere Vorhersagen gemacht werden konnten als mit komplexen statistischen Modellen (?, ?). Komplexe Modelle können sehr gute Vorhersagen liefern innerhalb des Trainings-

datensatz, doch oft scheitert die Vorhersage beim Versuch, diese zu generalisieren. Der zweite Teil befasst sich mit dem Problem der Überanpassung komplexer Modelle und diskutiert mögliche Lösungsansätze.

Automatische Modellwahlverfahren

In der psychologischen Forschung kommen oft Ex-post-facto-Designs zum Einsatz. Dies insbesondere weil viele Daten mit geringem Aufwand erhoben werden können. Oft werden Daten gleich für mehrere Studien erhoben und in anderen Studien verwendet. Daraus resultieren viele potenzielle Prädiktorvariablen für neue Fragestellungen. Es gilt das “beste” Modell berechnen, oft ohne dass es theoretische Gründe gibt nach denen die Auswahl der Prädiktoren einzuschränken ist. “Alle Modelle sind falsch, aber einige sind nützlich” (?, ?, p.202). Box will damit hervorheben, dass obschon in der Literatur oft vom “besten” oder “wahren” Modell gesprochen wird, dies nur ein Approximation darstellt (?, ?, p.172).

Exhaustiv Regression

Eine naive Herangehensweise ist alle möglichen Modelle welche mit p Prädiktoren möglich sind durchzurechnen. Zur Beurteilung der Modellgüte kann die mittlere quadratische Abweichung herangezogen werden, welche durch das S_p -Kriterium ausgedrückt wird. Das Modell mit der kleinsten Fehlerquadratsumme SSE_p wird als das “beste” Modell bezeichnet (?, ?, p. 6).

$$S_p = \frac{SSE_p}{(n - p - 2)(n - p)} \quad (S_p\text{-Kriterium})$$

Da alle Möglichen Kombinationen durchgerechnet werden, wird das “beste” Modell, also das welches den Trainingsdatensatz am besten Vorhersagen kann, gefunden. Entsprechend wird in der Literatur auch oft vom optimalen Modell gesprochen. Es liegt jedoch auf der Hand, dass dies zwar die interne Validität erhöht, aber die externe Validität gefährdet. Das optimale Modell kann, insbesondere bei hoher Komplexität, zu fest auf die Referenzdaten behaftet sein und dadurch

schlecht generalisierbar sein. Auf diesen Effekt der Überanpassung wird im zweiten Teil dieser Arbeit noch näher eingegangen. ? (? , p.6) sieht einzig den Nachteil darin, dass der Rechenaufwand exponentiell mit der Anzahl zu berücksichtigender Prädikatoren steigt. Es müssen immer $2^p - 1$ Modelle berechnet werden, bei 5 Prädikatoren sind dies 31 Modelle, bei 10 bereits 1023 usw. Während früher eingeschränkte Rechenkapazität oft ein ökonomischer Faktor war - es musste Rechenzeit in einem Rechenzentrum reserviert werden, spielt die Rechengeschwindigkeit auf modernen Systemen eine untergeordnete Rolle, da insbesondere in der Psychologie oft nur eine Handvoll Prädikatoren durchgerechnet werden müssen.

Schrittweise Regression

Das optimale Modell beinhaltet jeden Prädiktor, der die Voraussage bezüglich des getesteten Datensatzes auch nur minimal verbessert. Es stellt sich die Frage ob diese minimale Verbesserung noch nützlich ist. Die "stepwise"-Verfahren arbeiten wesentlich liberaler, in dem Prädikatoren hinzugefügt oder eliminiert werden, je nach deren Relevanz für die Modellgüte. Es werden Kriterien festgelegt, nach welchen ein Modell als angemessen zu betrachten ist. Dies hat gegenüber der exhaustiven Verfahren den Vorteil, dass nicht alle Modelle berechnet werden müssen und entsprechend schneller Lösungen gefunden werden. Im Schnitt müssen xxxx Modelle berechnet werden, um eine adäquate Lösung zu finden (? , ?).

