Totgesagte leben länger - die lineare binäre Datamining-Regression im Lichte der Heidelberger Schlaganfallstudie von 1994

Hans-Peter Höschel

FairIsaac INFORMA Unternehmensberatung, Berliner Strasse 207-211, 65205 Wiesbaden
Telefon: 0611 / 978525 bzw. 0172 / 7247965
eMail: HHoeschel@informa.de

Abstract

Die fast 400 Datensätze der Studie [1] sind ein interessantes Analyseobjekt. Denn einerseits kann jeder Zuhörer seine persönlichen Risikofaktoren nach dem Vortrag etwas genauer einschätzen, insbesondere was Bluthochdruck, Rauchen und Saufen anbelangt.

Andererseits soll mit dem kleinen Analyse-Experiment verdeutlicht werden, dass die in der Praxis beliebten linearen Verfahren auch für binäre Zielvariable besser sind, als man nach der Theorie erwarten könnte. Denn tragende Voraussetzungen des linearen Standardmodells sind von vornherein nicht erfüllt, wenn die Zielvariable dichotom ist und die Einflussvariablen kategorial sind, und so findet man dazu in den Lehrbüchern nur logistische Regressionen.

Die Ergebnisse des Analyse-Experiments stimmen gut mit denen aus der fachlich korrekten bedingten logistischen Regression überein (vgl. [1]). Nebenbei werden selbst bei diesem relativ einfachen Datensatz Arbeitstechniken des Datamining erkennbar, wie sie im Prinzip auch bei grossen Datenbanken mit Millionen von Kunden im Direktmarketing und der Bonitätsprüfung üblich sind. Dazu gehört insbesondere, dass die End-Ergebnisse auch Abnehmern vermittelbar sind, die mit der Verkettung von Odds-Ratios ihre Schwierigkeiten hätten.

Es werden Vorteile und Grenzen beim Einsatz der Automatischen Response Analyse (ARA) diskutiert. Das Verfahren kann kurz als diskretisierte Dataminig Regression für binäre Zielvariablen mit vollständiger Wechselwirkungsanalyse charakterisiert werden. "Datamining" heisst dabei erstens, dass mit Trainings- und Testdateien gearbeitet wird. "Diskretisiert" heisst dabei, dass numerische Variablen automatisch diskretisiert werden, und ebenso wie die Merkmalsausprägungen der kategorialen Variablen in 0-1-Variable überführt werden. Gegebenenfalls werden auch sämtliche Wechselwirkungsterme erzeugt und "0" als Sondergruppe erfasst. "Datamining" heisst dabei zweitens, dass aus bis zu mehreren tausend Dummy-Variablen mit Verfahren der automatischen Modellwahl für lineare oder logistische Regression "signifikante" Prognose-Variable ausgewählt werden. Für numerische Variablen können damit auch nichtlineare Effekte erfasst werden.

