Variablenselektion im linearen Regressionsmodell mit der experimentellen Prozedur PROC GLMSELECT

Brigitte Hörmann
Institut für Biometrie,
Universität Ulm
Schwabstraße 13
89075 Ulm
brigitte.hoermann@gmx.de

PD Dr. Rainer Muche
Institut für Biometrie,
Universität Ulm
Schwabstraße 13
89075 Ulm
rainer.muche@uni-ulm.de

Zusammenfassung

Die ab der SAS-Version 9.1 experimentelle Prozedur PROC GLMSELECT bietet einen wesentlich erweiterten Umfang an Möglichkeiten bei der Variablenselektion im linearen Regressionsmodell als z. B. die Prozedur PROC REG. Neben den klassischen Möglichkeiten der Stepwise-Verfahren (Forward, Backward, Stepwise) werden die modernen Modellwahlverfahren LASSO und LAR bereitgestellt. Das Angebot an Modellgüte-Kriterien, die in der neuen Prozedur an verschiedenen Stellen während des Selektionsprozesses eingesetzt werden können, wird ausgeweitet. Gegenüber der Standard-Prozedur PROC REG kann man darüber hinaus wesentlich einfacher Variablenselektionen automatisieren, speziell bei der Einbindung von kategoriellen Variablen (Dummy-Kodierung) und der Angabe von Wechselwirkungen. Einige Einstellmöglichkeitern dieser neuen Prozedur PROC GLMSELECT werden an Hand eines Beispiels aus dem medizinischen Umfeld demonstriert, um einen Überblick zu geben, wie die neue Prozedur eingesetzt werden kann und welche Vorteile, aber auch welche Einschränkungen mit ihrem Einsatz verbunden sind.

Schlüsselwörter: Variablenselektion, PROC GLMSELECT, LASSO, LAR

1 Einleitung

Die folgende Ausführung soll eine Einführung in die neue SAS-Prozedur PROC GLMSELECT geben. Dabei wird zunächst kurz auf das Ziel sowie auf einen speziellen Anwendungsfall, die Einbindung kategorieller Variablen bei der Variablenselektion, eingegangen. Der Nennung der bereits bekannten SAS-Prozeduren für die Variablenselektion folgt die Beschreibung einiger wichtiger Funktionalitäten der neuen Prozedur PROC GLMSELECT. Die Einstellmöglichkeiten der neuen Prozedur werden anschließend an Hand eines Datensatzes aus dem medizinischen Umfeld in Form von drei Vergleichen demonstriert.

2 Variablenselektion

Das Ziel der Variablenselektion ist es, aus einer großen Auswahl von potentiell erklärenden Variablen in einem Datensatz diejenigen herauszufinden, mit denen ein Regressionsmodell die Vorhersage der Zielvariablen am Besten beschreibt. Die Variablenaus-

wahl erfolgt zum einen aus inhaltlichen Gesichtspunkten und zum anderen mit Hilfe unterschiedlicher statistischer Verfahren. Bekannte Selektionsmethoden sind beispielsweise die Backward-, Forward- oder Stepwise-Verfahren [4].

2.1 Einbindung von kategoriellen Variablen

In der Regressionsanalyse werden im Allgemeinen lineare Zusammenhänge zwischen der abhängigen und den unabhängigen Variablen untersucht. Dieser Zusammenhang ist bei Variablen mit ordinalen oder auch nominalen Ausprägungen nicht gegeben, da die Abstände zwischen den Ausprägungen nicht definiert oder auch nicht gleich sind. Mit Hilfe der Dummy-Kodierung können jedoch die Sprünge zwischen den Klassen einzeln modelliert werden [3]. Ob die Umsetzung der Kodierung in SAS „von Hand“ oder automatisch erfolgt, hängt vom Vorhandensein eines CLASS-Statements in der eingesetzten Prozedur ab. Für die Variablenselektion ist es wichtig, dass Dummy-Variablen gemeinsam überprüft und als Gruppe in das Modell aufgenommen oder aus dem Modell ausgeschlossen werden. Denn sind nur Teile einer Gruppe von Dummy-Variablen im Modell, lässt sich das Ergebnis nur schwer interpretieren.

2.2 SAS-Prozeduren für lineare Regression

Eine klassische Prozedur für lineare (multiple) Regressionsprobleme ist die Prozedur PROC REG, die verschiedene Verfahren zur Variablenselektion und weitere Regressionsdiagnostiken bereitstellt. Die Prozedur verfügt über kein CLASS-Statement, so dass Dummy-Kodierungen daher separat in einem vorherigen Data-Step erfolgen müssen.
Eine weitere Prozedur für die lineare Regression ist PROC GLM, die jedoch keine Möglichkeit zur automatischen Variablenselektion bietet, deren Vorteil eines CLASS-Statements jedoch in die neue Prozedur übernommen wurde, was in der Prozedurendokumentation folgendermaßen beschrieben ist: „The REG procedure supports a variety of model-selection methods but does not support a CLASS statement. The GLM procedure supports a CLASS statement but does not include effect selection methods. The GLMSELECT procedure fills this gap.” [1]

3 PROC GLMSELECT

Seit August 2005 bietet SAS die neue Prozedur PROC GLMSELECT an. Bei dieser Prozedur handelt es sich um eine noch experimentelle Version, die über die SAS-Homepage zum Download zur Verfügung steht. Voraussetzung für die Installation ist eine Windows-Umgebung und eine SAS 9.1-Version. Zusätzlich steht auch ein Beispielprogramm und die 104-seitige Prozeduren-Dokumentation zum Download bereit [5].
Neben zahlreichen Möglichkeiten, wie der einfachen Einbeziehung von Wechselwirkungen, der Erzeugung von Makrovariablen und der Ermöglichung der Selektion einer großen Anzahl von Effekten (>1000), sind folgende Punkte von besonderer Bedeutung und sollen bei der weiteren Betrachtung der Prozedur im Vordergrund stehen.