Innerhalb der schrittweisen Verfahren unterscheidet man zwischen *forward selection* und *backward elimination*. Ausgehend vom leeren Modell werden in der ersten Variante schrittweise weitere Variable der Nützlichkeit nach in das Modell integriert, bis eine Abbruchbedingung erfüllt ist.

In der zweiten Variante werden alle Prädikatoren in das Modell integriert und schlechte entfernt, wiederum bis das Kriterium erreicht ist.

Die Aufnahme einer neuen Variable kann dazu führen, dass eine bereits im Modell vorhandene Variable obsolet wird. Um diesem Umstand Rechnung zu tragen werden oft forward selection und backward elimination kombiniert (Bortz & Schuster, 2011, p. 461).

In seltenen Fällen kann es vorkommen, dass zwei Variablen für sich das Kriterium in die Regressionsgleichung aufgenommen zu werden nicht erfüllen, jedoch zusammen zum Vorhersagepotential einen substantiellen Beitrag leisten (?, ?, p.261). Schrittweise Verfahren sind entsprechend nicht in der Lage solche Effekte mit zu berücksichtigen.

Entgegen dem exhaustiven Verfahren besteht bei schrittweisen Verfahren das Problem, dass unter Umständen nicht das optimale Modell gefunden wird. Da die Nützlichkeitsunterschiede, welche oft nur geringe statistische Bedeutung haben, das Modell bestimmen, bezeichnet Bortz und Schuster (2011, p. 462) dieses Verfahren eher zu den explorativen gehörend. ? (? , p. 56ff) lehnt das Verfahren gar ab und führt ins Feld, dass sämtliche statistischen Prinzipien verletzt würden. ? (?) zeigte jedoch in einem Vergleich, dass die durchschnittliche Differenz der Fehlerquadratsummen zwischen exhaustiven und schrittweiser Regression kaum 7% übertrifft.

