Logistische Regression zur Modellierung von Binärdaten

Armin Koch
Abteilung Medizinische Biometrie
Universität Heidelberg
Im Neuenheimer Feld 305

69120 Heidelberg

1. Einleitung

Dieser Beitrag soll eine anwendungsorientierte Einführung in das Modellieren von Binärdaten mit Hilfe der logistischen Regression geben, die nach Ansicht des Autors eines der wichtigsten Auswertungsverfahren der angewandten Statistik ist. Einschränkend muß bemerkt werden, daß diese Einschätzung eventuell einer gewissen Verzerrung unterliegt. Viele Antwortvariablen in der Medizin sind binär (pro Beobachtungseinheit wird, eventuell in einem übertragenen Sinne, ein "Erfolg" oder ein "Mißerfolg" festgestellt) und die Parameter des Modells sind gut interpretierbar.

Anhand eines Beispieldatensatzes aus einer großen Beobachtungsstudie im Herzinfarktbereich, soll das Modell und die Interpretation der Parameter vorgestellt werden. Dabei werden das Prinzip der Adjustierung und der Effektmodifikation vorgestellt und der Umgang mit vielen Variablen und Modellsuchstrategien vorgestellt. Ein kurzer Abschnitt über Prädiktion mit Hilfe von logistischer Regression beschließt diese Einführung.

Für alle Auswertungen des Beispiels werden, gegebenenfalls in Auszügen, sowohl der notwendige SAS-Code, als auch der entsprechende Ergebnisausdruck beigefügt, sodaß der Text auch als Einführung in die Benutzung der entsprechenden SAS-Prozeduren dienen kann.

Dieser Text bietet keine neuen Ergebnisse über die logistische Regression. Wenn der Leser mit den oben angesprochenen Themen vertraut ist, oder die exzellente Monographie von Hosmer & Lemeshow[5] gelesen hat, so kann ihm dieser Text außer einem hoffentlich interessanten Beispiel nur wenig neues bieten.

2. Das Beispiel

2.1. Medizinischer Hintergrund

Das 60-Minuten-Herzinfarktprojekt[6; 7] ist eine multizentrische prospektive Beobachtungsstudie in der zwischen 1992 und 1994 an 136 Kliniken in Deutschland etwa 13000 Patienten nach einem Herzinfarkt registriert wurden. Ziel der Studie war es, Informationen über die Prähospitalzeit, also die Zeit zwischen dem Infarktereignis und der Klinikaufnahme, und das therapeutische Verhalten der Ärzte beim Akutinfarkt unter den Bedingungen des "klinischen Alltags" zu gewinnen. Die Prähospitalzeit beschreibt nicht nur die Qualität der Bevölkerungsaufklärung über die richtige Verhaltensweise beim Akutinfarkt, sondern auch die Rettungskette in der Akutversorgung. Darüber hinaus ist die Prähospitalzeit auch wichtig für die Therapieentscheidung. Oberstes Ziel der therapeutischen Bemühungen ist es, die verschlossenen Versorgungsgefäße des Herzens,

die die Ursache für den Infarkt sind, möglichst schnell wieder zu eröffnen, um eine weitergehende Schädigung des Herzmuskels zu vermeiden. Wichtigster Therapieträger ist die Thrombolyse (medikamentöse Auflösung des Gefäßverschlusses), die umso besser funktioniert, je schneller nach dem Gefäßverschluß sie eingesetzt wird, und die, wie beinahe alle Therapien nicht ohne Nebenwirkungen ist: mit der Thrombolyse ist eine Herabsetzung der Blutgerinnung verbunden, die bei gefährdeten Patienten zu einer Blutung (z.B. Schlaganfall) führen kann. Mit zunehmendem Abstand des Therapiebeginns zum Infarkt werden folglich die wünschenswerten Effekte der Therapie abnehmen und die Wahrscheinlichkeit für ein unerwünschtes Ereignis zunehmen.

