Automatische Verfahren zur Prädiktorauswahl in Regressionsmodellen

Literaturarbeit vorgelegt von
Markus Graf (markus.graf@uzh.ch)
am Psychologischen Institut der Universität Zürich
Betreut durch Dr. Christina Werner
22. Juni 2013

Ziel der multiplen Regression ist es Kriteriumsvariablen durch mehrere Prädiktorvariablen möglichst gut vorherzusagen. In diesem Kontext kommen automatische Modellwahlverfahren zur Anwendung, wenn für die Schätzung des Modells viele potentielle Prädiktoren zur Auswahl stehen, insbesondere wenn theoretische Grundlagen fehlen. Das exhaustive Verfahren in Kombination mit der Kreuzvalidierung ist momentan die einzige Technik, die das beste und stabilste Modell findet. Schrittweise Verfahren kommen zur Anwendung bei kleinem Stichprobenumfang. Während früher aus Mangel an Rechenleistung standardmässig schrittweise Verfahren angewandt wurden, soll heutzutage das rechenintensive exhaustive Verfahren bevorzugt werden.

Inhaltsverzeichnis

Einführung

Das Standardverfahren, um eine Kriteriumsvariable durch Prädiktorvariablen vorherzusagen, 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 (Gauss, 1809).

Im Unterschied zur einfachen linearen Regression werden in einem multiplen Regressionsmodell mehrere Prädiktoren p mit einbezogen. Es resultiert eine Regressionsgleichung, welche zur Vorhersage einer Kriteriumsvariable y_i aufgrund mehrerer Prädiktorvariablen genutzt wird (Bortz & Schuster, 2011, S. 448).

$$y_i = \beta_0 + \beta_1 \cdot x_{i1} + ... + \beta_p \cdot x_{ip} + \epsilon_i \qquad \text{(multiple lineare Regression)}$$

Beim klassisch hypothesengeleiteten Vorgehen wird eine Hypothese definiert, welche empirisch getestet wird. Der empirische Test wiederum ist ein Modell, in unserem Fall eine Regressionsgleichung, welche aufgrund theoretischer Überlegungen erstellt wurde. Wenn es jedoch keine klaren theoretischen Gründe gibt potentielle Prädiktorvariablen in das Modell aufzunehmen, werden mehrere Modelle geschätzt und jenes mit der besten und stabilsten Vorhersagekraft verwendet. Bei komplexen Modellen mit vielen Prädiktoren werden Modelle mittels automatischer Verfahren geschätzt und selektiert. Ein zentrales Problem solcher Verfahren ist, dass sie dazu neigen, zu komplexe Modelle zu schätzen. Komplexe Modelle können sehr gute Vorhersagen innerhalb des Trainingsdatensatzes liefern, doch scheitern gern beim Versuch, generelle Vorhersagen zu treffen.

Im folgenden wird diskutiert, wann und weshalb automatische Verfahren zur Modellwahl eingesetzt werden. Anschliessend wird das exhaustive und schrittweise Verfahren vorgestellt und kritisch diskutiert. Die Frage nach der Generalisierbarkeit automatisch geschätzter Modelle wird im Anschluss besprochen und die Kreuzvalidierung als Lösungsansatz genannt.

Recherche

Hauptquelle der Literatur-Recherche waren Artikel, die via Google Scholar gefunden wurden. Als Stichworte zu nennen sind *stepwise model selection, stepwise regression criteria, model selection paradigm, cross validation, overfitting.* Grundlagen zu den Verfahren wurden mittels Bortz und Schuster (2011) und Cohen, Cohen, West und Aiken (2003) erarbeitet. Der Fokus bei der Recherche wurde auf Artikel im Bereich der Psychologie gerichtet. Es fanden sich jedoch auch viele Artikel in anderen Fachbereichen, die mit den selben Problemen konfrontiert sind. Kurz vorgestellt wurden die Verfahren von Carolin Strobl in der Vorlesung "160 Psychologische Methoden: Datenerhebung, Analyse und Darstellung" im Rahmen des Psychologie-Aufbaustudiums der Universität Zürich. Interessante Hinweise und praktische Beispiele fanden sich ausserdem in den Manuals von R (R Development Core Team, 2011).

Sinn und Zweck automatisierter Modellwahl

In psychologischen Fragestellungen kommt es vor, dass viele Prädiktoren in ein Modell einfliessen oder potentiell für ein Modell in Frage kommen. Die Frage, welche Prädiktoren nun ein Modell am besten beschreiben ist dabei die eigentliche Gretchenfrage.