- Vorhandensein eines CLASS-Statements
- Neue Selektionsmethoden LASSO und LAR (neben Backward-, Forward- und Stepwise-Selektion)
- Vielfältige Auswahl an Modellgüte-Kriterien als
  - Selektionskriterium
  - Stoppkriterium
  - Auswahlkriterium
- Validierungsmöglichkeiten
- Grafische Darstellung des Selektionsprozesses

Um die folgenden Erläuterungen der Prozeduren-Statements (hier dick gedruckt) in die allgemeine GLMSELECT-Syntax einordnen zu können, ist diese mit einigen Options-Möglichkeiten dargestellt:

```
PROC GLMSELECT DATA=... TESTDATA=... VALDATA=... PLOTS=... ;
     BY Variablen ;
     CLASS Variablen / PARAM=... REF=... SPLIT ;
     FREQ Variable ;
     MODEL Zielvariable = Variablen
           / SELECTION =...(SELECT=... STOP=... CHOOSE=...);
     OUTPUT OUT =... ;
     PARTITION FRACTION(TEST=... VALIDATE=...);
     PERFORMANCE;
     SCORE;
     WEIGHT;
RUN;
```

3.1 CLASS-Statement

PROC GLMSELECT bietet über das CLASS-Statement die Möglichkeit, aus kategoriellen Variablen automatisch Dummy-Variablen zu generieren. Nach welchem Kodierungsverfahren dabei vorgegangen werden soll, kann über die Option PARAM= gewählt werden. Die Gruppe der Dummy-Variablen wird gemeinsam in das Modell aufgenommen oder aus ihm entfernt, außer man möchte eine separate Selektionsmöglichkeit über die Option SPLIT erreichen. Bei den neuen Verfahren LASSO und LAR werden alle Variablen jedoch immer einzeln selektiert, so dass eine gemeinsame Betrachtung von Dummy-Variablen nicht möglich ist.

3.2 Neue Selektionsmethoden LASSO und LAR

Als unterschiedliche Selektionsmethoden sind in der neuen Prozedur zusätzlich zu den Standardverfahren Backward-, Forward- und Stepwise-Selektion auch die modernen Verfahren LASSO und LAR implementiert. Das LASSO-Verfahren (Least absolute shrinkage and selection ooperator) wurde 1995 erstmals von Tibshirani vorgestellt [6]. Die Berechnung der Regressionskoeffizienten, die auf Shrinkage-Methoden beruht, ge-

schieht über die Kleinste-Quadrate-Schätzung, jedoch mit der Bedingung, dass die Summe der absoluten Werte der Regressionskoeffizienten kleiner/gleich einem vorgegebenen Wert t ist.
Die Berechnung der LASSO-Methode führt zu einem quadratischen Programmierproblem, deren Lösung aufwendig ist. Ein effizienter Ansatz zur Berechnung des LASSO ist aber mit Hilfe der LAR-Methode möglich. Über diesen Algorithmus erhält man automatisch alle möglichen Werte von t. Dieser Ansatz, entwickelt von Efron et al. (2004) [2], wurde zur Implementierung des LASSO in der neuen Prozedur PROC GLMSELECT verwendet.
Das LAR-Verfahren hat eine gewisse Ähnlichkeit zur Forward-Selektion, nimmt jedoch im Gegensatz dazu die Variablen nicht von Anfang an mit dem vollen Wert ihrer berechneten Koeffizienten ins Modell auf. Die Berechnung erfolgt mit standardisierten Variablen. Zunächst werden alle Regressionskoeffizienten auf 0 gesetzt. Danach wird der Vektor derjenigen Variablen ausgewählt, die am Höchsten mit der Zielvariablen korreliert. Der zur ausgewählten Variable dazugehörige Koeffizient wird angepasst, bis eine weitere Variable genauso hoch mit den bisherigen Residuen korreliert wie das aktuelle Modell. Diese Variable wird dann neu in das Modell aufgenommen. Alle im Modell enthaltenen Variablen werden gleichmäßig angepasst, bis eine weitere Variable genauso hoch mit den bisherigen Residuen korreliert. Nach diesem Algorithmus wird fortgefahren, bis alle Variablen im Modell sind oder die Selektion durch das gewählte Stoppkriterium beendet wird. Der LAR-Algorithmus ist für das LASSO-Verfahren jedoch an einer Stelle modifiziert: Nimmt ein standardisierter Regressionskoeffizient während der Selektionsberechnung wieder den Wert Null an, so wird diese Variable im Gegensatz zur LAR-Methode aus dem Modell entfernt. In späteren Schritten kann sie aber wieder hinzugefügt werden.