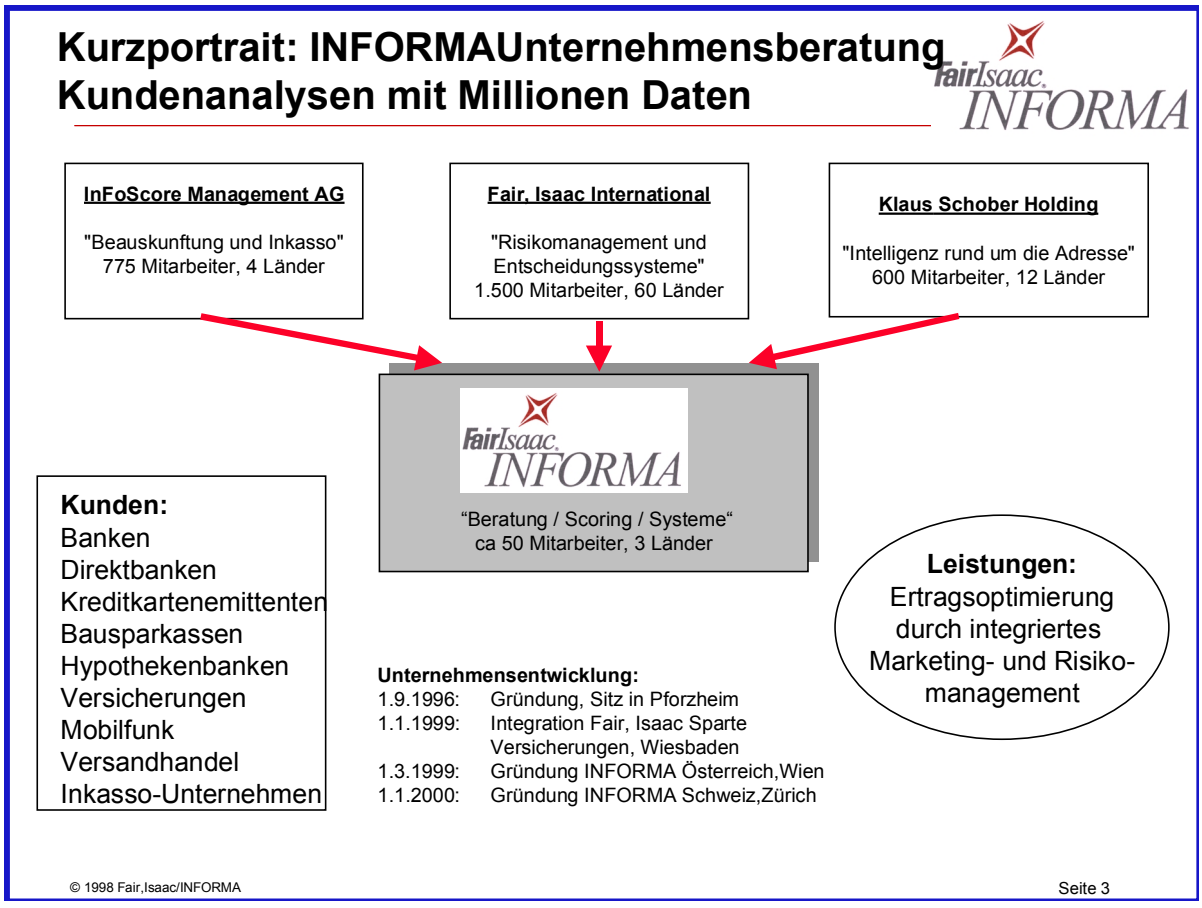
Eine Variante von ARA ist einschliesslich automatisierter geschichteter Stichproben, Potentialschätzung und Scoring in der ScoreXpert Software implementiert. Damit wird leistungsfähiges und kostengünstiges Scoring auch für grosse und grösste Datenmengen bei einfachster Bedienung und hoher Produktivität ermöglicht.

Inhalt

1. Für allgemein-medizinisch Interessierte - zum Thema Schlaganfall mit Schlussfolgerungen für die persönliche Lebensgestaltung insbesondere für den heutigen Abend

2. Für statistisch Interessierte - ein kleines statistisches Experiment zur Rehabilitation der linearen Regression für eine binäre Zielvariable

3. Für Softwareinteressierte - kurze Informationen über das verwendete Werkzeug, die ScoreXpert Software



Schlaganfall - der Blitz aus heiterem Himmel

Nach Angaben der Schlaganfall-Stiftung erleiden jährlich mehr als 250.000 Deutsche einen Schlaganfall.

Jeder Dritte der Betroffenen stirbt an den Folgen (>90.000), ein weiteres Drittel bleibt oft schwer behindert. Der Schlaganfall ist damit nach Krebs und anderen Herz-Kreislauf-erkrankungen die dritthäufigste Todesursache in Deutschland und die häufigste Ursache für eine Behinderung, noch vor Herzinfarkt (vgl. [3], [4]).

Im Vergleich dazu gibt es pro Jahr etwa 8000 Verkehrstote in Deutschland (vgl. [5]).

Täglich sterben fast 250 Menschen an Schlaganfall, über doppelt soviel wie beim Zugunglück von Eschede.

