Elastic Net und Lasso: lassen Sie in unübersichtlichen Situationen Software statistische Modelle finden

Bernd Heinen
SAS Institute GmbH
In der Neckarhelle 168
Heidelberg
Bernd.heinen@jmp.com

Zusammenfassung

Viele Entwicklungen haben dazu geführt, dass immer häufiger Datensätze analysiert werden müssen, die nicht gut strukturiert sind. Oft steht man vor der Aufgabe, aus vielen Variablen diejenigen herauszufinden, die für die Anpassung eines Modells überhaupt benötigt werden. Korrelationen zwischen den Faktoren erschweren die Modellbildung zusätzlich. Außerdem ist nicht selten die Zahl der Variablen größer als die Zahl der Beobachtungen. In diesen Situationen sind Elastic Net und Lasso effiziente Verfahren, um Modelle zu bilden und zu vergleichen. Beispielanwendungen mit JMP Pro zeigen die Einsatzbereiche und Effizienz dieser Verfahren.

Schlüsselwörter: Lineares Modell, Variablenauswahl, Elastic Net, Lasso, JMP Pro

1 Einleitung

Die umfassende digitale Abwicklung und Steuerung vieler Prozesse, automatisierte Analyseverfahren und preiswerte Datenspeicher, die das Sammeln all dieser Daten überhaupt erst ermöglichen, führen dazu, dass für Datenanalysen häufig umfangreiche Datenmengen zur Verfügung stehen. Wenn diese aus unterschiedlichen Quellen kombiniert wurden, entstehen schnell auch schlecht strukturierte Datensätze, d.h. solche mit hoher Korrelation unter den Variablen und vielen fehlenden Werten. In den meisten Fällen ist man bei einer Analyse ja nicht nur an einer univariaten Beschreibung der einzelnen Variablen interessiert, sondern eher an der Bildung von Modellen, die Zusammenhänge zwischen den Variablen analysieren lassen, und die für Prognosen oder zur Optimierung genutzt werden können. Streng genommen erfordern diese Bedingungen eine sorgfältige Vorverarbeitung der Daten, Bereinigung, Transformation, eventuell das Ersetzen fehlender Werte. Stehen sehr viele Variablen zur Verfügung eventuell sogar mehr als Beobachtungen, steht die Auswahl der relevanten Variablen an erster Stelle. Klassische lineare Modelle können in dieser Situation gar nicht angewandt werden. Im Folgenden werden mit LASSO und Elastic Net Verfahren zur Bildung linearer Modelle beschrieben, die robuste und zuverlässige Modelle erstellen, teilweise auch eine Variablenauswahl vornehmen und selbst bei „unsauberen“ Datensätzen anwendbar sind.

2 Herleitung

Im linearen Modell y = Xβ +e wird üblicherweise der kleinste Quadrate Schätzer ($\hat{\beta}_{KQ}$) zur Schätzung der Parameter herangezogen. Bekanntermaßen ist er der beste lineare erwartungstreue Schätzer, aber er hat auch die Eigenschaft, die Parameter tendenziell zu groß zu schätzen, d.h. es gibt immer Schätzer $\hat{\beta}$, deren euklidische Norm kleiner als die des kleinste Quadrate Schätzers ist: $\|\hat{\beta}_{KQ}\| \geq \|\hat{\beta}\|$. Wenn man außerdem nicht erwartungstreue Schätzer zulässt, lässt sich die mittlere quadratische Abweichung weiter verkleinern, so dass trotz Bias eine bessere Prognose möglich ist. Der kleinste Quadrate Schätzer liefert einen Wert für jeden Modellparameter, er führt keine Variablenselektion durch. Die Schätzung erfolgt durch Minimierung der Summe der kleinsten Quadrate: $\hat{\beta}_{KQ} = \arg\min_{\beta}\{|y - X\beta|^2\}$.