3.3 Vielfältige Auswahl an Modellgüte-Kriterien

Die folgenden Modellgüte-Kriterien in Tabelle 1 können sowohl als Selektions-, Stopp- oder Auswahlkriterium zur Beeinflussung des Selektionsprozesses eingesetzt werden:

Tabelle 1: Modellgüte-Kriterien

ADJRSQ	Adjusted R-square statistic
AIC	Akaike Information Criterion
AICC	Corrected Akaike Information Criterion
BIC	Sawa Bayesian Information Criterion
CP	Mallow C(p) statistic
CVPRESS	Cross Validation predicted residual sum of squares statistic
PRESS	predicted residual sum of squares statistic
RSQUARE	R-square statistic
SBC	Schwarz Bayesian Information Criterion
SL	significance level of the F statistic
VALIDATE	average square error over the validation data

Selektionskriterium (SELECT=…)
Mit SELECT= kann bei den Standard-Verfahren gewählt werden, nach welchen Kriterien die einzelnen Selektionsschritte erfolgen sollen. Somit können neben dem bekannten Selektionskriterium des Signifikanzniveaus (SELECT=SL), wie bei PROC REG implementiert, auch andere Kriterien den Selektionsalgorithmus beeinflussen. Bei LASSO und LAR gibt es diese Option nicht, da der Algorithmus grundsätzlich auf einem anderen Selektionsprinzip (Korrelation der Residuen mit einer Variablen) beruht.

Stoppkriterium (STOP=...)
Bei allen Selektionsverfahren ist die Angabe eines Stoppkriteriums möglich. Dementsprechend stoppt die Selektion an dem Punkt, an dem das weitere Hinzufügen oder Entfernen einer Variablen den Wert des jeweiligen Stoppkriteriums verschlechtern würde. Ausschlag für die Beendung des Selektionsprozesses gibt also ein lokales Minimum des Kriteriumwertes.

Auswahlkriterium (CHOOSE=...)
Zusätzlich kann man über die CHOOSE-Option ein Auswahlkriterium festlegen. Damit wird aus dem gesamten durchlaufenen Selektionsprozess dasjenige Modell als Endmodell betrachtet, dessen Auswahlkriterium den besten Wert aufweist.

Die einzelnen Optionen könnten beispielsweise bei einer Backward-Selektion folgendermaßen gewählt und kombiniert werden:

```
SELECTION=BACKWARD SELECT=SL SLS=0.1 STOP=AIC CHOOSE=SBC
```

Dieses Beispiel bewirkt eine Variablenselektion basierend auf dem Signifikanzniveau, das zusätzlich mit SLS= festgelegt werden muss, falls nicht die Grenze der Standardeinstellung (0,15) gewünscht ist. Die Selektion stoppt, wenn das Entfernen einer weiteren Variablen den AIC-Wert des Modells erhöhen würde. Am Ende werden die einzelnen schrittweise erzeugten Modelle verglichen und das Modell als Endmodell gewählt, das den niedrigsten SBC-Wert aufweist. Inwieweit die Kombination verschiedener Kriterien bei den einzelnen Optionen in einem Selektionsprozess sinnvoll und empfehlenswert ist, kann hier jedoch nicht beantwortet werden.

3.4 Modellvalidierung

Unterschiedliche Prinzipien der Modellvalidierung [3] können wie oben bereits aufgelistet zur Steuerung des Selektionsprozesses eingesetzt werden. PROC GLMSECT ermöglicht sowohl Data-Splitting als auch Kreuzvalidierung. Beim Data-Splitting wird der Datensatz in einen Trainings-, Test- und Validierungs-Datensatz aufgeteilt. Die Modellschätzung wird ausschließlich mit den Trainings-Daten berechnet. Der Validierungs-Datensatz kann dabei zur Ermittlung von Selektions-, Stopp- und Auswahlkriterien dienen (SELECT=VALIDATE, STOP=VALIDATE, CHOOSE=VALIDATE). Das mit dem Trainings-Datensatz erhaltene Modell wird auf den Test-Datensatz angewendet, um festzustellen, ob es auf diese von der Regressionsschätzung unabhängigen Daten übertragbar ist.

Für die Aufteilung der Daten gibt es unterschiedliche Möglichkeiten. Die verschiedenen Datensätze können beispielsweise explizit im GLMSELECT-Statement mit DATA=, TESTDATA= und VALDATA= angegeben werden. Eine alternative Möglichkeit ist die automatische Aufteilung der Daten über das PARTITION-Statement. Darin wird prozentual angegeben, wie der Ausgangsdatensatz aufgeteilt werden soll. Folgendes Statement führt zur zufälligen Unterteilung der Daten in 50% Trainingsdaten und jeweils 25% Test- und Validierungsdaten: `PARTITION FRACTION(TEST=0.25 VALIDATE=0.25)`
Das Prinzip der Kreuzvalidierung kann ebenfalls als Selektions-, Stopp- oder Auswahlkriterium genutzt werden. Zur Durchführung der k-fold Kreuzvalidierung kann die Methode zur Aufteilung der Daten in k gleich große Teile mit der Option CVMETHODS= im Model-Statement angegeben werden. Dabei kann mit BLOCK(), SPLIT(), RANDOM() oder VARIABLE bestimmt werden, wie die Datenaufteilung erfolgen soll. Voreingestellt ist die zufällige Aufteilung mit k=5 (CVMETHODS= RANDOM(5)). Die Bezeichnung der k-fold crossvalidation für die Verwendung als SELECT-, STOP- und CHOOSE-Kriterium heißt CV.
Genauso kann als Kriterium auch die leave-one-out Kreuzvalidierung (PRESS) eingesetzt werden, bei der die Anzahl der Gruppen der Anzahl an Beobachtungen entspricht.