Unterteilen lässt sich die automatisierte Modellwahl in (a) eine explorative, und (b) eine optimierende Anwendung. Im Falle der explorativen Anwendung fehlen grösstenteils theoretische Begründungen für die Auswahl bestimmter Prädiktoren. Ein Beispiel für eine solche Anwendung liefert eine Studie, die Prädiktoren des Alltagstransfers eines stationär erlernten Entspannungstrainings suchte (Bernardy, Krampen & Köllner, o. J.). Oft werden Daten gleich für mehrere Studien erhoben und in anderen Studien verwendet. Diese systematische Anwendung automatischer Modellwahlverfahren, mit dem Ziel neue Muster zu erkennen, ist damit eine Aufgabenstellung des Datamining. Die so gewonnen Daten können zu neuen Fragestellungen führen.

Der zweite Anwendungsfall ergibt sich, wenn bereits ein Modell vorhanden ist. Insbesondere komplexe Modelle sind meist schlecht generalisierbar. Automatisierte Verfahren können helfen, Prädiktoren zu erkennen, welche die Komplexität unnötig erhöhen.

Automatische Verfahren zur Prädiktorauswahl

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 Langlebigkeit 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 (Armstrong, 2012, p.2). Jeder durchschnittliche Computer erledigt dies heutzutage in Sekundenbruchteilen. Mit dem technischen Fortschritt einhergehend wurden Verfahren entwickelt, welche alle möglichen Kombinationen von Prädiktoren, inklusive ihrer Interaktionen, berücksichtigen und gegeneinander testen.

Es gilt also, das "beste" Modell zu schätzen. Gemeint ist mit dem "besten" Modell das, das innerhalb des Trainingsdatensatzes die beste Vorhersage liefert. Anhand des Trainingsdatensatzes wurde das Modell jedoch auch geschätzt. Entsprechend kann es Modelle geben, die in der Gesamtpopulation bessere Vorhersagen liefern. "All models are wrong, but some are useful" (Box & Draper, 1987, p.424). Box will damit hervorheben, dass obschon in der Literatur oft vom "besten" oder "wahren" Modell gesprochen wird, dies nur eine Approximation der Wirklichkeit darstellt (Weakliem, 2004, p.172).

Exhaustive Schätzung

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. Das Modell mit der kleinsten

Fehlerquadratsumme SSE_p wird als das optimale Modell bezeichnet (Thompson, 1978, p. 6).

$$SSE_p = \sum_{i=1}^{n} (y_{ip} - \hat{y}_p)^2 \qquad \text{(Fehlerquadratsumme)}$$

Da alle möglichen Kombinationen durchgerechnet werden, wird das Modell gefunden, das den Trainingsdatensatz am besten vorhersagt. Thompson (1978, p.6) sieht einzig den Nachteil darin, dass der Rechenaufwand exponentiell mit der Anzahl zu berücksichtigender Prädiktoren steigt. Es müssen immer $2^p - 1$ Modelle berechnet werden: Bei 5 Prädiktoren 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.

Schrittweise Verfahren

Das optimale Modell beinhaltet jeden Prädiktor, der die Voraussage auch nur minimal verbessert. Es stellt sich die Frage, ob diese minimale Verbesserung auch nützlich ist oder einfach durch Zufall entstanden ist. Schrittweise Verfahren arbeiten wesentlich liberaler. Prädiktoren werden hinzugefügt oder eliminiert, je nach deren Relevanz für die Modellgüte. Es werden Kriterien festgelegt, nach welchen ein Modell als angemessen zu betrachten ist. Dies hat gegenüber dem exhaustive Verfahren den Vorteil, dass nicht alle Modelle berechnet werden müssen und entsprechend schneller Lösungen gefunden werden.

Innerhalb der schrittweisen Verfahren unterscheidet man zwischen *Forward Selection* und *Backward Elimination*. Ausgehend vom leeren Modell werden in der ersten Variante schrittweise weitere Variablen der Nützlichkeit nach in das Modell integriert. Dies dauert so lange an, bis kein Prädiktor mehr gefunden wird, der ein gewisses Kriterium erfüllt.