Die Daten

Die im Vortrag erwähnten Schlaganfalldaten sind erstmals von Grau et al. [1] diskutiert worden.
Die Studie umfasste je 197 Probanden mit und ohne Schlaganfall, wobei die Kontrollgruppe ohne Schlaganfälle so ausgewählt wurde, dass Alter und Wohnort möglichst dem jeweiligen Schlaganfall-Patienten entsprechen sollten.
Es gab Folgearbeiten bis 1997 mit umfangreichen medizinischen Hintergrund-Analysen, insbesondere im Hinblick auf das Zusatz-Risiko durch Infekte. Die folgenden Datenanalyse wurde als Privatinitiative des Referenten mit freundlicher Genehmigung des Dateneigners durchgeführt.

Ist die Lineare Regression nicht geeignet?

Üblicherweise wird für binäre Zielvariablen ausschliesslich logistische Regression eingesetzt. Grund dafür ist, dass die theoretischen Voraussetzungen der Anwendung der linearen Regression von vornherein nicht erfüllt sind:
„There are several difficulties with using ordinary least squares. ...
Conditions that make least squares estimates optimal are not satisfied here.
For instance the variance of Y is $p(x)(1-p(x))$, which is not constant over the range of explanatory variables.
Standard distributional statements for estimators do not apply, since Y is dichotomous rather than normally distributed. ... The model itself is likely to be inaccurate in certain regions, however, if some Xi are quantitative. This follows because the model predicts the impossible values $p < 0$ and $p > 1$ for sufficiently large or sufficiently small values of Xi ...“ (vgl. [2]).
Trotzdem ist aus praktischen Anwendungen der ScoreXpert Software bekannt, dass die lineare Regression mit diskretisierten numerischen Variablen zum Teil besser abschneidet als die entsprechende logistische Regression.

Data-Mining für nur 394 Datensätze?

Biometrisch-statistische Analyse

Logistische Regression :

Vorteil: Ausnutzen der Zusatzinformation von Kontrollpaaren mit bedingter logististischer Regression

Nachteil: kombinierte Risiken mit Odds-Ratios komplizierter. Solche Ergebnisse wie z. B. „OR - odds ratio“ Chancen-Verhältnis; z. B. „Infection remained .... a significant risk factor in a logistic model (OR, 4.6; 95% CI, 1.9 to 11.3).“ sind durch Endanwender schwer interpretierbar.

Binäre Datamining-Regression (Automatische Response Analyse)

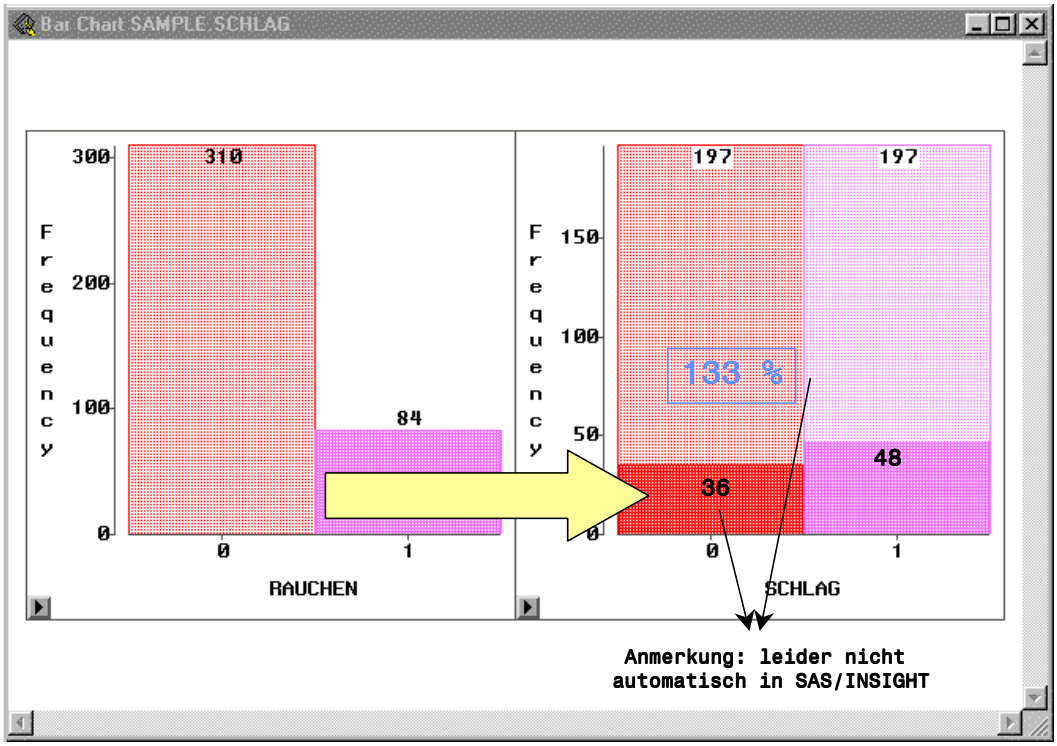
Nachteil: Kontrollpaare nicht nutzbar und weniger Trainingsdaten