Um die angesprochenen Schwächen des KQ Schätzers zu beheben, führt man Nebenbedingungen für β in diese Minimierungsaufgabe ein: $\hat{\beta} = \arg\min_{\beta}\{|y - X\beta|^2 + L(\lambda)\}$. Generell spricht man bei diesen Verfahren von pönalisierter Regression. Die hier beschriebenen Schätzverfahren unterscheiden sich in der Wahl der Nebenbedingung, sie lauten

$$\hat{\beta}_{Lasso} = \arg\min_{\beta}\left\{|y - X\beta|^2 + \lambda \sum_{j=1}^{p}|\beta_j|\right\}$$

$$\hat{\beta}_{Elastic\ Net} = \arg\min_{\beta}\left\{|y - X\beta|^2 + \lambda_1 \sum_{j=1}^{p}|\beta_j| + \lambda_2 \sum_{j=1}^{p}{\beta_j}^2\right\}$$

Die Nebenbedingung für das Lasso Verfahren wird L_1 Pönalisierung genannt. Für das elastische Netz wird dieselbe Nebenbedingung genutzt wie für Lasso und ein weiterer quadratischer Term hinzugefügt. Die Lösungen beider Verfahren setzen keinen vollen Rang von X voraus, sind also auch anwendbar, wenn korrelierte Variablen in das Modell eingehen. In beiden Fällen gibt es keine geschlossene Darstellung der Lösung, die Optimierungsaufgaben werden iterativ gelöst. Zur Bestimmung der λ Parameter wird eine Suche über ein feines Raster durchgeführt und die Wahl durch das Validierungsverfahren getroffen. Die Lösung der Minimierungsaufgabe bedeutet, dass die Schätzer selbst klein werden müssen, gegenüber dem kleinste Quadrate Schätzer werden sie also geschrumpft, woraus sich auch der englische Oberbegriff Shrinkage für diese Verfahren ableitet.

3 LASSO

Bei Betrachtung der Nebenbedingung für die LASSO Schätzung wird auch die Herkunft der Bezeichnung **L**east **A**bsolute **S**hrinkage and **S**election **O**perator intuitiv klar [1].

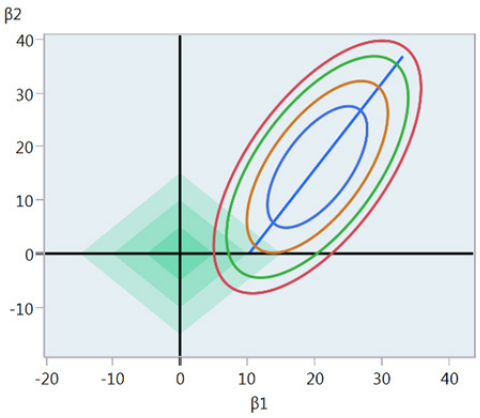

Abbildung 1: Minimierung unter Nebenbedingungen

Für p=2, d.h. $\beta = (\beta_1, \beta_2)^T$ lassen sich Nebenbedingung und Schätzverfahren anschaulich darstellen. RSS(β), die Residuenquadratsumme, stellt für jede Konstante c mit RSS(β) = c eine Ellipse dar. Die abgestuften Quadrate zeigen die Nebenbedingungen für verschiedene λ. Die Lösung des Schätzproblems ist dort, wo eine Ellipse zum ersten Mal auf eine Kante der Nebenbedingung trifft. Fällt der Schnittpunkt auf eine Ecke, wird der entsprechende Koeffizient mit Null geschätzt, im übertragenen Sinn aus dem Modell entfernt.