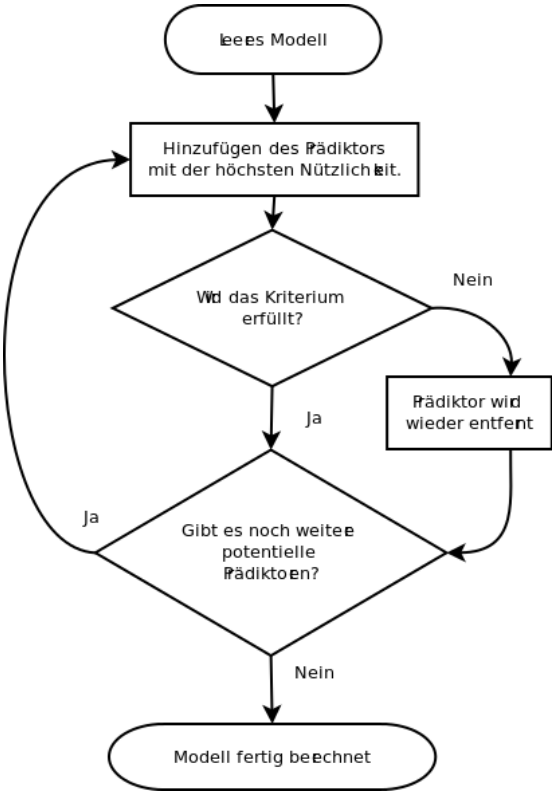

Abbildung 1. Forward Selection. Das Flussdiagramm beschreibt den schrittweisen Aufbau eines neuen Modells aus dem leeren Modell durch Hinzufügen potentieller Prädiktoren.

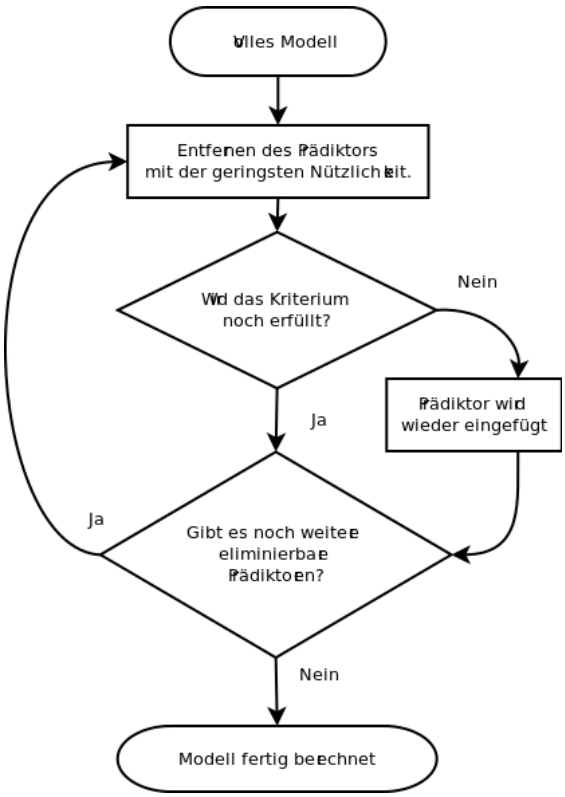

Abbildung 2. Backward Elimination. Das Flussdiagramm beschreibt die schrittweise Elimination von unnützen Prädiktoren aus dem vollen Modell.

In der zweiten Variante werden alle Prädiktoren in das Modell integriert und sukzessive nacheinander entfernt. Wiederum endet das Verfahren, sobald kein Prädiktor mehr weggelassen werden kann, ohne dass ein gewisses Kriterium unterschritten wird.

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 in die Regressionsgleichung aufgenommen, die Vorhersage kaum verbessern und das Kriterium nicht erfüllen. Zusammen leisten sie jedoch einen substantiellen Beitrag (Cohen et al., 2003, p.261). Schrittweise Verfahren mittels Forward Selection sind entsprechend nicht in der Lage solche Effekte mit zu berücksichtigen, wogegen Backward Elimination robuster gegen solche Supressionseffekte ist (Shieh, 2006).

Zentrales Element der schrittweisen Regression ist, das Kriterium zur Beurteilung der Modellanpassung, welches besagt, weshalb und wann ein Modell als akzeptabel zu betrachten ist. Als Folge dessen, wird damit auch die Anzahl relevanter Prädiktoren bestimmt. Im Laufe der Zeit wurden diverse Kriterien definiert, welche alle für sich ihre Berechtigung haben. Einteilen lassen sie sich in Kriterien, welche (a) sich auf die Beurteilung innerhalb des Trainingsdatensatzes beschränken oder (b) die Generalisierbarkeit ausserhalb des Trainingsdatensatzes zu berücksichtigen versuchen. Letztere werden im Abschnitt des Overfittings beschrieben.

Interessiert uns die Modellgüte innerhalb des Trainingsdatensatzes bietet das Bestimmtheitsmass erste Hinweise. Im Fall der multiplen Regression entspricht dies dem Quadrat des multiplen Korrelationskoeffizienten R^2 und besagt, wie viel systematische Varianz durch das Modell 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).

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

In schrittweisen Verfahren wird nicht einzig aufgrund von R^2 selektiert, sondern es wird zusätzlich getestet, ob Verbesserungen nicht durch Zufall entstanden sind.