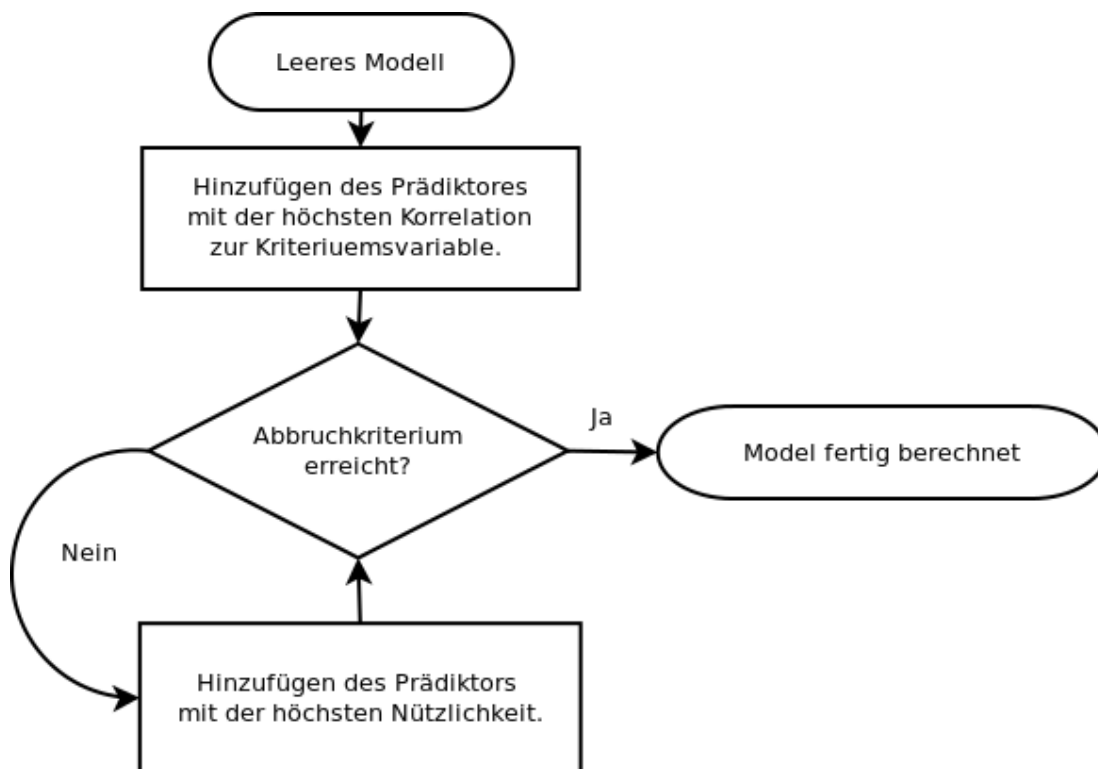


Abbildung 1. : Forward Selection

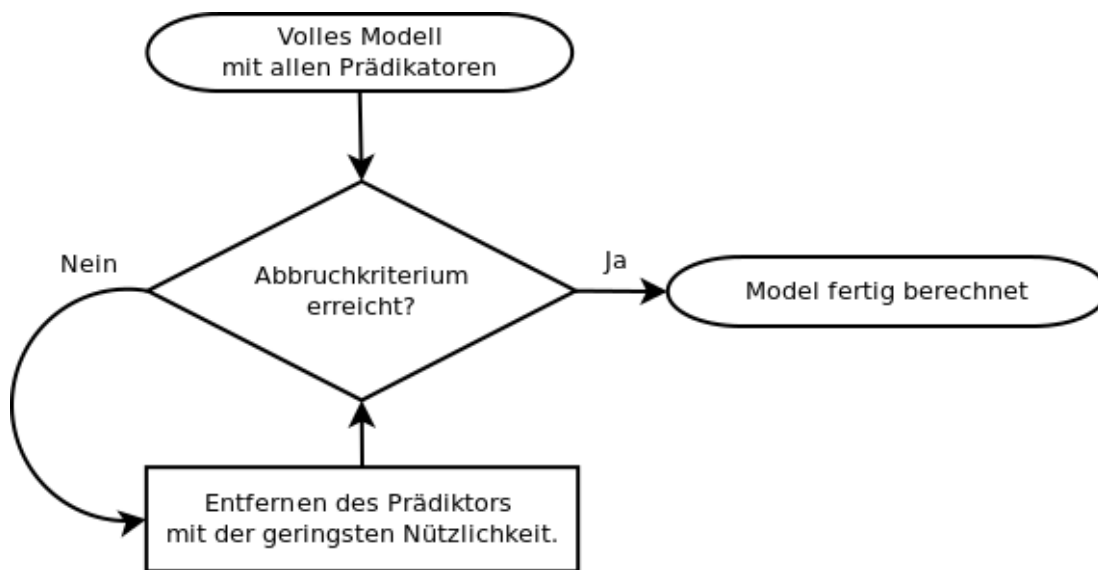


Abbildung 2. : Backward Elimination

Zentrales Element der schrittweisen Regression ist das Kriterium zur Beurteilung der Modellanpassung, welche besagt, weshalb und wann ein Modell als akzeptabel zu betrachten ist. Als Folge dessen wird damit auch die Anzahl relevanter Prädiktoren bestimmt.

Das Quadrat des multiplen Korrelationskoeffizienten R^2 besagt wie viel systematische Varianz aufgeklärt wird. Das Bestimmtheitsmass steigt mit der Anzahl der Prädiktoren p , insbesondere bei kleinem Stichprobenumfang n , weshalb eine Schrumpfungskorrektur vorgenommen werden muss (Bortz & Schuster, 2011, p. 451).

$$\bar{R}^2 = 1 - \frac{\frac{1}{n-p-1} \sum_{i=1}^n (Y_i - \hat{Y}_i)^2}{\frac{1}{n-1} \sum_{i=1}^n (Y_i - \bar{Y})^2} \quad (\text{korrigiertes Bestimmtheitsmass})$$

Insbesondere bei kleinem R^2 sollte auf Signifikanz getestet werden, entsprechend werden Modelle in schrittweisen Verfahren nicht einzig aufgrund von R^2 selektiert.