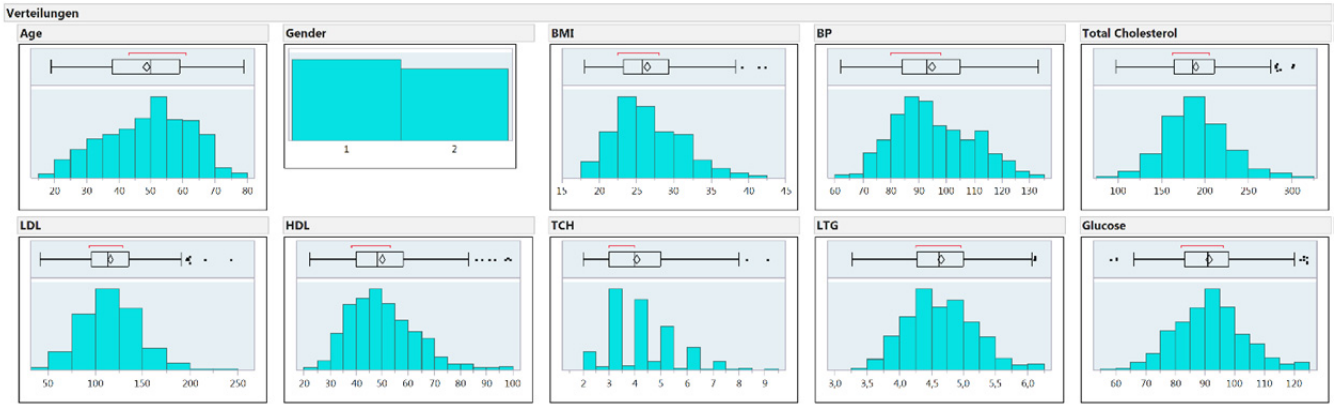

Abbildung 2: Verteilung der Variablen aus dem Beispieldatensatz

Zur Veranschaulichung der Anwendung habe ich einen bekannten Datensatz [2] über die Entwicklung von Diabeteserkrankungen nach einem Jahr herangezogen. Die unabhängigen Variablen zeigen verschiedene Verteilungen, die ohne weitere Transformation in die Analysen eingehen.

Der JMP Dialog für „Modell anpassen“ wird einfach mit Zielgröße und interessierenden Variablen ausgeführt und die „Verallgemeinerte Regression“ gestartet (nur in JMP Pro verfügbar). Ein Auswahlfenster lässt wiederholt die Art der pönalisierten Regression auswählen und die Validierungsmethode bestimmen. Damit kann man unter denselben Bedingungen mehrere Regressionsverfahren anwenden und deren Ergebnisse unmittelbar vergleichen. Das Ergebnis wird jeweils in einer Grafik und in Tabellen angezeigt.

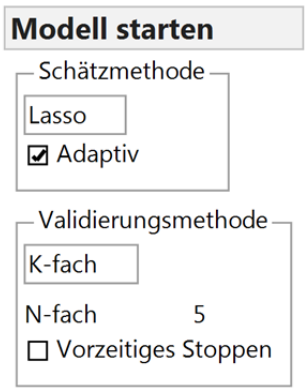

Abbildung 3: Methodendialog in JMP Pro

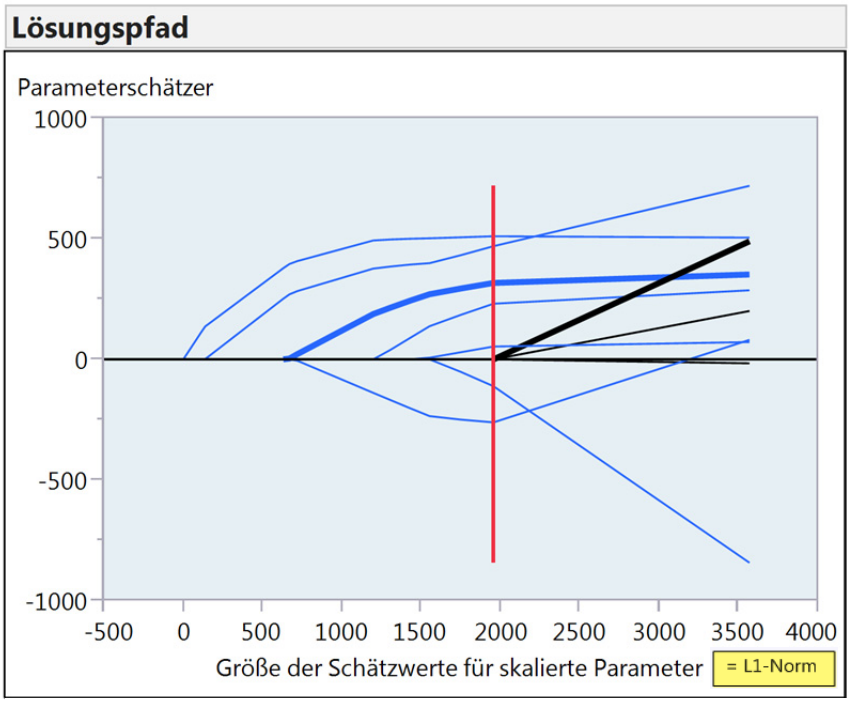

Abbildung 4: Grafische Ergebnisdarstellung LASSO

Die Grafik zeigt, wie im Verlauf des Schätzprozesses die einzelnen Parameterschätzer schrumpfen (Y-Achse). Auf der X-Achse ist die L1 Norm aufgetragen, d.h. die Summe der absoluten Parameterschätzer. Die rote Linie markiert die Lösung für das Modell.
Die Tabelle enthält die entsprechenden Parameter, Teststatistiken und Konfidenzintervalle. Wählt man Terme der Tabelle aus, werden die zugehörigen Pfade in der Grafik hervorgehoben und umgekehrt.