Beim Signifikanztest als Kriterium wird das Verfahren beendet, wenn kein Prädiktor mehr hinzugefügt werden kann, der das Vorhersagepotential signifikant erhöht (Bendel & Afifi, 1977, p.48). Das Vergleichen zweier Regressionsgleichungen mittels Signifikanztest bedingt, dass diese geschachtelt sein müssen, das heisst, das kleinere Modell muss im grösseren enthalten sein (Cohen et al., 2003, p. 508). Das gewählte Signifikanzniveau ist eigentlich arbiträr gewählt und eher als Konvention zu betrachten. Das Signifikanzniveau soll grundsätzlich die relativen Kosten der Irrtumswahrscheinlichkeit beschreiben (Hansen, 1999, p. 196). Derksen und Keselman (1992, p. 269) diskutieren mehrere Empfehlungen für Signifikanzniveaus und weisen darauf hin, das sich über mehrere Tests der α-Fehler kumuliert. In Simulationen mit artifiziellen Daten zeigen Mundry und Nunn (2009) das Problem multipler Tests beispielhaft auf. Daraus resultierend lehnen sie die Verwendung der schrittweisen Regression mittels Signifikanztest gar ab.

Eine weitere Schwäche des Hypothesentestens ist der Einfluss der Stichprobengrösse. So kann bei genügend grossem Umfang nahezu jedes Modell signifikant werden, was wiederum zu komplexen und überangepassten Modellen führt (Weakliem, 2004, p.173).

Entgegen dem exhaustiven Verfahren besteht bei Schrittweisen das Problem, dass unter Umständen nicht das optimale Modell gefunden wird. Bortz und Schuster (2011, p. 462) bezeichnen diese Verfahren eher als explorativ, da die Nützlichkeitsunterschiede, die oft nur geringe statistische Bedeutung haben, das Modell bestimmen. Harrell (2001, p. 56ff) lehnt das Verfahren gar ab und führt ins Feld, dass sämtliche statistischen Prinzipien verletzt würden. Berk (1978)

zeigte in einem Vergleich, dass die durchschnittliche Differenz der Fehlerquadratsummen zwischen exhaustiver und schrittweiser Regression kaum 7% übertrifft. Bei Modellen, welche im Allgemeinen bereits viel Varianz aufklären, mag dies akzeptabel sein. Wenn hingegen zu befürchten ist, dass wenig Varianz aufgeklärt werden kann, könnte diese Differenz jedoch das Zünglein an der Waage sein.

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 Prädikator und dem Kriterium steig und (d) selbst wenn die Korrelation zwischen konfundierendem Prädikator und Kriterium klein ist, besteht die Gefahr, dass korrekt schwach signifikante Prädiktoren nicht signifikant werden (Graham, 2003, p. 2810).

Es gilt die Voraussetzung, dass alle Prädiktoren erhoben wurden um das Modell zu definieren und diese sauber gemessen wurden. Bei Verletzung dieser Voraussetzung sind die Residuen nicht unabhängig und die Regressionskoeffizienten sind verzerrt (Cohen et al., 2003, p. 119). Doch gerade in der psychologischen Forschung lassen sich Abhängigkeiten zwischen Prädiktoren 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 mehrmals erhoben. Der Zusammenhang zwischen einer Prädikatorvariable i und der vorhergesagten Kriteriumsvariable lässt sich mit dem Strukturkoeffizienten c_i ausdrücken.

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

r_{ic} beschreibt den einzelnen Korrelationskoeffizient zwischen dem Prädikator i und dem Kriterium c und R den multiplen Korrelationskoeffizient. Gilt es kollineare Prädiktoren zu entfernen, kann so jener eliminiert werden, welcher schlechter mit dem Kriterium korreliert (Bortz & Schuster, 2011, S. 453).

Overfitting

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 oft das Umgekehrte der Fall ist (Armstrong, 2012, p.3). Insbesondere Gigerenzer demonstrierte eindrucksvoll, wie mit einfachen Rekognitionsheuristiken bessere Vorhersagen gemacht werden konnten als mit komplexen statistischen Modellen (Borges, Goldstein, Ortmann & Gigerenzer, 1999). Komplexe Modelle können sehr gute Vorhersagen innerhalb des Trainingsdatensatzes liefern, doch scheitern oft beim Versuch der Generalisierung.