Zur Illustration der folgenden Ausführungen wird ein Subkollektiv von rund 700 Patienten herangezogen, bei denen vor Behandlungsbeginn ein Herzstillstand aufgetreten ist und die mechanisch reanimiert werden mußten. Es handelt sich hierbei um eine Patientengruppe, die unter einem extrem hohen Risiko steht, den aktuellen Infarkt nicht zu überleben. Obgleich der Nutzen der Thrombolyse beim Herzinfarkt außerhalb jeder Diskussion steht, ist die mechanische Reanimation (wegen des hohen Blutungsrisikos nach diesem Vorgang) lange Zeit als Kontraindikation angesehen worden. Randomisierte Studien in dieser Patientengruppe sind extrem schwer durchführbar, da sich die Patienten eben in einem sehr schlechten Zustand befinden. Da im Rahmen des 60-Minuten-Herzinfarktprojektes auch Patienten registriert wurden, bei denen nach der Entscheidung des behandelnden Arztes nach einer mechanischen Reanimation eine Lyse durchgeführt wurde, bestand die Frage, ob man mit Hilfe dieser Daten eine Aussage über den Nutzen der Lysetherapie für diese spezielle Patientengruppe treffen kann.

2.2. Deskription und Beschreibung von Effekten in der Vierfeldertafel

Werden pro Proband im Rahmen einer Untersuchung zwei dichotome Variablen erhoben (hier z.B. ob der Patient eine Lyse erhalten hat oder nicht und ob der Patient im Krankenhaus verstorben ist oder nicht), so kann das Untersuchungsergebnis in einer Vierfeldertafel zusammengefaßt werden:

Tabelle 1: Vierfeldertafel

		Status	
		verstorben	überlebt
Population	Lyse	a	b
	keine Lyse	c	d

Das entsprechende Ergebnis erhält man in SAS mit *proc freq*, wobei hier die Variablen lyse und klintod, die vermerken, ob der Patient eine Lysetherapie erhalten hat und ob der Patient während des Krankenhausaufenthaltes verstorben ist, als 0/1-Variablen codiert wurden.

Programmcode:

```
;proc freq data=a.herzmass
;     tables lyse*klintod / cmh nocol
;run
;
```

Ergebnis (in Auszügen):

```
LYSE(Thrombolyse)
          KLINTOD(in der Klinik verstorben)
```

Frequency| Percent | Row Pct	0	1	Total
0	135	223	358
	20.49	33.84	54.32
	37.71	62.29	
1	157	144	301
	23.82	21.85	45.68
	52.16	47.84	
Total	292	367	659
	44.31	55.69	100.00

Obgleich schon zu Studienbeginn bekannt war, daß sich die Einschätzung, daß die Lyse bei Patienten nach Reanimation kontraindiziert ist, verändert hat, sind die teilnehmenden Ärzte doch überrascht gewesen, daß beinahe die Hälfte der Patienten (46%) eine Lyse erhalten haben. Deutlich wird auch, die schlechte Prognose der Patienten nach einer mechanischen Reanimation, von denen 56% den Krankenhausaufenthalt nicht überleben.

In randomisierten klinischen Studien wird der Therapieeffekt (als Unterschied zwischen behandelten und unbehandelten Patienten) üblicherweise durch die Risikodifferenz oder das Risikoverhältnis beschrieben. Bezeichnen $\hat{p}_0 = b/(a+b)$ und $\hat{p}_1 = d/(c+d)$ die Schätzwerte für die Wahrscheinlichkeit ohne bzw. mit Therapie im Krankenhaus zu versterben, so ergibt sich für die Risikodifferenz $\hat{\Delta} = \hat{p}_1 - \hat{p}_0 = 0{,}48 - 0{,}62 = -0{,}14$, was dann üblicherweise so interpretiert wird, daß bei einer Behandlung von 100 Patienten 14 Leben mehr gerettet werden können. Für das Risikoverhältnis erhält man $\hat{RR} = \hat{p}_1 / \hat{p}_0 = 0.77$, also das Risiko nach einem Herzinfarkt (mit einer Reanimation) im Krankenhaus zu versterben ist unter einer thrombolytischen Therapie geringer.