Parameterschätzwerte für zentrierte und skalierte Prädiktoren

Term	Schätzer	Std.-Fehler	Wald-Chi-Quadrat	Wahrsch. > Chi-Quadrat	Untere Grenze 95%	Obere Grenze 95%
Achsenabschnitt	152,13	2,56	3538,2	<,0001*	147,12	157,15
Age	0,00	0,00	0,00	1,0000	0,00	0,00
Gender[1]	185,24	58,53	10,02	0,0016*	70,52	299,96
BMI	520,78	68,30	58,13	<,0001*	386,91	654,65
BP	290,64	65,57	19,64	<,0001*	162,12	419,17
Total Cholesterol	-88,78	70,99	1,56	0,2111	-227,9	50,35
LDL	0,00	0,00	0,00	1,0000	0,00	0,00
HDL	-219,9	65,85	11,15	0,0008*	-348,9	-90,78
TCH	0,00	0,00	0,00	1,0000	0,00	0,00
LTG	505,65	81,68	38,33	<,0001*	345,57	665,73
Glucose	48,54	63,31	0,59	0,4433	-75,55	172,63
Skala	53,77	1,64	1077,4	<,0001*	50,56	56,98

Abbildung 5: Schätzer und Statistiken des LASSO

Gleichzeitig werden die Terme in allen anderen Modellen markiert, die in demselben Analysefenster angepasst wurden, sowie in den zugehörigen Spalten der Datentabelle. Das erleichtert den Vergleich der Terme über verschiedene Modelle hinweg ebenso wie die weitere Bearbeitung dieser Variablen in anderen Plattformen. Die Schätzer für Age, LDL und TCH werden hier auf Null geschrumpft, d.h. diese Variablen nicht in das Modell aufgenommen, Ergebnis der Variablenselektion, die LASSO auch liefert.

4 Elastic Net

Die Gleichungen in Absatz zwei lassen leicht erahnen, dass man verschiedene Nebenbedingungen formulieren kann, um damit Schätzer unterschiedlicher Güte und Qualität zu erhalten. Die Nebenbedingung für das LASSO Verfahren kann man mit guter Berechtigung durch den Term $\lambda \sum_{j=1}^{p} \beta_j^2$ ersetzen. Das ist als Ridge-Regression bekannt, die oftmals besser angepasste Modelle liefert, aber keine Variablenauswahl trifft und daher nicht Gegenstand dieses Vortrags ist. Das elastische Netz [3] kombiniert beide Bedingungen und Qualitäten. Die Durchführung in JMP und der Ergebnisbericht sind gleich wie beim LASSO Verfahren. Die Ergebnisse für die ausgewählten Daten sind auch ähnlich.