Vorteile:
- schnelle automatische Analyse und anschauliche Punktebewertung
- anschauliche Güteschätzung der Prognose auf Prüfstichprobe
- Hochrechnung auf Gesamtbevölkerung einfach möglich
- Stabilität der Prognose bei verschiedenen Stichproben wird deutlich

=> Ein Versuch wert !

„Reaktionsthermometer“ mit SAS/INSIGHT - Derzeitiges Rauchen

FairIsaac. INFORMA



 Seite 12

Punktbewertung Schlaganfallrisiko mit Automatischer Response Analyse

- Bluthochdruck 132
- Rauchen 150
- Diabetes 204
- Infekt 252
- Hirnrisiko 262

- lineares Modell (binäre lineare Data-Minining Regression), Signifikanz 0.1, ohne Alter, 30%Teststichprobe, Stichprobe 1
- Univariat - multivariat: Rauchen stärker als bei univariater Betrachtung, Alkohol fällt weg!!
- Hinweis: in ScoreXpert automatisch 0-1000 Punkte.
- Stimmt mit Basis-Artikel ganz gut überein. Risiken dort Odds-Ratios: Hochdruck: 1.7 Rauch: 1.9 Diab: 3.4 Infekt: 4.6 Hirn: 4.6 Herz: 2.2

Individuelle Risikoerhöhung - die Risikoleiter (ScoreXpert Software)

FairIsaac. INFORMA

ScoreXpert Software - ARA-Output

C:\ScoreXpert\Projects\Schlag\L_1_1R.lst

```
**********************************************************************;

.....

Estimated Responses for 100K Population
Estimated Rates from Test(Validation)-Data
Expected overall Response Rate is 0.305 %
   Using Y O U R  Size Numbers  Respondents/ALL =  250000/82000000
   Supplementary Variable = ALTER
```

Deutschland
Schlaganfälle / Bevölkerung

Ausfallrisiko / %-Erhöhung / Punktwert

Scoreclass with %	High-Scores 100K	Hi-Scores Resp-Rate 100K	Response-Rate Index	Cut-off-Score
0: 100%	100000	0.3	100	0.00
45: 55%	56010	0.4	145	79.11
47: 53%	56010	0.4	145	132.39
67: 33%	33996	0.6	194	142.23
69: 31%	22148	0.8	268	149.07
79: 21%	22148	0.8	268	186.51
81: 19%	18747	0.9	286	252.76
82: 18%	13647	0.9	310	281.46
88: 12%	13647	0.9	310	336.58
91: 9%	3462	2.4	783	393.98
98: 2%	3462	2.4	783	401.83
99: 1%	1762	4.1	1347	523.31
100: 0%	0	.	.	1000.00
			Average	

(Anmerkung: Grobe Schätzung, wegen Methodik der Stichprobe)