3.5 Grafische Darstellung des Selektionsprozesses

Die neue Prozedur ermöglicht durch besondere Grafiken eine Visualisierung des Selektionsprozess. Unter Einbindung der Prozedur in das ODS GRAPHICS-Statement können vier unterschiedliche Grafiken mit der PLOT-Option im GLMSELECT-Statement angefordert werden: CandidatesPlot, ASEPlot, CoefficientPanel und CriterionPanel. Die beiden letzteren werden im Folgenden kurz dargestellt.
Das CoefficientPanel ist in zwei Teilgrafiken unterteilt. Im oberen Teil ist die jeweilige Variable zu jedem Selektionsschritt mit dem dazugehörigen standardisierten Regressionskoeffizienten dargestellt. Die Standardisierung ermöglicht eine leichtere Beurteilung der Bedeutsamkeit der einzelnen Variablen für das Modell zu jedem Zeitpunkt des Selektionsprozesses. Es ist genau zu erkennen, in welchem Schritt welche Variable ins Modell aufgenommen wird und welche Auswirkung dies für die restlichen Variablen im Modell hat. In Abbildung 1 bei einer Forward-Selektion mit SELECT=CP STOP=AIC und CHOOSE=SBC ist beispielsweise zu erkennen, dass der Einfluss der Variablen Var2 durch die Aufnahme von Var4 im zweiten Schritt der Selektion verringert wird. (Die Aufnahme des Intercepts wird als Schritt 0 angenommen.)

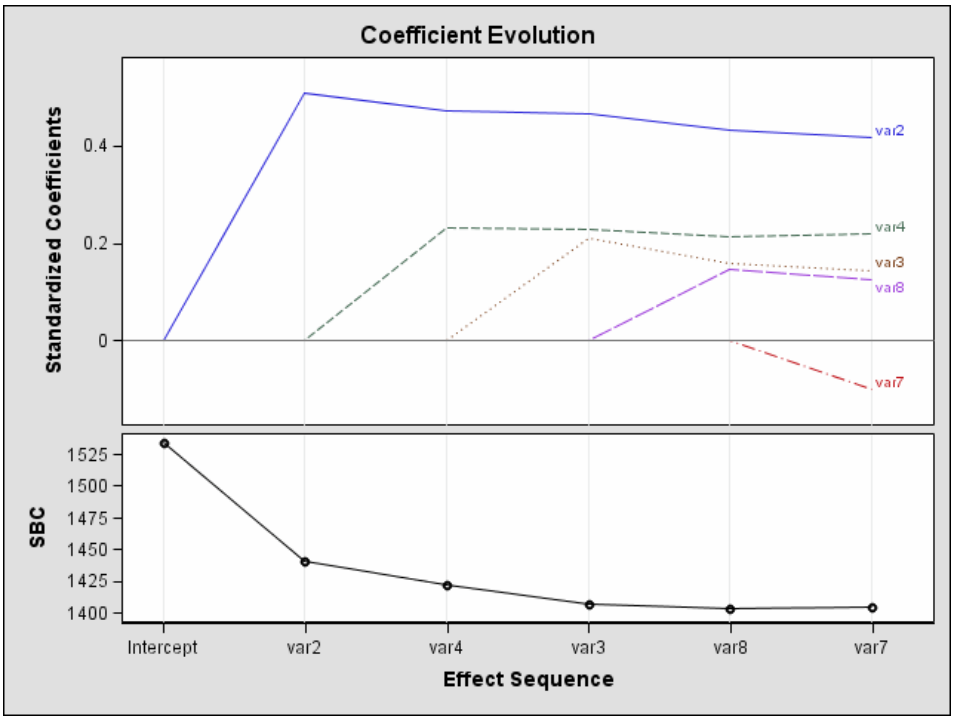

Abbildung 1: CoefficientPanel zur Darstellung der Variablen-Aufnahme zu jedem Selektionsschritt der Forward-Selektion

Der untere Teil der Grafik zeigt zu jedem Selektionsschritt den aktuellen Wert des Auswahlkriteriums an. Im Beispiel ist dies das SBC-Auswahlkriterium (CHOOSE=SBC). Außerdem wird die Variable, die in diesem Schritt in das Modell aufgenommen wurde dargestellt. Ist kein Auswahlkriterium angegeben, wird bei den Standardverfahren das Selektionskriterium, bei LAR und LASSO das Stoppkriterium (Standard: SBC) angezeigt. Nicht eindeutig zu erkennen ist, dass die Variable Var8 den geringsten SBC-Wert aufweist, was sich jedoch in einer späteren Darstellungsform noch zeigen wird. Als Endmodell wird folglich das Modell ohne die Variable Var7 gewählt, da durch deren Aufnahme der SBC-Wert des Modells schlechter werden würde.
Eine weitere Darstellungsform ist das CriterionPanel, das die Werte verschiedener Fit-Statistiken für jeden einzelnen Schritt während des Selektionsprozesses darstellt. Standardmäßig werden der AIC-, AICC-, SBC- und ADJRSQ-Wert augegeben. Werden aber in den Sub-Optionen (SELECT=, STOP=, CHOOSE=) weitere Kriterien genannt oder mit Hilfe von STATS= angefordert, wird das CriterionPanel entsprechend ergänzt. Für obiges Beispiel erhält man die in Abbildung 2 dargestellte Grafik.
Die übersichtliche Nebeneinanderstellung der unterschiedlichen Bewertungskriterien für jeden Schritt kann bei der schwierigen Auswahl der Selektions-, Stopp- und Auswahlkriterien helfen. Die Anzahl der dargestellten Selektionsschritte spiegelt den Prozess bis zum Abbruch der Selektion durch das Stoppkriterium wieder. Der Stern markiert den besten Wert des jeweiligen Kriteriums im Selektionsprozess. Im Beispiel muss dementsprechend der letzte Wert beim AIC-Kriterium als der beste identifiziert sein. Die Abbildung zeigt, dass die Selektion bei der Wahl des Stoppkriteriums SBC schon einen Schritt früher beendet gewesen wäre. Da im Beispiel jedoch die Angabe der Option

CHOOSE=SBC angegeben ist, wird genau dieses beim Selektionsschritt 4 errechnete Modell als Endmodell bestimmt.