Das Verfahren wird beendet, wenn kein Prädiktor mehr vorhanden ist, der das Vorhersagepotential signifikant erhöht (?, ?, p.48). Das vergleichen zweier Regressionsgleichungen mittels Signifikanztest bedingt, dass diese geschachtelt sein müssen, das kleinere Modell muss im grösseren enthalten sein (?, ?, p. 508).

Das gewählte Signifikanzniveau ist eigentlich unbegründet gewählt (?, ?, p. 174). ? (?, p. 269) diskutieren mehrere Empfehlungen für Signifikanzniveaus und weisen darauf hin, dass sich über mehrere Tests der α -Fehler kumuliert. In Simulationen mit artifiziellen Daten zeigen ? (?) das Problem multipler Tests beispielhaft auf. Daraus resultierend lehnen sie die Verwendung der schrittweisen Regression gar ab.

Ein weitere Schwäche des Hypothesentestens ist der Einfluss der Stichprobengrösse. So kann bei genügend grossem Umfang nahezu jede Hypothese abgelehnt werden, was wiederum zu komplexen und überangepassten Modellen führt (?, ?, p.173).

Multikollinearität

Multikollinearität, als hohe Korrelationen zwischen mehreren Prädiktoren, führt zu Problemen bei der automatischen Modellwahl. In schrittweisen Verfahren ist es in solchen Fällen häufig vom Zufall abhängig welche der beteiligten Variablen als erste weggelassen beziehungsweise aufgenommen wird. Ein Anstieg der korrelation zwischen Prädiktoren hat zur Folge, dass (a) die p-Werte bei Signifikanztests sinken, (b) schwache Prädiktoren entgegen den starken eher fälschlicherweise ausgeschlossen werden, (c) korrekt hoch signifikante Prädiktoren werden eher ausgeschlossen wenn die korrelation zwischen einem konfundierenden Prädiktor und dem Kriterium steigt und (d) selbst wenn die korrelation zwischen konfundierendem Prädiktor und Kriterium klein ist, besteht die Gefahr, dass korrekt schwach signifikante Prädiktoren nicht signifikant werden (?, ?, p. 2810).

Grundsätzlich gilt die Unabhängigkeit der Prädiktoren als Voraussetzung für die multiple Regression, doch gerade in der psychologischen Forschung lassen sich Korrelationen meist schlecht vermeiden. Um bereits im Vorfeld Hinweise auf

potentielle Kollinearität zu erhalten, empfiehlt es sich die Kovarianzmatrix zwischen allen Prädiktoren vor der eigentlichen Auswahl zu betrachten. Eventuell wurde das selbe Merkmal mehrmal erhoben. Der Zusammenhang zwischen einer Prädiktorvariable i und der vorhergesagten Kriteriumsvariable lässt sich mit dem Strukturkoeffizienten c_i ausdrücken.

$$c_i = \frac{r_{ic}}{R} \quad (\text{Strukturkoeffizient})$$

r_{ic} beschreibt den einzelnen Korrelationskoeffizient zwischen dem Prädiktor i und dem Kriterium c und R den multiplen Korrelationskoeffizient. So kann auf die Prädiktoren eingeschränkt werden, welche am besten das Kriterium vorhersagen (Bortz & Schuster, 2011, S. 453).

Überanpassung - Der einfachste Ausdruck verwickelter Probleme ist immer der beste

Der Mensch strebt nach Kontrolle und verfällt dabei oft der Illusion, alles bis ins letzte Detail kontrollieren zu müssen. Entsprechend wurden im Zuge der technischen Revolution hoch komplexe Systeme und Maschinen entwickelt, welche jeder Anwendung gerecht werden sollten. Insbesondere im Bereich des Produktdesigns hat in den letzten Jahren ein radikales umdenken statt gefunden: “Simplicity is about subtracting the obvious, and adding the meaningful.” (?, ?). Einfachheit wurde gelebt was zahlreiche unumstritten grossartige Produkte hervor brachte.