Die Illusion der Komplexität ist auch in der Statistik anzutreffen (Armstrong, 2012, p. 3). Im Kontext der Modellwahl äussert sich dies in Form des Overfittings. Das heisst, dass das Modell zu sehr an den Trainingsdatensatz angepasst ist. Insbesondere Modellwahlverfahren, welche die Anzahl der Prädikatoren nicht bestrafen, sind davon betroffen. Als Einflussfaktoren seitens der Daten sind Repräsentativität und Stichprobengrösse zu nennen. Mit steigendem Stichprobenumfang und höherer Repräsentativität sinkt das Overfitting 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 von Overfitting beurteilen nicht nur die Güte des Modells, sondern strafen auch Komplexität ab. 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 aufbaut und sowohl die Prädikatoranzahl p als auch die Stichprobengrösse n berücksichtigt.

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

Angestrebt wurde dabei, alle wichtigen Prädikatoren zu berücksichtigen. Das "beste" Modell ist das mit (a) dem niedrigsten C_p-Wert, der (b) möglichst gleich p ist (Gilmour, 1996). Angewendet wird dieses Kriterium insbesondere in Kombination mit dem exhaustiven Verfahren.

Die bisher auf der Methode der kleinsten Quadrate basierenden schrittweisen Verfahren bedingen, dass die Prädikatoren geschachtelt sind. Das soll heissen, dass jeweils alle Prädiktoren des kleineren Modells im grösseren enthalten sein müssen. Dies wird für die beiden Kennwerte, Akaikes Informationskriterium (AIC) und Bayessches Informationskriterium (BIC) nicht vorausgesetzt. Beide Kennwerte basieren auf der Maximum-Likelihood-Methode L und berücksichtigen die Anzahl Prädiktoren p. Dem Prinzip der Sparsamkeit entsprechend ergeben kleinere Modell bei gleicher Vorhersagekraft bessere Kennwerte (Cohen et al., 2003, p. 509). Bei Regressionsmodellen mit normalverteilten Fehlern entspricht die Wahrscheinlichkeitsfunktion L der des quadrierten Standardfehlers der Regression σ^2 (Weakliem, 2004, p. 169).

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

Gegenüber AIC ist BIC konservativer, denn es wird zusätzlich der Stichprobenumfang n stärker berücksichtigt (Weakliem, 2004, p. 169).

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

Durch die Berücksichtigung der Komplexität im Kriterium ist es möglich innerhalb des schrittweisen Verfahren, die Anzahl der Prädiktoren zu verringern.

Berücksichtigt werden dabei jedoch nur Daten aus dem Trainingsdatensatz. Entsprechend fehlt die Möglichkeit, diese Vereinfachung empirisch zu rechtfertigen. Das soll heissen, dass man nicht sagen kann, ob ein komplexeres und entsprechend exaktes Modell gerade in diesem Fall wirklich schlechter generalisierbar wäre. Um dies zu bewerkstelligen müssen die Vorhersagen der Modelle mittels unabhängiger Datensätze verglichen werden, was mittels Kreuzvalidierung erreicht werden kann.

Kreuzvalidierung

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

Die Idee hinter der Kreuzvalidierung liegt darin, die Daten aufzuteilen. Zum einen in eine Trainingsstichprobe, anhand deren die Gleichung geschätzt wird, zum anderen in eine oder mehrere zusätzliche Teststichproben, mittels derer die Stabilität validiert wird. Kennwert der Stabilität ist meist die durchschnittliche Fehlerquote der einzelnen Vorhersagen (Arlot & Celisse, 2010, p. 3).

Die Frage stellt sich, welchen Platz die Kreuzvalidierung in der automatisierten Modellwahl einnimmt. Die Kreuzvalidierung kann über ein Set von n Regressionsgleichungen durchgeführt werden, beispielsweise potentielle Modelle nach einer schrittweisen Regression (Arlot & Celisse, 2010, p. 12). Wird der Kreuzvalidierung das exhaustive Verfahren vorangestellt, macht es wenig Sinn alle $2^p - 1$ Modelle mit einzubeziehen, da die meisten das Kriterium sehr schlecht vorhersagen werden. Es soll entsprechend nur eine Hand voll der vielversprechendsten Modelle beachtet werden. Die 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 das (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. Arlot und Celisse (2010, 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, die keine genügenden Werte aufweisen, werden verworfen. Ein Nachteil hierbei ist der höhere Rechenaufwand, da jedes Modell zusätzliche Durchgänge benötigt.