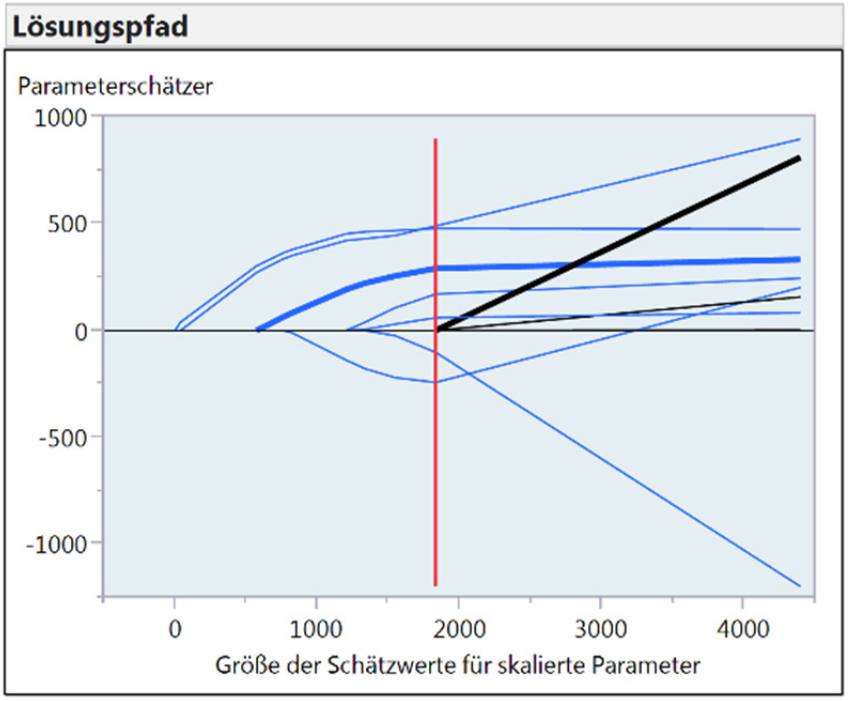

Abbildung 6: Grafische Ergebnisdarstellung Elastic Net

Parameterschätzwerte für zentrierte und skalierte Prädiktoren

Term	Schätzer	Std.-Fehler	Wald-Chi-Quadrat	Wahrsch. > Chi-Quadrat	Untere Grenze 95%	Obere Grenze 95%
Achsenabschnitt	152,71	2,90	2780,98	<,0001*	147,03	158,38
Age	5,41	62,83	0,01	0,9314	-117,74	128,57
Gender[1]	191,30	66,80	8,20	0,0042*	60,39	322,22
BMI	435,08	77,24	31,73	<,0001*	283,69	586,46
BP	308,73	80,16	14,83	0,0001*	151,62	465,84
Total Cholesterol	-103,93	79,89	1,69	0,1933	-260,51	52,65
LDL	0,00	0,00	0,00	1,0000	0,00	0,00
HDL	-249,66	78,64	10,08	0,0015*	-403,80	-95,52
TCH	0,00	0,00	0,00	1,0000	0,00	0,00
LTG	541,72	89,52	36,62	<,0001*	366,26	717,17
Glucose	69,01	68,54	1,01	0,3140	-65,32	203,35
Skala	54,08	1,83	872,57	<,0001*	50,49	57,67

Abbildung 7: Schätzer und Statistiken des Elastic Net

5 Vergleich

In diesem Fall hat LASSO einen Parameter mehr ins Modell aufgenommen als das Elas-

LASSO

Maß	Training	Validierung
Zeilenanzahl	354	88
Summe der Häufigkeiten	354	88
-LogLikelihood	1903,8746	484,77492
BIC	3860,5729	1009,8459
AIC	3825,7493	987,54985

Elastisches Netz

Maß	Training	Validierung
Zeilenanzahl	353	89
Summe der Häufigkeiten	353	89
-LogLikelihood	1901,4007	487,83969
BIC	3861,4661	1020,5657
AIC	3822,8014	995,67937

Maximum Likelihood

Maß	Training
Zeilenanzahl	442
Summe der Häufigkeiten	442
-LogLikelihood	2385,9929
BIC	4845,0814
AIC	4795,9857

tische Netz. Die Modellanpassungen beider Modelle sind gleich gut und beide Modelle sind deutlich besser angepasst als der Maximum Likelihood Schätzer, der von einem Validierungsverfahren nicht profitieren würde und daher aus allen Daten berechnet wird.