Die Illusion der Komplexität ist auch in der Statistik anzutreffen (?, ?, p. 3). Im Kontext der Modellwahl äussert sich dies in Form der Überanpassung, wenn also das Modell zu sehr an die Testdaten angepasst ist. Insbesondere Modelwahlverfahren, welche die Anzahl der Prädiktoren nicht abstrafen, sind davon betroffen. Als Einflussfaktoren seitens der Daten sind Repräsentativität und Stichprobengrösse zu nennen. Mit steigendem Stichprobenumfang und höheren Repräsentativität sinkt die Überanpassung und steigt die stabilität der Vorhersage.

Kriterienbasierte Strategien

Die bisher beschriebenen Modellwahlverfahren fokussieren sich darauf anhand der gegebenen Daten das “beste” Modell zu finden. Kriterienbasierende Strategien zur Vermeidung der Überanpassung strafen Komplexität ab, in dem das Kriterium diese berücksichtigt. Je komplexer ein Regressionsmodell wird, desto besser muss die Vorhersage stimmen um die Komplexität zu rechtfertigen.

Colin Lingwood Mallows entwickelte das C_p Kriterium, dass auf der Methode der kleinsten Quadrate (S_p -Kriterium) aufbaut und sowohl die Prädiktoranzahl p als auch die Stichprobengrösse n berücksichtigt.

$$C_p = \frac{SSE_p}{\sigma^2} - n + 2p \quad (\text{Mallows's } C_p)$$

Angestrebt wurde dabei alle wichtigen Prädiktoren zu berücksichtigen. Das “beste” Modell ist das mit (a) dem niedrigsten C_p -Wert, der (b) möglichst gleich p ist, ist (? , ?).

Die bisher auf der Methode der kleinsten Quadrate basierenden Verfahren bedingen, dass die Prädiktoren geschachtelt sind. Das soll heissen, dass jeweils alle Prädiktoren des kleineren Modells im grösseren enthalten sein müssen. Dies wird für die beiden Kennwerte, Akaike's Informationskriterium (AIC) und Bayessche Informationskriterium (BIC) nicht vorausgesetzt. Beide Kennwerte basieren auf der Maximum-Likelihood-Methode L und berücksichtigen die Anzahl Prädiktoren p . Dem Prinzip der Sparsamkeit entsprechend geben kleinere Modelle bei gleicher Vorhersagekraft, bessere Kennwerte (? , ?, p. 509). Bei Regressionsmodellen mit normalverteilten Fehlern entspricht die Wahrscheinlichkeitsfunktion L der des quadrierten Standardfehlers der Regression σ^2 (? , ?, p. 169).

$$AIC = n \log(\sigma^2) + 2p \quad (\text{AIC})$$

Gegenüber AIC ist BIC konservativer, denn es wird zusätzlich der Stichprobenumfang n berücksichtigt (? , ?, p. 169).

$$BIC = n \log(\sigma^2) + p \log(n) \quad (\text{BIC})$$

Kreuzvalidierung

Die Stabilität eines Modelles lässt sich durch den Vergleich mit zusätzlichen Stichproben ermitteln. Zu diesem Zweck kommen sogenannte Kreuzvalidierungsverfahren zum Einsatz.

Die Idee hinter der Kreuzvalidierung liegt darin, die Daten aufzuteilen. Zum einen in eine Referenzstichprobe, anhand der die Gleichung geschätzt wird, zum anderen in eine oder mehrere zusätzliche Teststichproben, anhand derer die Stabilität als durchschnittliche Fehlerquote berechnet wird (Lütkepohl, 2000, p. 3).