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 Untermengen unabhängig und gleich verteilt sind. Um die Gleichverteilung der Untermengen zu gewährleisten, werden diese gelegentlich auch stratifiziert (Diamantidis, Karlis & Giakoumakis, 2000). Bei k-facher Kreuzvalidierung wird der Datensatz in k möglichst gleich grosse Teile aufgeteilt und k Testläufe durchgeführt. Bei jedem Durchlauf wird einer der k Teile als Testdatensatz herangezogen und die restlichen $k - 1$ Datensätze bilden, als Trainingsdatensatz, die Grundlage für die Schätzung des Models. Jede der k Teile ist dabei einmal Testdatensatz. Bei jedem Durchgang wird (1.) das Modell geschätzt und (2.) aufgrund dessen die Fehlerquote des Testdatensatzes bestimmt. Ist das Kriterium eine metrische Variable, kann die Fehlerquote anhand der durchschnittlichen Abweichung zum vorhergesagten Wert berechnet werden. Bei nicht metrischem Kriterium kann die Fehlerquote als Fehlklassifikationsrate betrachtet werden. Der Durchschnitt aus allen Fehlerquoten der k Durchläufe entspricht der Gesamtfehlerquote des Modells (Arlot & Celisse, 2010, p. 14). Je niedriger die Gesamtfehlerquote, desto stabiler ist die Regressionsgleichung. Weitere Verfahren und deren Vergleiche finden sich bei Arlot und Celisse (2010).

Die Kreuzvalidierung ist ein gutes Mittel um Overfitting entgegen zu wirken. Die Stabilität ist ein guter Indikator für die Generalisierbarkeit. Sie kann immer auf ein Set potentieller Modelle angewandt werden, unabhängig vom Modellselekti-

onsverfahren. Dies ermöglicht es auch, verschiedene Verfahren zur Modellselektion gegeneinander zu vergleichen. Bedingung ist jedoch, dass dafür zusätzlich Datensätze zur Verfügung stehen, was im Bereich der psychologischen Forschung durchaus nicht immer gegeben ist. Wird beispielsweise davon ausgegangen, dass der Testdatensatz jeweils m Stichproben beinhalten soll, so werden $n = k \cdot m$ Stichproben benötigt um eine k-fachen Kreuzvalidierung durchzuführen.

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 (R Development Core Team, 2011) und setzt momentan den Standard 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 Teetor (2011) empfehlenswert.

Modellselektion

Das exhaustive Verfahren wurde im Paket "leaps" von Alan Miller (2013) 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 (Venables & Ripley, 2002). Ausführlicher werden die kriteriumsbasierenden Verfahren mit R bei Faraway (2002) beschrieben.

Kreuzvalidierung

Für die k-fache Kreuzvalidierung bietet sich die Funktion $CVlm(...)$ aus dem Paket "DAAG" an (Maindonald & Braun, 2013). Die Funktion bietet über die reine Berechnung hinaus die Möglichkeit, die k Durchgänge grafisch auszugeben.

[1] http://tryr.codeschool.com

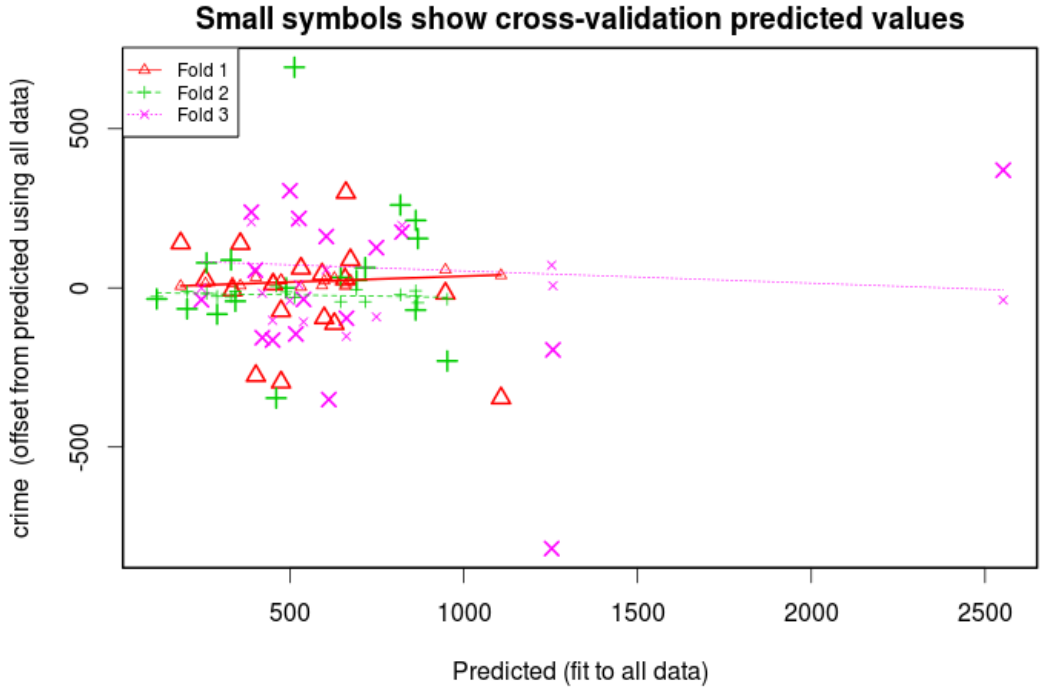

Abbildung 3. Residualplot einer 3-fachen Kreuzvalidierung. Die Abszisse beschreibt die vorausgesagten Werte. Auf der Ordinate ist die Abweichung der Vorhersage zum wahren Wert des Kriteriums abgetragen. Hier Gewaltdelikte pro 100'000 Einwohner.