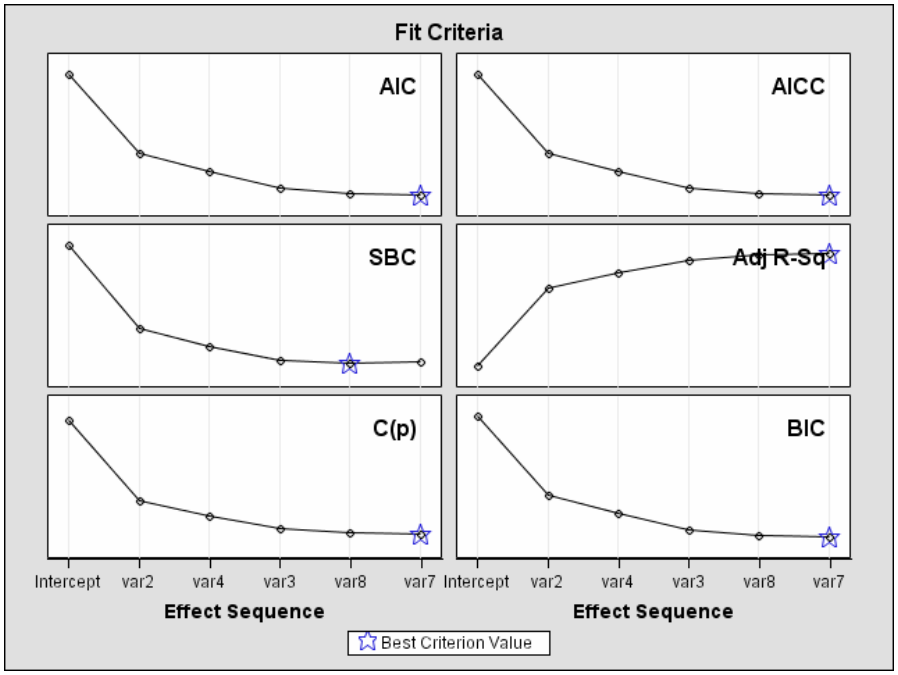

Abbildung 2: CriterionPanel zur Darstellung der Modellgüte-Kriterien

4 Vergleiche von Einstellungen an Hand eines medizinischen Beispiels

Durch die vielfältigen Möglichkeiten sowohl bei der Wahl der Selektionsmethoden als auch bei den Einstellungen von Selektions-, Stopp- und Auswahlkriterien in der Prozedur PROC GLMSELECT ergeben sich enorm viele Kombinationsmöglichkeiten, die hier nicht alle betrachtet werden können. Im Folgenden werden daher Ergebnisse an Hand einiger ausgewählter Einstellmöglichkeiten verglichen, um unterschiedliche Auswirkungen leicht verständlich an Beispielen darzustellen. Die Daten stammen aus einer Studie zur Untersuchung des frühen Auftretens von Typ-2-Diabetes und seiner Vorstufen bei übergewichtigen Kindern und Jugendlichen [7]. Die Fragestellung im hier vorgestellten Anwendungsbeispiel ist, welche Variablen einen Einfluss auf die Zielgröße Insulin haben. Es soll damit nicht die medizinisch inhaltliche Interpretation eines Modells im Vordergrund stehen, sondern ein Gefühl vermittelt werden, wie vielseitig man den Selektionsprozess mit der neuen Prozedur steuern kann.

4.1 Vergleich aller Selektionskriterien

Als erstes Beispiel soll hier ein Vergleich aller in der Prozedur zur Verfügung stehender Selektionsverfahren dargestellt werden. Um mit einem bekannten Selektionsalgorithmus zu beginnen, wird zum Vergleich der verschiedenen Verfahren als Selektionskriterium (SELECT=) die Durchführung auf Basis des Signifikanzniveaus (SL) gewählt, so wie die Selektion in PROC REG implementiert ist. Als SLS bzw. SLE wird einheitlich der

Wert 0,15 gesetzt, was auch der Standard-Einstellung in PROC GLMSELECT entspricht. Bei den Methoden LAR und LASSO, deren spezieller Algorithmus keine SELECT-Option erlaubt, entspricht die Option STOP=SL am ehesten einer Modellfindung auf Basis des Signifikanzniveaus und wird dort explizit angegeben. Das Stoppkriterium bei den anderen Verfahren entspricht ohne zusätzliche Angabe automatisch dem des Selektionskriteriums und ist daher auch einheitlich SL. Da die neuen Methoden LAR und LASSO keinen gemeinsamen Ein- bzw. Ausschluss von jeweils zusammengehörigen Dummy-Variablen ermöglichen, wird bei den Standardverfahren zu Vergleichszwecken ebenfalls die getrennte Selektion der einzelnen Dummies betrachtet.