Es fragt sich, welchen Platz die Kreuzvalidierung in der Modellselektion einnimmt. Kreuzvalidierungen können über ein Set von n Regressionsgleichungen durchgeführt werden, beispielsweise potentielle Modelle nach einer schrittweisen Regression (Lütkepohl, 2000, p. 12). n potentiellen Modelle werden validiert, wobei unter Umständen das stabilste nicht gleich dem vielversprechendsten aus der vorangegangenen Selektion ist. Überangepasste Modelle können somit eliminiert werden und an deren Platz rücken einfachere und stabilere Modelle. Der Vorteil dieses Vorgehens liegt darin, dass (a) die Validierung komplett von der Modellselektion getrennt werden kann und (b) es nur n Durchgänge benötigt. In der Modellselektion wird jedoch die Stabilität nicht berücksichtigt. Bei der schrittweisen Regression haben wir gesehen, dass das "beste" Modell nicht zwangsläufig gefunden wird. Entsprechendes gilt für die Stabilität, was zur Konsequenz führen kann, dass zwar gute Modelle gefunden werden, diese jedoch allesamt nicht stabil genug sind oder das stabilste schlicht nicht gefunden wird. Lütkepohl (2000, p. 12) nennen noch die Möglichkeit, die Stabilität in die Modellselektion zu integrieren und als Kriterium zu berücksichtigen. Zu jeder Modellschätzung wird deren Stabilität berechnet und Modelle, welche keine genügenden Werte aufweisen, werden abgewiesen. Ein Nachteil dessen ist der höhere Rechenaufwand, da für jedes Modell zusätzliche Durchgänge benötigt werden.

Kreuzvalidierungsverfahren unterscheiden sich in erster Linie anhand der Strategie, mit der die Daten “getrennt” werden. In der Regel wird dafür ein genug grosser Datensatz herangezogen und unterteilt, wobei vorausgesetzt wird, dass die Teilstichproben unabhängig und gleich verteilt sind. Bei k -Facher-Kreuzvalidierung wird der Datensatz in k möglichst gleich grosse Teile aufgeteilt und k Testläufe durchgeführt. Der Durchschnitt aus den Einzelfehlerquoten der k Durchläufe entspricht der Gesamtfehlerquote welche für jede Lösung verglichen werden kann (, , p. 14). Je niedriger die Gesamtfehlerquote, desto stabiler ist die Regressionsgleichung. Um der gleichverteilung der Untermengen gerecht zu werden diese gelegentlich auch stratifiziert (,). Weitere Verfahren und deren Vergleiche finden sich bei ? ().

Die Kreuzvalidierung ist ein gutes Mittel um Überanpassung entgegen zu wirken. Darüber hinaus ist die Stabilität ein guter Indikator für die Generalisierbarkeit. Bedingung ist jedoch, dass dadurch zusätzlich Datensätze zur Verfügung stehen, was im bereich der Psychologischen Forschung durchaus nicht immer gegeben ist.

Software zu den vorgestellten Verfahren

Die bisher vorgestellten Verfahren sind in den meisten grösseren Statistikprogrammen bereits integriert oder können als Erweiterung hinzugefügt werden. Insbesondere wenn es darum geht verschiedene Verfahren der Modellselektion zu beurteilen bietet sich R an.

R ist eine frei verfügbare Programmiersprache für statistisches Rechnen und setzt momentan den Standart im Bereich der Rechnergestützten Datenanalyse. Eine guter Einstieg in R, mit vielen interaktiven Übungen, bietet der Kurs “tryR”¹ von code school. Für das tägliche Arbeiten mit R ist ? () empfehlenswert.

¹<http://tryr.codeschool.com>

Modellselektion

Das exhaustive Verfahren wurde im Paket “leaps” von ? (?) implementiert. Ausgegeben werden kann Mallows’s C_p , R^2 oder auch das adjustierte Bestimmtheitsmass \bar{R}^2 .

Die schrittweise Regression ist ein fester Bestandteil von R (?, ?) und ermöglicht eine bestehende Gleichung schrittweise vorwärts, rückwärts oder beidseitig zu durchsuchen. Als Kriterium wird dabei Akaikes Informationskriterium verwendet, da *step(object,...)* eine vereinfachte Implementierung der Funktion *stepAIC(object,...)* aus dem Paket “MASS” darstellt (?, ?). Ausführlicher werden die kriteriumsbasierende Verfahrens mit R bei ? (?) beschrieben.

Kreuzvalidierung

Für die k -Fache-Kreuzvalidierung bietet sich die Funktion *CVlm(...)* aus dem Paket “DAAG” an (?, ?). Die Funktion bietet über die reine Berechnung hinaus die Möglichkeit die k Durchgänge grafisch auszugeben.