Der Chancenquotient (das Odds-Ratio bzw. das Kreuzproduktverhältnis) ist ursprünglich für Fall-Kontroll-Studien entwickelt worden. Es wird wie folgt aus den Schätzungen für die Erfolgswahrscheinlichkeiten in den beiden Gruppen berechnet:

$$\hat{OR} = \frac{\hat{p}_1/(1-\hat{p}_1)}{\hat{p}_2/(1-\hat{p}_2)} = \frac{a \times d}{b \times c} = \frac{135 \times 144}{157 \times 233} = 0{,}55$$

Offensichtlich sind Odds-Ratio und Risikoverhältnis ähnlich, wenn man seltene Ereignisse betrachtet. In dieser Situation ist $1 - \hat{p}_i (i = 1{,}2)$ klein. Die Interpretation ist, zumindest für Deutsche, die ja nicht so wettbegeistert sind, wie ihre englischen Nachbarn, etwas schwierig: in unserer Sprache ist häufig der Begriff Chance das Gegenteil des negativen Begriffs Risiko. Mathematisch ist eine Chance, also z.B. $\hat{p}_1/(1-\hat{p}_1)$ nichts anderes als der Quotient aus Wahrscheinlichkeit und Gegenwahrscheinlichkeit ("Wie steht es, daß der Patient wieder gesund wird?" - "Ja, also, fifty - fifty"). Der Chancenquotient vergleicht (aus der Sicht des Glücksspielers) die Gewinnchancen zweier verschiedener Glücksspiele (ist er größer als 1, so würde ich das Glücksspiel, dessen Erfolgwahrscheinlichkeit im Zähler des Odds-Ratio eingetragen ist, bevorzugen).

Üblicherweise werden zu Schätzern Konfidenzintervalle (im SAS-Ausdruck: 95% confidence bounds) angegeben, die einen Bereich angeben, in dem der wahre Wert unter der Hypothese, daß eine Zufallsstichprobe aus einer Grundgesamtheit gezogen wurde, mit einer vorgegebenen Wahrscheinlichkeit (hier 95%) liegt. Dies kann man auch als eine Art Genauigkeitsaussage für die Schätzung verstehen. Man erhält sie mit der Option */cmh* in der obigen *tables*-Anweisung.

Ergebnis (in Auszügen):

```
Estimates of the Common Relative Risk (Row1/Row2)
                                                    95%
Type of Study    Method              Value   Confidence Bounds
-----------------------------------------------------------------
Case-Control     Mantel-Haenszel     0.555      0.407    0.757
  (Odds Ratio)   Logit               0.555      0.407    0.758

Cohort           Mantel-Haenszel     0.723      0.609    0.858
  (Col1 Risk)    Logit               0.723      0.609    0.858

Cohort           Mantel-Haenszel     1.302      1.133    1.496
  (Col2 Risk)    Logit               1.302      1.129    1.502
```

Seine große Bedeutung hat das Odds-Ratio dadurch gewonnen, daß man es mit statistischen Modellen untersuchen kann. Hiervon handelt der folgende Beitrag.

Betrachtet man den Vergleich lysierter und nicht lysierter Patienten aus inhaltlicher Sicht, so wäre es ein hervorragendes Ergebnis für die Thrombolyse, wenn der beobachtete Behandlungseffekt von 62%-48%=14% ein "Therapieeffekt wäre, daß also durch die Behandlung 14 Prozent mehr Patienten gerettet werden könnten. Bedauerlicherweise sind die Gruppe der lysierten und die Gruppe der nicht-lysierten Patienten schon a priori verschieden (siehe Tabelle 2).

Tabelle 2: Ausgangssituation

	Lyse nein	 ja
% Männer	68,6	77,9
% Prähospitalzeit<4h	80,3	88,0
% Alter>75	28,4	10,1
% Linksschenkelblock	14,7	8,0
% Vorderwandinfarkt	51,9	53,4
% syst. Blutdruck<90	48,1	41,9
% Reinfarkte	26,3	22,4

Auch die Begleitmedikation ist in den beiden Gruppen nicht gleich: Acetylsalicylsäure (ASS) haben 53 bzw. 76 Prozent der Patienten erhalten, beim Beta-Blocker sind es 8,7 bzw. 11,4 Prozent und eine Nitro-Therapie haben 50 bzw. 48 Prozent der Patienten erhalten. Offensichtlich besteht in Abhängigkeit von verschiedenen Patientencharakteristika eine unterschiedliche Chance, daß ein Patient eine Lysetherapie erhält. So waren die Patienten, die eine Lyse erhalten haben, jünger und haben das Krankenhaus schneller erreicht. Zusätzlich beeinflussen alle diese Variablen in unterschiedlichem Ausmaß die Wahrscheinlichkeit, nach einem Infarkt mit Herzstillstand zu versterben.