Goback

Kleiner Vergleich Linear - Logistisch

FairIsaac INFORMA

- Güte für Segmente 1.+3.Dezil für 4 Stichproben Heidelberger Schlaganfallstudie

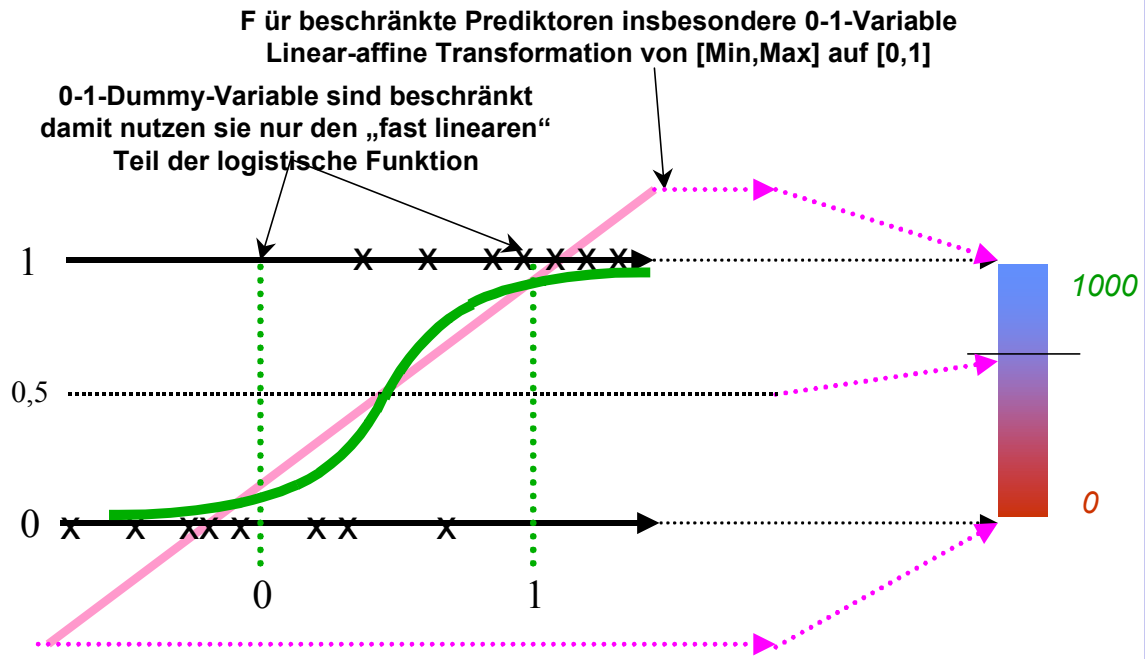
	Linear	Logistisch	Linear mit Wechselwirkungen	Logistisch	
1	194;783	194;783	126;287	190;429	
2	182;470	182;470	211;462	182;262	
3	221;282	377;180	280;347	237;372	
4	298;265	289;298	260;199	215;239	
Basis ohne WW			Basis mit WW		
1	238;416	238;543	231;347	209;331	8Stunden
2	199;455	199;455	199;455	194;494	
3	251;396	251;396	251;396	226;435	8Stunden
4	231;259	231;259	184;259	199;298	

- **Güte: linear ist vergleichbar gut (+/ - /“ “) , bei „echten“ grossen Projekten ist die lineare Regression praktisch gleichwertig**
- **Rechenzeit: Nachteil für logistische Regression, vor allem bei grosser Anzahl von Merkmalen. Konvergenzprobleme, nicht vorhersehbar.**