Tabelle 2: Variablen im Endmodell bei unterschiedlichen Selektionsverfahren

Einflussgröße	Backward	Forward	Stepwise	LASSO/LAR
adiponectin	X	X	X	X
alter	X	X	X	X
crp				
geschlecht	X			
gewichtsklasse_2				X
gewichtsklasse_3				X
harnsaure	X	X	X	X
hdl				X
hypertonie	X	X	X	X
iad	X	X	X	X
ldl	X	X	X	X
leptin	X	X	X	X
steatosis_hep_1	X	X	X	X
steatosis_hep_2	X	X	X	X
steatosis_hep_3	X	X	X	X
testosteron	X			
triglyzeride	X	X	X	X
whr				
Gütekriterien				
p-Wert	<0,0001	<0,0001	<0,0001	<0,0001
R-Square	0,4283	0,4234	0,4234	0,4266
ADJRSQ	0,4112	0,4089	0,4089	0,4081
AIC	1843,29320	1843,13562	1843,13562	1846,66209
SBC	1900,79152	1892,41990	1892,41990	1908,26743
CP	10,69251	10,40266	10,40266	14,01064

Tabelle 2 gibt einen Überblick, welche Variablen bei den unterschiedlichen Verfahren im Endmodell enthalten sind. Dabei ist zu sehen, dass sich die Standard-Verfahren kaum unterscheiden. Nur bei der Backward-Selektion werden zwei zusätzliche Variablen, *geschlecht* und *testosteron*, mit ins Modell aufgenommen.
Am meisten Variablen enthält das Endmodell bei den Methoden LAR und LASSO, die im Gegensatz zu den Standard-Verfahren auch die Variablen *gewichtsklasse* und *hdl* im Endmodell behalten. 13 der insgesamt 18 Variablen werden von allen hier gezeigten

Methoden bezüglich Aufnahme und Ausschluss gleich behandelt. Das Ergebnis von LAR und LASSO ist identisch, was zeigt, dass mit dem Datensatz beim vorliegenden Beispiel kein Regressionskoeffizient während der Berechnung wieder den Wert Null annimmt, was den entscheidenden Unterschied zwischen den Verfahren ausmacht (siehe Abschnitt 3.2). Die Parameterschätzer bei LAR und LASSO sind durch das Shrinkage kleiner und weisen eine geringere Varianz auf, was in diesem Beispiel eher die Aufnahme von mehr Variablen ins Endmodell begünstigt. Die große Anzahl von Variablen im Endmodell erklärt die Gütekriterien, die bei LAR und LASSO bis auf R^2 allesamt etwas schlechter ausfallen als bei den anderen Selektionsverfahren. Diese Abweichungen liegen jedoch bis auf CP im Promille-Bereich.
Abschließend kann man zu diesem Vergleich sagen, dass die neuen Methoden LASSO und LAR bei Durchführung der Selektion bis zu einem Stoppkriterium auf Basis des Signifikanzniveaus mehr Variablen ins Modell aufnehmen als die Standardverfahren, wie sie beispielsweise in PROC REG implementiert sind.

4.2 Vergleich an Hand verschiedener Auswahlkriterien

Ein weiterer Vergleich soll nun die Unterschiede bei der Modellwahl an Hand unterschiedlicher Einstellungen bei der CHOOSE-Option aufzeigen. Der Vergleich der Modellgüte-Kriterien als Auswahlkriterium soll mit der Backward-Selektion erfolgen.

Tabelle 3: Auflistung der Variablen im Endmodell bei unterschiedlichen Auswahlkriterien beim Backward-Verfahren mit Angabe des gewählten Selektionsschrittes

Einflussgröße	ADJRSQ	AIC	AICC	BIC	CP	CV	SBC
adiponectin	X	X	X	X	X	X	X
alter	X	X	X	X	X	X	X
crp							
geschlecht	X	X					
gewichtsklasse_2	X						
gewichtsklasse_3							
harnsaure	X	X	X	X	X	X	X
hdl							
hypertonie	X	X					
iad	X	X	X	X	X		
ldl	X	X	X	X	X		
leptin	X	X	X	X	X	X	X
steatosis_hep_1	X	X	X	X	X	X	X
steatosis_hep_2	X	X	X	X	X	X	X
steatosis_hep_3	X	X	X	X	X		X
testosteron	X	X					
triglyzeride	X	X	X	X	X	X	X
whr							
Selektionsschritt	**4**	**5**	**8**	**8**	**8**	**11**	**10**

Der Vergleich wurde unter Angabe der Optionen SELECTION=BACKWARD SELECT=SL STOP=NONE und unterschiedlicher Auswahlkriterien (ADJRSQ, AIC, AICC, BIC, CP, CV, SBC) bei der CHOOSE-Option (CHOOSE=) durchgeführt. Dies bedeutet, dass der Selektionsprozess jeweils bis zum Ende durchgeführt wird, also bis alle Variablen im Modell aufgenommen sind. Erst dann wird aus den Modellen der einzelnen Selektionsschritte dasjenige Modell ermittelt, das jeweils gemäß der CHOOSE-Option als das Beste erkannt wird, was in Tabelle 3 dargestellt ist. Die Zahl des gewählten Selektionsschrittes gibt die Anzahl der entfernten Variablen an.
Bei ADJRSQ und AIC als Auswahlkriterium befinden sich mehr Variablen im Endmodell als bei AICC, BIC und CP. Der Einsatz der Kreuzvalidierung (CV), die auf mehrfacher Wiederholung der Modellschätzung auf unterschiedlichen Teildatensätzen basiert, führt zum kleinsten Modell. Das SBC-Kriterium kann zudem als „strengster" Maßstab unter den Kriterien mit Strafterm betrachtet werden.