Zu klären ist also, ob die Sterbewahrscheinlichkeit in der Lyse-Gruppe tatsächlich geringer ist, oder ob der beobachtete Effekt eine Folge der Unterschiede in den Patientengruppen ist (z.B. weil in der Lysegruppe bei weniger Patienten der aktuelle Infarkt bereits ein Reinfarkt gewesen ist).

3. Regression für Binärdaten

Häufig wird multiple lineare Regression benutzt, wenn es darum geht, den Einfluß einer Reihe von unabhängigen Einflußgrößen auf eine Zielvariable darzustellen. So kann man mit dem

$$\text{Modell: Blutdruck} = \beta_0 + \beta_1 \times \text{Alter} + \beta_2 \times \text{Geschlecht}$$

untersuchen, ob der Blutdruck vom Alter und dem Geschlecht abhängt. Die Variablen $\beta_i (i = 0,1,2)$ heißen dabei Regressionskoeffizienten. In der hier betrachteten Situation besteht das Problem, daß eine direkte Modellierung von Erfolgsraten unter Umständen nicht möglich ist, da ja die vorhergesagten Werte (die Mortalität nach Herzinfarkt ist eine Wahrscheinlichkeit) das Intervall [0,1] nicht verlassen dürfen.

Deshalb führt man eine sogenannte logit-Transformation durch, bei der das Intervall [0,1] auf den ganzen Zahlenstrahl abgebildet wird:

$$\text{logit}(p) = \log\left(\frac{p}{1-p}\right)$$

Diese Abbildung ist eineindeutig. Tabelle 3 stellt an einigen Beispielen den Effekt der logit-Transformation dar.

Tabelle 3: Effekt der Logit-Transformation

p	0,01	0,05	0,10	0,25	0,50	0,75	0,90	0,95	0,99
logit	-4,60	-2,94	-2,20	-1,10	0,00	1,10	2,20	2,94	4,60

Für den logit kann man dann ein lineares Modell "rechnen". Wollte man im einfachsten Fall den Einfluß der Variablen Alter auf die Mortalität untersuchen, so würde das Modell

$$\text{logit}(p) = \beta_0 + \beta_1 \times \text{Alter}$$

lauten, wobei p die Wahrscheinlichkeit bezeichnet, daß ein Patient unserer Untersuchungsgruppe in der Klinik verstirbt. Hier sind jedoch auch die Einflußvariablen dichotom. Bezeichnet (Lyse) eine 0/1-Variable, die anzeigt, ob ein Patient eine Lyse-Therapie erhalten hat, so lautet das Modell:

$$\text{logit}(p) = \beta_0 + \beta_1 \times Lyse$$

und man erhält:

$$\begin{aligned}\text{logit}(p_{(Lyse)}) - \text{logit}(p_{(\neg Lyse)}) &= \log\left(\frac{p_{(Lyse)}}{1-p_{(Lyse)}}\right) - \log\left(\frac{p_{(\neg Lyse)}}{1-p_{(\neg Lyse)}}\right) \\ &= \log(OR_{(Lyse)}) = \beta_1\end{aligned}$$

Dabei bezeichnet (¬Lyse) eine Indikatorvariable, die genau dann den Wert 1 annimmt, wenn die ursprüngliche Indikatorvariable (Lyse) den Wert 0 hat und umgekehrt. Die Differenz der logits ist also gerade der Logarithmus des Odds-Ratios für den Vergleich der Patienten mit bzw. ohne Lysetherapie. Der Koeffizient einer Variable im logistischen Regressionsmodell entspricht gerade dem Logarithmus des entsprechenden Odds-Ratios. Zur Auswertung mit SAS sieht dies wie folgt aus:

Programmcode:

```
;proc logistic descending data=a.herzmass
;     model klintod=lyse / risklimits
;run
;
```