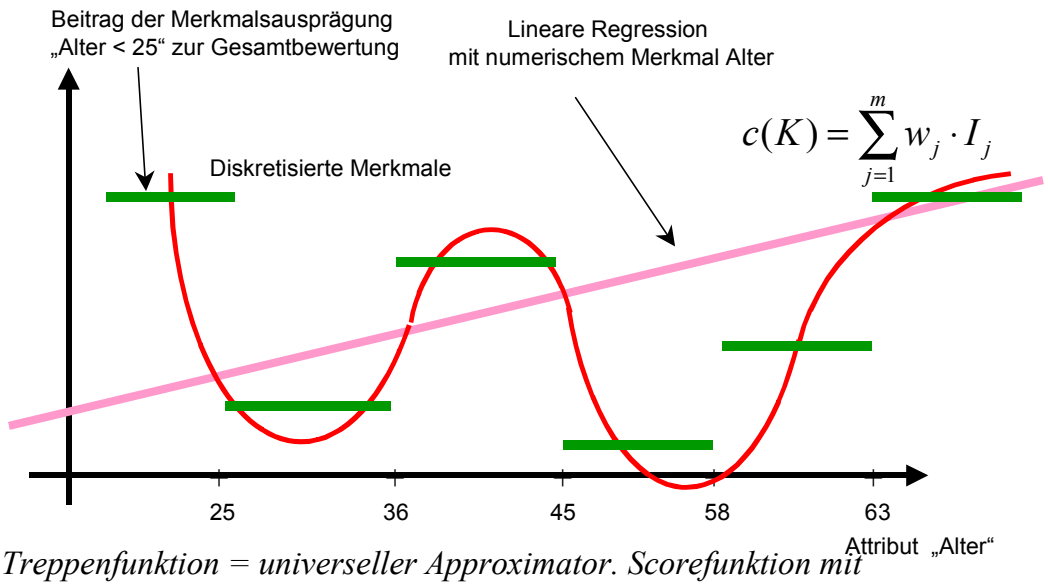
====> „lineare“ diskretisierte binäre Datamining-Regression ist das Rückgrat des Datamining

Lineare Regression für binäre Zielvariable - wieso ist sie so gut gegen die logistische?

FairIsaac INFORMA



Nichtlinearität mit diskretisierten numerischen Variablen



Treppenfunktion = universeller Approximator. Scorefunktion mit Wechselwirkungen. Achtung: nur die signifikanten s_{ij}, w_{ijkl} sind ≠ 0

$$s(K) = s_0 + \sum_{i=1}^{m}\sum_{j=1}^{p_i} s_{ij} \cdot I_{ij} + \sum_{i=1}^{m}\sum_{j=1}^{p_i}\sum_{k=1}^{m}\sum_{l=1}^{p_k} w_{ijkl} \cdot I_{ij} I_{kl}$$

 Seite 17

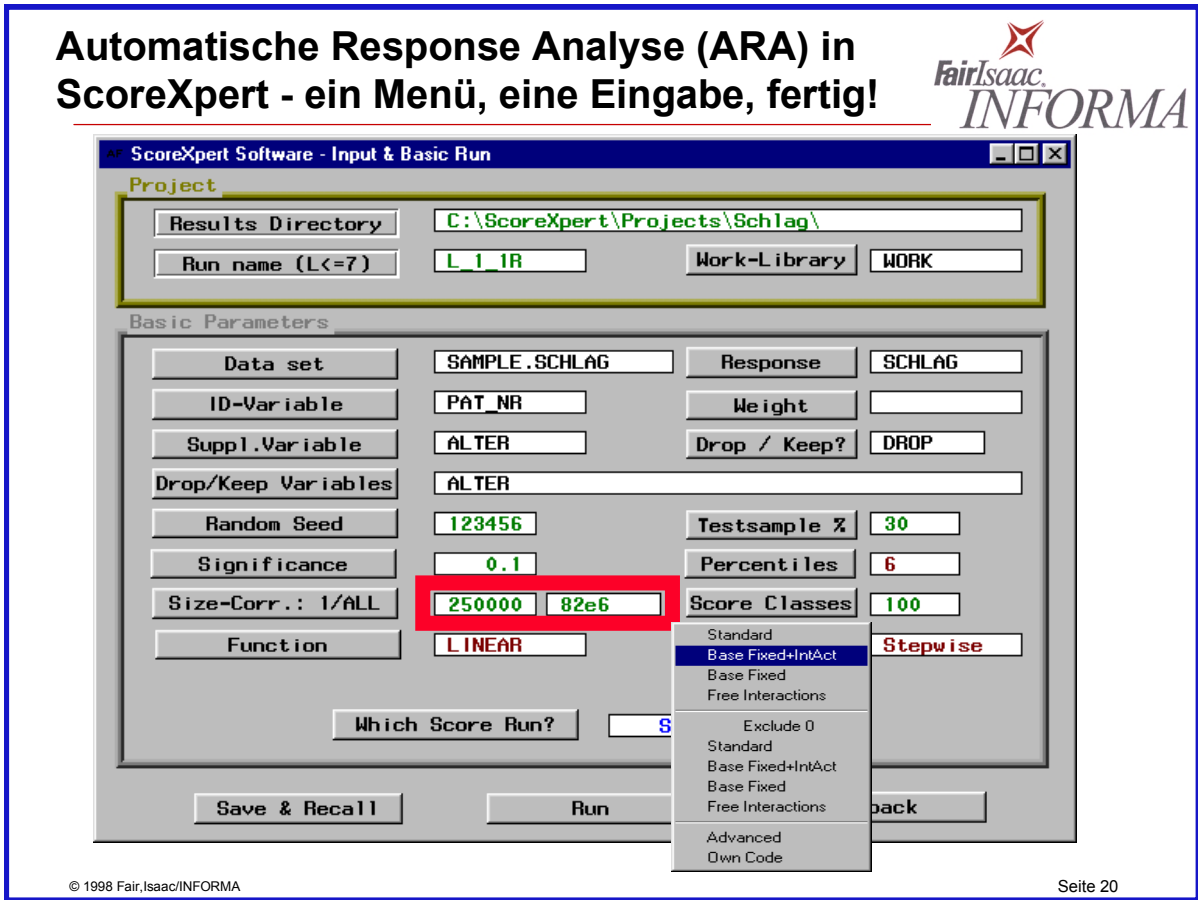
ARA - Eine Wechselwirkungsanalyse mit linearer Zwangsregression