Zusammenfassung / Diskussion

Der Einsatz automatisierter Verfahren zur Bestimmung von Regressionsgleichungen ist umstritten. Während die einen Autoren dies als ein probates exploratives Mittel ansehen lehnen andere insbesondere die schrittweise Regression ab.

? (?, p. 57) erwähnt dabei eine ganz generelle Schwierigkeit: “It allows us to not think about the problem.”. Moderne leicht zu bedienende Statistikprogramme gepaart mit der Möglichkeit schier grenzenloser Rechenkapazität verführen einem nach Effekten zu fischen und Datamining zu betreiben. In der Grundlagenforschung der Psychologie gilt es jedoch hypothesengeleitet vorzugehen. Es muss zuvor über das Problem nachgedacht werden bevor mittels automatisierte Verfahren explorativ Hypothesen erarbeitet werden. Mögliche Abhängigkeiten

zwischen potentiellen Prädiktoren sind im Vorfeld zu beachten. Dies zeigt und insbesondere das Problem der Kolinearität.

Es wurden zwei Verfahren zur automatisierten Modellwahl vorgestellt, exhaustive und schrittweise Verfahren. Der Vorteil exhaustiver Verfahren liegt darin, dass durch komplettes durchrechnen das zum jeweiligen Trainingsdatensatz gefundene optimale Modell gefunden wird. Im Anschluss

Glossar

Datamining Systematische Anwendung statistischer Methoden auf einen grossen Datenbestand mit dem Ziel, neue Muster zu erkennen. 13

Literatur

- Alan Miller, T. L. (2013). leaps: regression subset selection [Software-Handbuch]. Verfügbar unter <http://cran.r-project.org/web/packages/leaps> (R package version 2.9 using Fortran)
- Arlot, S. & Celisse, A. (2010). A survey of cross-validation procedures for model selection. *Statistics Surveys*, 4, 40–79.
- Armstrong, J. (2011). Illusions in regression analysis. *Available at SSRN 1969740*.
- Bendel, R. B. & Afifi, A. A. (1977). Comparison of stopping rules in forward “stepwise” regression. *Journal of the American Statistical Association*, 72 (357), 46–53.
- Berk, K. N. (1978). Comparing subset regression procedures. *Technometrics*, 20 (1), 1–6.
- Borges, B., Goldstein, D. G., Ortmann, A. & Gigerenzer, G. (1999). Can ignorance beat the stock market. *Simple heuristics that make us smart*, 59, 72.
- Bortz, J. & Schuster, C. (2011). *Statistik für Human-und Sozialwissenschaftler* (Bd. 6). Springer.
- Box, G. E. (1979). *Robustness in the strategy of scientific model building*. (Bericht). DTIC Document.
- Derksen, S. & Keselman, H. (1992). Backward, forward and stepwise automated subset selection algorithms: Frequency of obtaining authentic and noise variables. *British Journal of Mathematical and Statistical Psychology*, 45 (2), 265–282.
- Diamantidis, N., Karlis, D. & Giakoumakis, E. (2000). Unsupervised stratification of cross-validation for accuracy estimation. *Artificial Intelligence*, 116 (1), 1–16.
- Faraway, J. (2002). Practical regression and anova using r: <http://cran.r-project.org/doc/contrib.Faraway-PRA.pdf>.
- Gauss, C. F. (1809). *Theoria motus corporum coelestium in sectionibus conicis solem ambientium*.
- Gilmour, S. G. (1996). The interpretation of mallows’s c_p-statistic. *The Statistician*, 49–56.
- Graham, M. H. (2003). Confronting multicollinearity in ecological multiple regression. *Ecology*, 84 (11), 2809–2815.
- Harrell, F. E. (2001). *Regression modeling strategies: with applications to linear models, logistic regression, and survival analysis*. Springer.
- Jacob Cohen, P. C., West, L. S. A., Cohen, J. & West, S. G. A. (2003). *Applied multiple*