Ergebnis (in Auszügen):

```
Analysis of Maximum Likelihood Estimates

                      Parameter         Standard          Odds
Variable      DF       Estimate            Error         Ratio
INTERCPT       1         0.5019           0.1090             .
LYSE           1        -0.5883           0.1588         0.555
```

und man erhält: $OR = e^{(-0.5883)} = 0.555$. Ein Konfidenzintervall für den Logarithmus des Odds-Ratios erhält man unter der Annahme, daß die Schätzung für den Paramter normalverteilt ist, wie üblich durch:

$$\begin{aligned} KI_{0.95}(\log(OR)) &= [\log(OR) \pm 1.96 \times se(\log(OR))] \\ &= [-0.5883 \pm 1.96 \times 0.1588] \\ &= (-0{,}899;\ -0{,}277) \end{aligned}$$

Das Konfidenzintervall für das Odds-Ratio erhält man, indem man e hoch die entsprechenden Grenzen nimmt. Mit der */risklimits*-Option in der *model*-Anweisung rechnet das SAS aber auch selbst aus:

Ergebnis (in Auszügen):

```
Conditional Odds Ratios and 95% Confidence Intervals

                                    Wald
                                Confidence Limits
                      Odds
Variable     Unit     Ratio    Lower    Upper
LYSE       1.0000     0.555    0.407    0.758
```

Bis hierher hat man also (fast) dasselbe wie mit *proc freq* erhalten. Das besondere ist jedoch, daß es mit Regressionsmodellen möglich ist, gleichzeitig den Einfluß von mehreren Variablen auf das Zielereignis zu untersuchen. Im folgenden Abschnitt wird beschrieben, wie die Parameter in dieser Situation zu interpretieren sind.

4. Interpretation der Parameter

4.1. Adjustierung

Aufgezeigt wurde, daß lysierte und nicht-lysierte Patienten eine unterschiedliche Altersstruktur aufweisen. Unter "Adjustierung" versteht man nun die Umrechnung des Lyse-Effektes in eine Population, in der lysierte und nicht lysierte Patienten eine gleiche Altersstruktur haben. Mit *proc freq* und einer *by*-Anweisung kann man die Mortalität in den beiden Altersgruppen unter Lyse bzw. wenn keine Lyse gegeben wurde, getrennt berechnen. Dabei entsteht eine 2×2×2-Kontingenztafel.

Ergebnis (in Auszügen):

```
CONTROLLING FOR ALTER75=0
LYSE(Thrombolyse)
          KLINTOD(in der Klinik verstorben)
Frequency|
Percent  |
Row Pct  |       0|       1|  Total
---------+--------+--------+
       0 |    107 |    149 |    256
         |  20.34 |  28.33 |  48.67
         |  41.80 |  58.20 |
---------+--------+--------+
       1 |    145 |    125 |    270
         |  27.57 |  23.76 |  51.33
         |  53.70 |  46.30 |
---------+--------+--------+
Total         252      274      526


CONTROLLING FOR ALTER75=1
LYSE(Thrombolyse)
          KLINTOD(in der Klinik verstorben)
Frequency|
Percent  |
Row Pct  |       0|       1|  Total
---------+--------+--------+
       0 |     28 |     74 |    102
         |  21.05 |  55.64 |  76.69
         |  27.45 |  72.55 |
---------+--------+--------+
       1 |     12 |     19 |     31
         |   9.02 |  14.29 |  23.31
         |  38.71 |  61.29 |
---------+--------+--------+
Total          40       93      133
```

Sowohl in der Altersgruppe unter 75 Jahren, als auch in der Gruppe der älteren Patienten gilt, daß die Mortalität unter der Lysetherapie um etwa 10% geringer ist. Bei älteren Patienten ist die Mortalität jedoch generell größer. Da in logistischen Regressionsmodellen im logit-Maßstab gerechnet wird, sind in Tabelle 4 noch einmal die entsprechenden Ergebnisse zusammengestellt.