4.3 Vergleich von SBC bei STOP- und CHOOSE-Option

Nach der Betrachtung der unterschiedlich gewählten Endmodelle mittels der CHOOSE-Option, soll im nächsten Vergleich die Auswirkung eines Gütekriteriums dargestellt werden, je nachdem ob es als Stopp- oder Auswahlkriterium eingesetzt wird. Dafür wurde das neue LASSO-Verfahren gewählt und der Vergleich an Hand des relativ strengen SBC-Kriteriums durchgeführt, das bei der Prozedur PROC GLMSELECT als Standardeinstellung bei der STOP-Option implementiert ist. Für den ersten Durchlauf der Prozedur wurden im MODEL-Statement die Optionen STOP=NONE und CHOOSE=SBC angegeben und für einen weiteren Durchlauf die Option STOP=SBC.

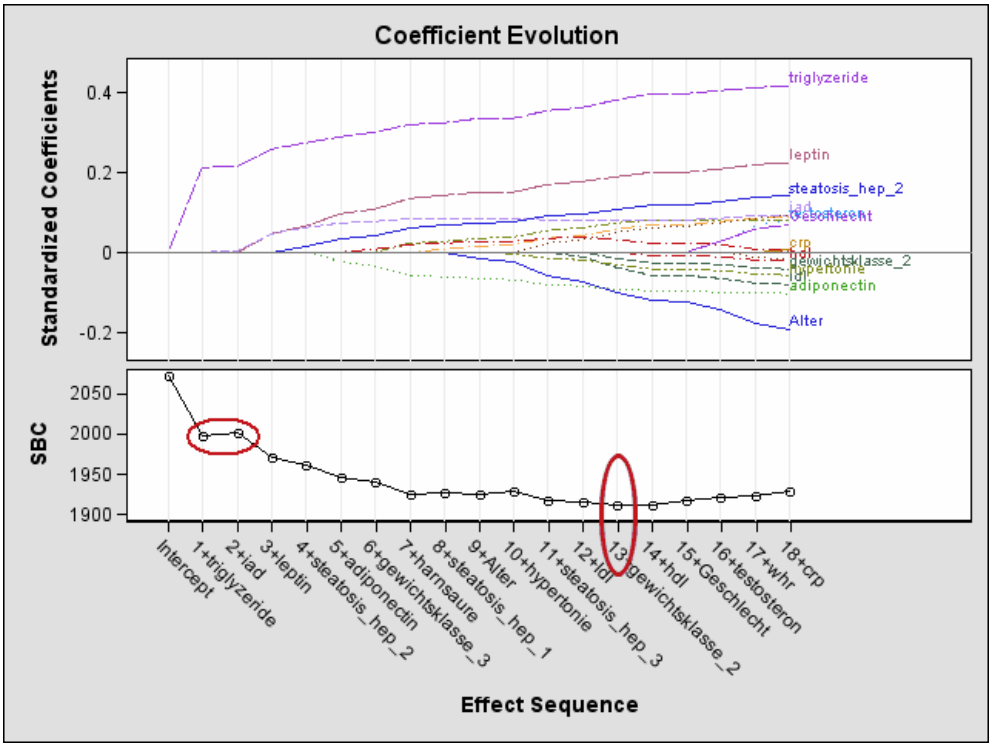

Abbildung 3: CoefficientPanel für die LASSO-Selektion mit STOP=NONE und CHOOSE=SBC

Die Ergebnisse dieser unterschiedlichen Verwendung des SBC-Kriteriums weichen stark voneinander ab. In Abbildung 3 ist der Selektionsprozess für den ersten Fall grafisch dargestellt. Nach der Berechnung aller Regressionskoeffizienten (STOP= NONE) wird am Ende das beste Modell bezüglich des SBC-Kriteriums ausgewählt. Das globale Minimum ist in diesem Fall das Modell zum Schritt 13, da dieses Modell den geringsten SBC-Wert aufweist.
Setzt man hingegen das SBC-Kriterium bei der STOP-Option ein, ist der Selektionsprozess schon nach dem ersten Schritt beendet und folglich befindet sich nur eine Variable im erzielten Modell. Das CoefficientPanel für diesen Durchlauf der Prozedur ist in Abbildung 4 dargestellt.