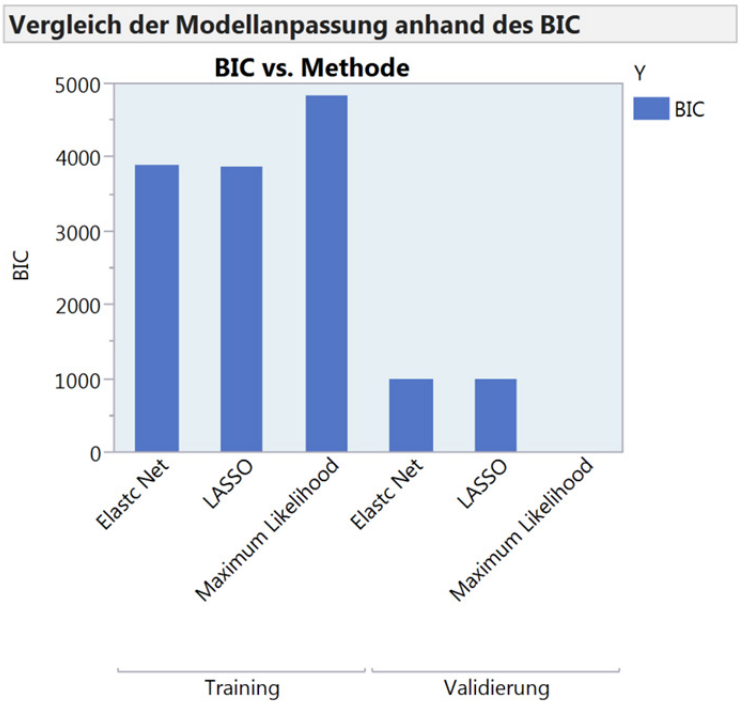

Abbildung 8: BIC pro Methode

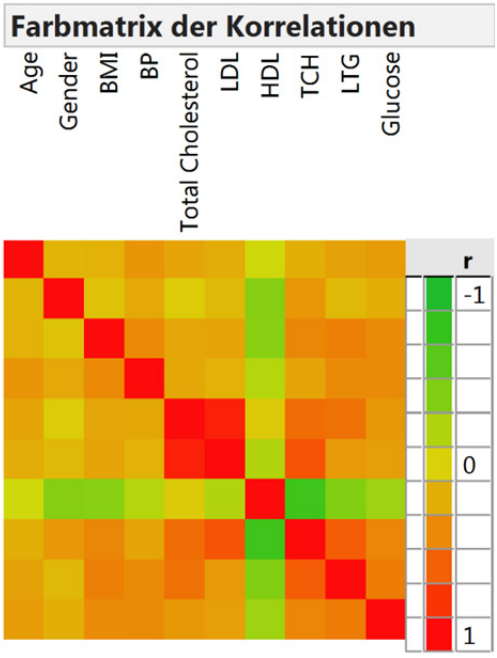

Abbildung 9: Korrelationen der Variablen aus dem Beispieldatensatz

Die Parameterschätzer unterscheiden sich ebenfalls deutlich von denen des ML-Verfahrens, einige unterscheiden sich auch zwischen LASSO und elastischem Netz, wobei die Schätzer des letztgenannten in der Tendenz eher kleiner sind.

Die Korrelationsstruktur der Variablen ist relativ einfach und verursacht daher keine großen Unterschiede bei beiden Verfahren. Generell tendiert allerdings das LASSO Verfahren dazu, bei Gruppen hochkorrelierter Variablen eine Variable auszuwählen und den Rest auszuschließen.