Tabelle 4: Mortalität für alte / junge Patienten

Alter	Lyse	p × 100%	logit(p)	Differenz
<75	nein	58,2	0,331	
	ja	46,3	-0,148	-0,479
>75	nein	72,6	0,974	
	ja	61,3	0,460	-0,514

Adjustierung bedeutet, daß ein gewichteter Mittelwert der Lyse-Effekte im logit-Maßstab berechnet wird:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{2} w_i \log(OR_i)}{\sum w_i}$$

Dabei ist $w_i = 1/Var(\log(OR_i))$, wobei die Varianz des Logarithmus des Odds-Ratios üblicherweise durch $1/a_i + 1/b_i + 1/c_i + 1/d_i$ geschätzt wird und a_i, b_i, c_i, d_i die Zellhäufigkeiten in der $i-$ten Vierfeldertafel bezeichnen. Grob gesprochen werden also die Odds-Ratios mit der Fallzahl in den beiden durch die Variable Alter gebildeten Gruppen gewichtet (die größere Gruppe erhält den größeren Einfluß bei der Mittelwertsbildung). Mit *proc logistic* erhält man hierfür:

Programmcode:

```
;proc logistic descending data=a.herzmass
;     model klintod=lyse alter75
;run
;
```

Ergebnis (in Auszügen):

```
Analysis of Maximum Likelihood Estimates

                Parameter   Standard      Wald         Pr >
Variable   DF    Estimate     Error    Chi-Square   Chi-Square

INTERCPT   1       0.3336     0.1220       7.4807       0.0062
LYSE       1      -0.4842     0.1629       8.8357       0.0030
ALTER75    1       0.6309     0.2134       8.7377       0.0031
```

Damit: OR=exp(-0,4842)=0,616 (im Vergleich zu: 0.555 unadjustiert). Durch das Hinzufügen eines weiteren erklärenden Faktors ist also der "Lyse-Effekt" kleiner geworden (er hat sich in Richtung auf die 1 bewegt). Man beachte auch, daß das Vorzeichen des Parameterschätzers negativ ist, wenn in der mit 1 codierten Gruppe der entsprechenden Variablen (hier die Lyse) die Mortalität geringer ist.

4.2. Interaktion

Im Unterschied zur Adjustierung, wo in den beiden zu vergleichenden Gruppen (hier bei den jungen und den alten Patienten) im wesentlichen derselbe Effekt beobachtet wurde, kann auch eine Interaktion (Effektmodifikation) auftreten. Das Beispiel hier ist der Linksschenkelblock (LSB, eine Reizleitungsstörung im Erregungsmechanismus des Herzmuskels) und die Lysetherapie. Betrachtet man ein einfaches logistisches Regressionsmodell, das neben der unabhängigen Variablen Lyse nun noch die Variable Linksschenkelblock mitbetrachtet, so erhält man:

Ergebnis:

```
Analysis of Maximum Likelihood Estimates

                Parameter   Standard      Wald         Pr >
Variable   DF    Estimate     Error    Chi-Square   Chi-Square

INTERCPT   1       0.3599     0.1150       9.7940       0.0018
LYSE       1      -0.5250     0.1618      10.5262       0.0012
LSB        1       0.8803     0.2750      10.2482       0.0014
```

Man würde schließen, daß in der Lysegruppe die Chance zu versterben geringer ist und der Linksschenkelblock mit einer erhöhten Mortalität einhergeht (das zugehörige Odds-Ratio ist 2,4). Die Mortalität bei Vorliegen eines Linksschenkelblocks ist also mehr als doppelt so groß, wie wenn dieser Risikofaktor nicht vorliegt). Betrachtet man jedoch die einzelnen Strata, so ergibt sich das folgende Bild:

Tabelle 5: Lyse bei Patienten ohne LSB

		verstorben	
		nein	ja
Lyse	nein	118 (40%)	181 (60%)
	ja	151 (56%)	119 (44%)

Tabelle 6: Lyse bei Patienten mit LSB

		verstorben	
		nein	ja
Lyse	nein	17 (31%)	37 (69%)
	ja	3 (13%)	21 (87%)

Während im Kollektiv der Patienten ohne Linksschenkelblock in der Lysegruppe die Mortalität wieder geringer ist, als in der Gruppe der nichtlysierten Patienten, erhält man für Patienten mit Linksschenkelblock ein gegenläufiges Ergebnis: unter Lyse ist die Mortalität deutlich größer.