Diskussion

Automatisierte Verfahren zur Bestimmung von Regressionsgleichungen werden eingesetzt, um Modelle zu optimieren oder explorativ neue Modelle zu generieren, wenn viele potentielle Prädikatoren involviert sind.

Im Falle der Optimierung gilt es insbesondere, ein zu stark an die Trainingsdaten angepasstes Modell zu vereinfachen. Einfachere Modelle sind in der Regel stabiler, was wiederum die Generalisierbarkeit erhöht. Ziel ist es hierbei, Prädikatoren mit geringer Vorhersagekraft zu eliminieren. Mittels exhaustiver Verfahren können sämtliche Kombinationen von den im Modell enthaltenen Prädikatoren geschätzt werden. Aus diesem Set meist einfacherer Modelle kann mittels Kreuzvalidierung die durchschnittliche Fehlvorhersage berechnet werden. Im Idealfall wird ein einfacheres Modell gefunden, das die Daten generell besser vorhersagt als das komplexe Modell. Der Vorteil exhaustiver Verfahren gegenüber den standardmässig eingesetzten schrittweisen Verfahren ist, dass das optimale Modell gefunden wird, insbesondere in Kombination mit der Kreuzvalidierung. Der erhöhte Rechenaufwand sollte heutzutage nicht mehr ins Gewicht fallen, insbesondere da in der Psychologie meist nur eine Handvoll Prädikatoren ein Modell beschreiben.

Die explorative Anwendung ist verbreiteter und dient dem Schätzen von Modellen ohne klare theoretische Begründung. Eine Auslese potentieller Prädikatoren sollte möglichst unkorreliert sein, um der Multikolinearität vorzubeugen. Bei genügend grossem Stichprobenumfang und einer mässigen Anzahl an potentiellen Prädiktoren führt auch hier das exhaustive Verfahren gefolgt von einer Kreuzvalidierung zur optimalen und stabilsten Vorhersage. In der psychologischen Forschung ist der Stichprobenumfang oft knapp bemessen. Kann aufgrund des Stichprobenumfangs eine Kreuzvalidierung nicht verlässlich geschätzt werden, kommen schrittweise Verfahren zum Einsatz. Dabei werden Prädikatoren schrittweise hinzugefügt oder eliminiert, bis ein zuvor bestimmtes Kriterium nicht mehr erfüllt werden kann. Das Kriterium besagt weshalb, und wann ein Modell als akzeptabel zu betrachten ist. Es soll ein Kriterium herbei gezogen werden, welches die Anzahl der Prädiktoren im Modell berücksichtigt um Overfitting entgegen zu

wirken. Akaikes Informationskriterium bietet sich hier als Kriterium der schrittweisen Regression an.

Harrell (2001, p. 57) erwähnte 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 dazu, nach Effekten zu fischen und Datamining zu betreiben. Der Einsatz automatisierter Verfahren zur Bestimmung von Regressionsgleichungen ist umstritten. Während die einen Autoren dies als ein probates Mittel ansehen, lehnen andere insbesondere die schrittweise Regression ab. Schrittweise Verfahren halten sich auch heute noch hartnäckig, obschon es keinen Grund gibt, nicht vollumfänglich alle Modelle durchzurechnen. Die Rechenkapazität sowie die nötige Software ist vorhanden. Forschende sollten sich vom Gedanken lösen, das Resultat gleich unmittelbar zu bekommen und dem Computer eine Kaffeepause lang die Möglichkeit geben, das optimale Resultat zu finden. Zu guter Letzt kann man auch hier die Faustregel anwenden, dass das simpelste Verfahren das beste Resultat ergibt.

Glossar

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

exhaustive Verfahren rechnen alle möglichen Modelle anhand der potentiellen Prädiktoren durch.

Kreuzvalidierung bezeichnet Verfahren, bei denen die Vorhersagezuverlässigkeit eines Modells anhand von unabhängigen Teilstichproben bestimmt wird.

Kriteriumsvariable Erklärte Variable, welche eine Wirkung misst.

Maximum-Likelihood-Methode wird benützt, um die unbekannten Parameter einer Funktion aus gemessenen Daten zu bestimmen.