Die Wechselwirkungsanalyse mit linearer Zwangsregression umfasste 144 automatisch erzeugte Wechselwirkungsterme.

+(259.752117996816)
+(HYPERTON='1')*(140.031231489277)
+(RAUCHEN ='1')*(148.093921855046)
+(INFEKT ='1')*(221.380483674011)
+(DIABETES='1')*(298.656155041128)
+(HIRNRISK='1')*(-14.7652160504472)

+(DIABETES='0')
(HIRNRISK='1')(410.402828384509)

- Beispiel aus der Wechselwirkungsanalyse, Signifikanz: 0.05, Stichprobe 1. Hierbei sind knapp 200 Wechselwirkungsterme automatisch erzeugt und in Stepwise Regression mit Zwangs-“Include“ für Basisvariable ausgewertet worden.
- Eigentlich haben 394 Fälle eine zu geringe Zellbesetzung für die Wechselwirkungs-analyse. Trotzdem können signifikante Terme in dem optimal ausgewählten Modell erscheinen. In der Praxis werden mehrere tausend Wechselwirkungsterme automatisch erzeugt und ausgewertet. Die Rechenzeit liegt bis etwa 5000 Dummy-Variable unter einer Stunde.



Schlussfolgerungen

1. Langfristig: Korrekte eigene Vorsorge mit persönlicher Risikoanalyse durch Stiftung Deutsche Schlaganfall-Hilfe
   http://www.schlaganfall-hilfe.de/risiko/frageb.htm

2. Kurzfristig: Schlussfolgerungen für den KSFE-Mixer
   Rauchen: Nein!, Saufen: Ja!

Literatur

[1] Stroke 1995 Mar;26(3):373-9, Recent infection as a risk factor for cerebrovascular ischemia. Grau AJ, Buggle F, Heindl S, Steichen-Wiehn C, Banerjeen T, Maiwald M, Rohlfs M, Suhr H, Fiehn W, Becher H, et al. Department of Neurology, University of Heidelberg, Germany

[2] Agresti, Alan (1984): Analysis of ordinal categorical data, Wiley, New York etc., p.105

Weitere Quellen

[3] http://www.statistik-bund.de/basis/d/gesu/gesutab3.htm

[4] http://www.schlaganfall-hilfe.de/ueber/ueb-fram.htm

[5] http://www.statistik-bund.de/presse/deutsch/pm/p9056191.htm