Tabelle 7: Mortalität und LSB

LSB	Lyse	p × 100%	logit(p)	OR
nein	nein	60,5	0,428	
	ja	44,1	-0,238	0,514
ja	nein	68,5	0,777	
	ja	87,5	1,168	3,126

Hieraus kann man nicht direkt schließen, daß die Lyse für Patienten mit Linksschenkelblock schlecht ist (sogenannte "ultima ratio"-Lyse: es wird alles unternommen, um einen gefährdeten Patienten doch noch zu retten). Dennoch ist das Ergebnis, wenn es genaueren Untersuchungen standhält, wegen der extremen Risikoerhöhung bedenklich. Wichtig ist, daß man dieses Ergebnis dem oben gerechneten logistischen Regressionsmodell nicht ansehen würde. Hier muß man in das logistische Regressionsmodell einen weiteren Term (einen sogenannten Interaktionsterm) mit aufnehmen, der es gestattet, den sonst in beiden Gruppen als gleich angenommenen Effekt der Lyse zu entkoppeln. Eine dritte Variable wird als Produkt der Indikatoren für die Lysetherapie und den Linksschenkelblock gebildet, die in allen Fällen den Wert 0 hat und nur für die Gruppe der lysierten Patienten, bei denen gleichzeitig ein Linksschenkelblock vorliegt, den Wert 1 annimmt:

Programmcode:

```
;data eins; set a.herzmass; intll=lyse*lsb; run
;proc sort; by lsb; run
;proc logistic descending
;      model klintod=lyse lsb intll / covb
;run;
```

Ergebnis (in Auszügen):

```
Analysis of Maximum Likelihood Estimates

                Parameter  Standard     Wald        Pr >     Odds
Variable   DF   Estimate    Error   Chi-Square  Chi-Square  Ratio

INTERCPT    1     0.4278    0.1183     13.0736      0.0003      .
LYSE        1    -0.6660    0.1704     15.2803      0.0001   0.514
LSB         1     0.3499    0.3160      1.2261      0.2682   1.419
INTLL       1     1.8342    0.7042      6.7850      0.0092   6.260
```

In Gegenwart von Interaktionstermen stimmen, wie der obige Ergebnisausdruck zeigt, die von SAS angegebenen Odds-Ratios nicht mehr (man vergleiche mit den in Tabelle 7 von Hand ausgerechneten Odds-Ratios). Jedoch muß man zugeben, daß hier ein Rechenprogramm wie SAS keine Chance hat: betrachtet man das Modell:

$$\text{logit}(p) = \beta_0 + \beta_1(\text{Lyse}) + \beta_2(\text{LSB}) + \beta_3(\text{Lyse} \times \text{LSB})$$

so muß man jetzt offensichtlich noch sagen, welche der betrachteten Variablen die Populationen bezeichnet, in denen ein Effekt untersucht werden soll (also ob der Effekt des Alters bei lysierten und nicht lysierten Patienten, oder der Effekt der Lyse bei alten und jungen Patienten dargestellt werden soll. Das Odds-Ratio für den Effekt der Lyse (Differenzbildung der beiden logits für Lyse=1 und Lyse=0) in der Population der Patienten ohne Linksschenkelblock (die Indikatorvariable LSB hat den Wert 0) erhält man wie üblich nach

$$\log(OR_{Lyse}, LSB = 0) = \beta_1$$

Schätzer und Konfidenzintervall können also wie immer aus dem Ausdruck entnommen werden. In der Population mit LSB erhält man dagegen:

$$\log(OR_{Lyse}, LSB = 1) = \beta_1 + \beta_3$$

also $OR_{Lyse} = \exp(-0{,}6660 + 1{,}8342) = \exp(1{,}1682) = 3{,}22$, was man schon aus der univariaten Auswertung kennt. Die Varianz, die man zur Berechnung eines Konfidenzintervalls benötigt, ergibt sich durch eine "üble Rechnerei", für die hier zunächst die allgemeine Formel angegeben werden soll (die auch gelten würde, wenn LSB ein stetiger Confounder wäre):

$$\begin{aligned}\text{var}(\log(OR_{Lyse}, LSB) = \text{var}(\beta_1)((Lyse) - (\neg Lyse))^2 \\ + \text{var}(\beta_3)(LSB((Lyse) - (\neg Lyse)))^2 \\ + 2\,\text{cov}(\beta_1, \beta_3)(LSB)((Lyse) - (\neg Lyse))^2\end{aligned}$$

Hier ergibt sich zur Berechnung:

$$\text{var}(\log(OR_{Lyse}, LSB = 1)) = \text{var}(\beta_1) + \text{var}(\beta_3) + 2\,\text{cov}(\beta_1, \beta_3)$$