- regression/correlation analysis for the behavioral sciences/*. Mahwah, NJ, USA: Lawrence Erlbaum Associates.
- Maeda, J. (2006). The laws of simplicity (simplicity: Design, technology, business, life).
- Maindonald, J. & Braun, W. J. (2013). Daag: Data analysis and graphics data and functions [Software-Handbuch]. Verfügbar unter <http://CRAN.R-project.org/package=DAAAG> (R package version 1.16)
- Mundry, R. & Nunn, C. L. (2009). Stepwise model fitting and statistical inference: turning noise into signal pollution. *The American Naturalist*, 173 (1), 119–123.
- R Development Core Team. (2011). R: A language and environment for statistical computing [Software-Handbuch]. Vienna, Austria. Verfügbar unter <http://www.R-project.org/> (ISBN 3-900051-07-0)
- Teetor, P. (2011). *R cookbook* (1. Aufl.). O'Reilly. Verfügbar unter <http://oreilly.com/catalog/9780596809157>
- Thompson, M. L. (1978). Selection of variables in multiple regression: Part I. A review and evaluation. *International Statistical Review/Revue Internationale de Statistique*, 1–19.
- Venables, W. N. & Ripley, B. D. (2002). *Modern applied statistics with s* (Fourth Aufl.). New York: Springer. Verfügbar unter <http://www.stats.ox.ac.uk/pub/MASS4> (ISBN 0-387-95457-0)
- Weakliem, D. L. (2004). Introduction to the special issue on model selection. *Sociological Methods & Research; Sociological Methods & Research*.

Anhang

Selbstständigkeitserklärung zur Literaturarbeit am Psychologischen Institut

Originalarbeit Ich erkläre ausdrücklich, dass es sich bei der von mir eingereichten schriftlichen Arbeit mit dem Titel “Automatische Verfahren zur Prädiktorauswahl in Regressionsmodellen” um eine von mir selbst und ohne unerlaubte Beihilfe sowie in eigenen Worten verfasste Originalarbeit handelt. Sofern es sich dabei um eine Arbeit von mehreren Verfasserinnen oder Verfassern handelt, bestätige ich, dass die entsprechenden Teile der Arbeit korrekt und klar gekennzeichnet und der jeweiligen Autorin oder dem jeweiligen Autor eindeutig zuzuordnen sind. Ich bestätige überdies, dass die Arbeit als Ganze oder in Teilen weder bereits einmal zur Abgeltung anderer Studienleistungen an der Universität Zürich oder an einer anderen Universität oder Ausbildungseinrichtung eingereicht worden ist noch inskünftig durch mein Zutun als Abgeltung einer weiteren Studienleistung eingereicht werden wird.

Verwendung von Quellen Ich erkläre ausdrücklich, dass ich sämtliche in der oben genannten Arbeit enthaltenen Bezüge auf fremde Quellen (einschliesslich Tabellen, Grafiken u. Ä.) als solche kenntlich gemacht habe. Insbesondere bestätige ich, dass ich ausnahmslos und nach bestem Wissen sowohl bei wörtlich übernommenen Aussagen (Zitaten) als auch bei in eigenen Worten wiedergegebenen Aussagen anderer Autorinnen oder Autoren (Paraphrasen) die Urheberschaft angegeben habe.

Sanktionen Ich nehme zur Kenntnis, dass Arbeiten, welche die Grundsätze der Selbstständigkeitserklärung verletzen – insbesondere solche, die Zitate oder Paraphrasen ohne Herkunftsangaben enthalten –, als Plagiat betrachtet werden und die entsprechenden rechtlichen und disziplinarischen Konsequenzen nach sich ziehen können (gemäss §§ 7ff der Disziplinarordnung der Universität Zürich sowie § 36 der Rahmenordnung für das Studium in den Bachelor- und Masterstudiengängen der Philosophischen Fakultät der Universität Zürich).

Ich bestätige mit meiner Unterschrift die Richtigkeit dieser Angaben.

Name: Markus Graf

Matrikelnummer: 08-91271-9

.....

Zürich, 23. Mai 2013