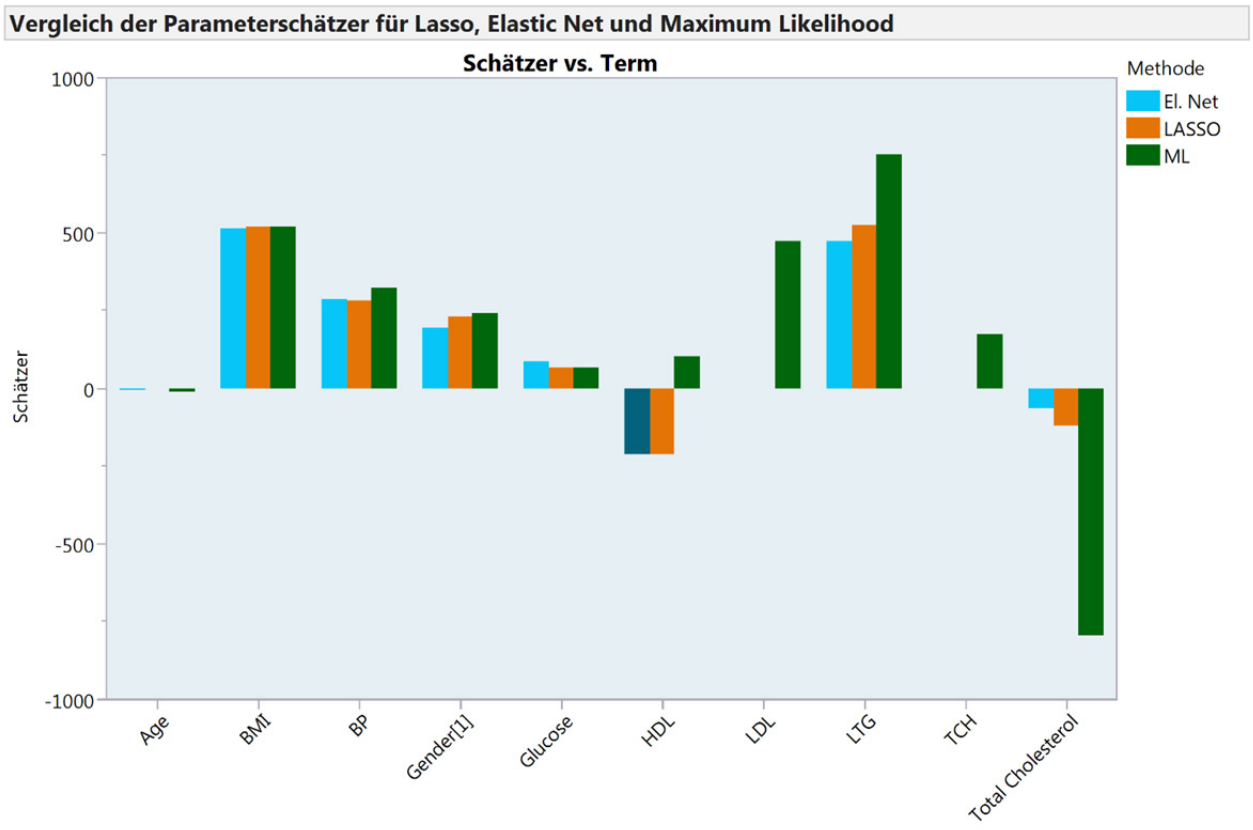

Abbildung 10: Parameterschätzer pro Schätzverfahren

Das elastische Netz neigt in dieser Situation eher dazu, alle Variablen der Gruppe ein- oder auszuschließen. Für den Fall, dass p, die Zahl der Variablen, größer ist als n, die Zahl der Beobachtungen, kann LASSO höchstens n Variablen auswählen, das elastische Netz kann auch p > n Variablen auswählen.

6 Fazit

Lineare Modelle erfreuen sich vor allem deshalb großer Beliebtheit, weil sie leicht nachvollziehbare und gut zu interpretierende Ergebnisse liefern. Allerdings stellen sie auch hohe Anforderungen an die Daten bezüglich Vollständigkeit, Unabhängigkeit und Umfang. Sind diese Voraussetzungen verletzt, insbesondere wenn p>n ist, bieten die vorgestellten Verfahren einen effizienten Weg, wesentliche von unwesentlichen Variablen zu trennen und möglichst einfache Schätzer zu finden. Das alles basierend auf linearen Modellen, sodass die leichte Verständlichkeit und Interpretierbarkeit erhalten bleiben obwohl der Anwendungsbereich deutlich erweitert ist. Es gibt auch andere Verfahren zur Variablenselektion, die aber entweder auch unter den Beschränkungen des linearen Modells und des Kleinste Quadrate Schätzers leiden oder aber rein datengetrieben sind, d.h. ohne Modellannahme auskommen. Die Ergebnisse lassen sich daher nicht inhaltlich interpretieren, was oft von Nachteil ist. Hier bieten pönalisierte Regressionsverfahren eine attraktive Alternative.

Literatur

[1] Jona Cederbaum, Handout zum Seminarvortrag Der LASSO-Schätzer, http://www.statistik.lmu.de/institut/lehrstuhl/semwiso/seminare/modellwahl_modelldiagnose_SoSe09/downloads/CEDERBAUM-Handout.pdf

[2] A. Hoerl, R. Kennard: Ridge regression: biased estimation for nonorthogonal problems. Technometrics, Vol. 12:55-67 (1970).

[3] Hui Zou and Trevor Hastie, Regularization and Variable Selection via the Elastic Net, Department of Statistics, Stanford University. December 5, 2003, http://www.stanford.edu/~hastie/Papers/elasticnet.pdf