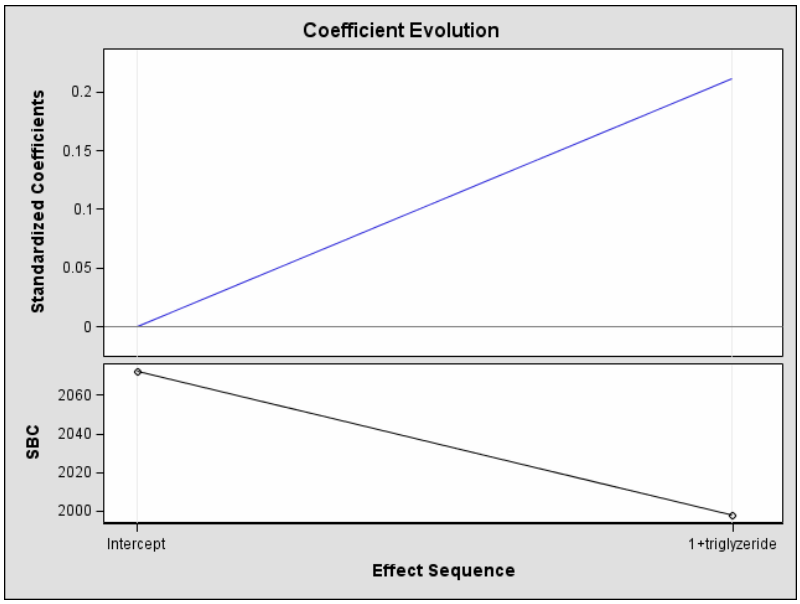

Abbildung 4: CoefficientPanel für die LASSO-Selektion mit STOP=SBC

Die Variable *triglyzeride* ist als einzige Variable im Endmodell, da die Hinzunahme einer weiteren Variablen den SBC-Wert des Modells verschlechtern würde. Betrachtet man nochmals Abbildung 3 für den ersten Durchlauf, kann man diese Tatsache an der grafischen Darstellung des SBC-Wertes ablesen. Man kann sehen, dass im ersten Schritt durch die Aufnahme der Variable *triglyzeride* der SBC-Wert für das Modell sinkt. Die Hinzunahme der Variablen *iad* im zweiten Schritt führt jedoch zu einer Erhöhung und damit Verschlechterung des SBC-Werts. Dieses lokale Minimum ist Auslöser, den Selektionsprozess an dieser Stelle zu beenden.
Prinzipiell kann man sagen, dass der Einsatz eines Gütekriteriums als STOP-Option eher ein kleineres Modell zur Folge hat als beim Einsatz bei der CHOOSE-Option, da der Selektionsprozess beendet wird, sobald das Kriterium, wenn auch nur kurzzeitig, einen schlechteren Wert annimmt. Es zählt also nur das lokale Minimum, wohingegen beim CHOOSE-Kriterium die Werte des gesamten Selektionsprozesses und somit das globale Minimum betrachtet wird. Um den gesamten Selektionsprozess verfolgen und analysieren zu können, empfiehlt sich daher eher, die Selektion mit STOP=NONE bis zum Ende durchlaufen zu lassen und dann mittels eines Auswahlkriteriums das Endmodell zu bestimmen.

5 Fazit

Die neue Prozedur bietet bezüglich der Variablenselektion vielfältigere Möglichkeiten als die herkömmlichen Prozeduren wie beispielsweise PROC REG. Genannt seien hier nochmals die Aufnahme der Selektionsmethoden LASSO und LAR, die in der Literatur immer mehr Einzug halten sowie die Auswahl unterschiedlicher Gütekriterien, die an verschiedenen Stellen im Selektionsprozess für Entscheidungen eingesetzt werden können. Als besonders hervorzuheben sind bei der neuen Prozedur die grafischen Darstellungsmöglichkeiten des Selektionsprozesses. Vor allem mit den Coefficient- und CriterionPanels wurde eine sehr schöne Visualisierung des Vorgehens gegeben.
Neben den genannten Vorteilen hat die Prozedur jedoch auch Nachteile, die sich vor allem bei der Anwendung der neuen Verfahren LASSO und LAR ergeben. Einige aufgeführte Funktionalitäten von PROC GLMSELECT sind gerade mit diesen Algorithmen nicht möglich. Als ein großes Manko kann dabei die fehlende blockweise Behandlung von Dummy-Variablen betrachtet werden. Unter anderem kann auch das Hierarchieprinzip, das beim Einschluss von Wechselwirkungen von Bedeutung ist, bei den neuen Verfahren nicht berücksichtigt werden, genauso wie die INCLUDE-Option, mit der Variablen festgelegt werden können, die auf alle Fälle ins Endmodell aufgenommen werden sollen.
Abschließend kann man die Prozedur PROC GLMSELECT trotz der aufgeführten Mängel und Probleme dennoch als sehr hilfreich und als komfortables Werkzeug für die Modellfindung mittels Variablenselektion im linearen Regressionsmodell betrachten. Sie darf aber nicht als Ersatz für die bekannten Prozeduren angesehen werden, die weiterhin für notwendige zusätzliche Untersuchungen (Multikollinearitätsdiagnostik, Ermittlung einflussreicher Beobachtungen, Residualanalyse u.s.w.) herangezogen werden müssen. Die neue Prozedur sollte durch ihre vielen neuen Möglichkeiten nicht dazu verleiten, Einstellungen unreflektiert vorzunehmen und die Ergebnisse ohne kritische Überprüfung zu verwenden. Trotz der vielen hilfreichen Automatismen bleibt die Variablenselektion ein manuell durchzuführender Prozess, der viel Erfahrung und Hintergrundwissen erfordert. Es ist aber sicherlich interessant, den Einsatz dieser Prozedur bei der Unterstützung dieser Aufgabe in der praktischen Anwendung zu verfolgen.

Literatur

[1] Cohen R. A.: *Introducing the GLMSELECT PROCEDURE forModel Selection.* SUGI 31 Proceedings, Paper 207-31, 2006.

[2] Efron, B; Hastie, T.; Johnstone, I. et al.: Least Angle Regression. In: *Annals of Statistics* Bd. 32, 2004, S. 407-499

[3] Harrell, F.E.: *Regression Modeling Strategies.* New York: Springer, 2001

[4] Muche, R.: *Variablenselektion in Kohortenstudien.* Universität Ulm: Dissertation zum Dr. hum. biol., 1995.

[5] *The GLMSELECT Procedure (Experimental).* SAS Inst. Inc., Cary, NC, USA, 2006, S. 1-102. – Download: *http://support.sas.com/rnd/app/da/glmselect.html*

[6] Tibshirani, R.: Regression Shrinkage and Selection via the LASSO. In: *Journal of the Royal Statistical Society* Bd. 58, 1996 (B (Methodological)), S. 267-288

[7] Wabitsch M., Hauner H., Hertrampf M., Muche R., Hay B., Mayer H., Kratzer W., Debatin K-M., Heinze E. (2004): Type 2 diabetes mellitus and impaired glucose regulation in Caucasian children and adolescents with obesity living in Germany. In: *Int J Obes Relat Metab Disord* Bd. 28, 2004, S. 307-313