Modellgüte beschreibt, wie gut ein Modell gegebene Daten vorhersagen kann.

Multikollinearität ist ein Problem der Regressionsanalyse und liegt vor, wenn zwei oder mehr Prädiktoren stark miteinander korrelieren.

Overfitting beschreibt eine mangelnde Generalisierbarkeit aufgrund eines Modells, das zu sehr an die Trainingsstichprobe angepasst ist.

Prädiktorvariable Unabhängige Variable, die einen zu bestimmenden Einfluss auf die Kriteriumsvariable ausübt.

Rechenaufwand beschreibt die Komplexität eines Verfahrens. Die Anzahl der Schritte, die für die Berechnung benötigt werden, dient als Kennzahl.

schrittweise Verfahren rechnen Modelle durch schrittweise Hinzunahme beziehungsweise Weglassen potentieller Prädiktoren.

Stabilität beschreibt, wie stark die Vorhersagen eines Modells im Generellen variieren.

Trainingsdatensatz liefert die Datenbasis für die Schätzung der Modellparameter.

Literatur

Alan Miller, T. L. (2013). *leaps: regression subset selection.* Zugriff auf `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. S. (2012). Illusions in regression analysis. *International Journal of Forecasting*. Zugriff auf `http://dx.doi.org/10.2139/ssrn.1969740` doi: 10.2139/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.

Bernardy, K., Krampen, G. & Köllner, V. (o. J.).

Borges, B., Goldstein, D. G., Ortmann, A. & Gigerenzer, G. (1999). *Can ignorance beat the stock market.* New York, NY: Oxford University Press.

Bortz, J. & Schuster, C. (2011). *Statistik für Human-und Sozialwissenschaftler* (6. Aufl.). Heidelberg: Springer.

Box, G. E. & Draper, N. R. (1987). Empirical model-building and response surfaces. *Willey series in probability and mathematical statistics*.

Cohen, J., Cohen, P., West, S. G. & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences.* Mahwah, NJ: Lawrence Erlbaum.

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*, 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 Aegression and Anova using R.* Zugriff auf `http://cran.r-project.org/doc/contrib/Faraway-PRA.pdf`

Gauss, C. F. (1809). *Theoria motus corporum coelestium in sectionibus conicis solem ambientium.* Hamburg: Frid. Perthes & I.H. Besser.

Gilmour, S. G. (1996). The interpretation of Mallows's Cp-statistic. *The Statistician*, *45*, 49-56.

Graham, M. H. (2003). Confronting multicollinearity in ecological multiple regression. *Ecology*, *84*, 2809-2815.

Hansen, B. E. (1999). Discussion of 'data mining reconsidered'. *The econometrics journal*, *2* (2), 192–201.

Harrell, F. E. (2001). *Regression modeling strategies: With applications to linear models, logistic regression, and survival analysis.* New York, NY: Springer.

Maindonald, J. & Braun, W. J. (2013). *Daag: Data analysis and graphics data and functions.* Zugriff auf `http://CRAN.R-project.org/package=DAAG` (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.* Vienna, Austria. Zugriff auf `http://www.R-project.org/`

Shieh, G. (2006). Suppression situations in multiple linear regression. *Educational and psychological measurement*, *66*, 435-447.

Teetor, P. (2011). *R cookbook* (1. Aufl.). Sebastopol, CA: O'Reilly. Zugriff auf `http://www.cookbook-r.com`

Thompson, M. L. (1978). Selection of Variables in Multiple Regression: Part I. A Review and Evaluation. *International Statistical Review/Revue Internationale de Statistique*, *46*, 1-19. Zugriff auf `http://www.jsor.org/stable/1402505`

Venables, W. N. & Ripley, B. D. (2002). *Modern Applied Statistics with S* (4. Aufl.). New York, NY: Springer. Zugriff auf `http://www.stats.ox.ac.uk/pub/MASS4`

Weakliem, D. L. (2004). Introduction to the Special Issue on Model Selection. *Sociological Methods & Research*, *33*, 167-187.

Anhang

R-Code für Abbildung 3

```
# Beispieldaten entlehnt von http://www.ats.ucla.edu/stat/r/dae/rreg.htm
# crime: Gewaltdelikte pro 100'000 Personen
# pctwhite: Prozent an Personen mit einer Hochschulausbildung
# single: Prozent alleinerziehender Personen

require(foreign)
require(MASS)
require(DAAG)
cdata <- read.dta("http://www.ats.ucla.edu/stat/data/crime.dta")
summary(cdata)
summary(lm(crime ~ pctwhite + single, data = cdata))
CVlm(cdata,form.lm = formula(crime ~ pctwhite + single), m=3, plotit="Residual")
```

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, 22. Juni 2013