Die Covarianzmatrix erhält man mit der */ covb*-Option in der Modellanweisung von *proc logistic*:

Ergebnis (in Auszügen):

```
Estimated Covariance Matrix

Variable      INTERCPT        LYSE         LSB       INTLL
INTERCPT        0.0140     -0.0140     -0.0140      0.0140
LYSE           -0.0140      0.0290      0.0140     -0.0290
```

```
LSB           -0.0140      0.0140      0.0998     -0.0998
INTLL          0.0140     -0.0290     -0.0998      0.4958
```

Die für die Berechnung erforderlichen Terme sind (hier) unterstrichen worden. Damit berechnet man die Varianz des Schätzers für den Lyse-Effekt im Kollektiv der Patienten mit einem Linksschenkelblock zu

$$\text{var}(\beta_1 + \beta_3) = 0.0290 + 0.4958 - 2 \times 0{,}0290 = 0{,}468$$

Damit berechnet man wieder das entsprechende Konfidenzintervall. Die Ergebnisse sind in Tabelle 8 zusammengefaßt. Es ergibt sich eine "nicht-signifikante", aber deutliche Erhöhung des Sterberisikos in der Gruppe der Patienten mit einem Linksschenkelblock, die jedoch, wie schon oben erwähnt, noch weiter untersucht werden muß. Hingegen ist der "Lyse-Effekt" im Kollektiv der Patienten ohne Linksschenkelblock deutlicher geworden.

Tabelle 8: Ergebnis

Population	Effekt	OR	95%-KI
kein LSB	Lyse	0,514	(0,368; 0,717)
LSB	Lyse	3,220	(0,841; 12,3)

5. Viele Einflußfaktoren

Besteht nun die Möglichkeit, mit Hilfe von logistischer Regression den Einfluß verschiedener unabhängiger Variablen auf eine Erfolgswahrscheinlichkeit zu modellieren, so stellt sich unmittelbar die Frage, wann ein Modell als "gut" angesehen werden soll. Die Antwort mag in verschiedenen Situationen unterschiedlich ausfallen, generell gilt jedoch, daß ein gutes Modell möglichst wenige erklärende Variablen beinhaltet. Dies wird dann deutlich, wenn man sich überlegt, daß das beste Modell zur Vorhersage der aktuellen Situation dasjenige Modell ist, in das man für jeden Patienten eine identifizierende Variable aufnimmt. Es ist perfekt zur "Vorhersage" der Situation im aktuellen Datensatz, man kann jedoch nichts daraus lernen, da es jeden Patienten für einzigartig erklärt und folglich keine Vorhersage für einen zukünftigen Patienten leisten kann.

Es gibt viele Methoden, wie man ein gutes Modell finden kann. Vielfach wird die sogenannte Rückwärtsselektion von Variablen empfohlen, bei der aus dem sogenannten vollen Modell, das alle Variablen enthält, die in der aktuellen Situation von Bedeutung sind, nacheinander Variablen entfernt werden.

Im hier betrachteten Beispiel der Modellierung der Mortalität nach Reanimation sind neben den bereits bekannten Variablen Alter, Linksschenkelblock und Lyse noch die Variablen Geschlecht, Prähospitalzeit kleiner als 4 Stunden, die Infarktlokalisation (VWI = Vorderwandinfarkt), ein Indikator für die Bedingung Aufnahmeblutdruck kleiner als 90 mm Hg (RRK90) und ein Indikator dafür, daß der aktuelle Infarkt bereits ein Reinfarkt ist, mit in das Modell aufgenommen worden.

Programmcode:

```
;proc logistic descending data=a.herzmass
;     model klintod= geschl alter75 phzs4 LSB VWI RRK90 Fruehmi
                     lyse / risklimits
;run;
```

Ergebnis (volles Modell):

Analysis of Maximum Likelihood Estimates

Variable	DF	Parameter Estimate	Standard Error	Wald Chi-Square	Pr > Chi-Square
INTERCPT	1	0.1433	0.2402	0.3558	0.5508
GESCHL	1	-0.00996	0.1941	0.0026	0.9591
ALTER75	1	0.5363	0.2306	5.4077	0.0200
PHZS4	1	-0.4045	0.1722	5.5154	0.0188
LSB	1	0.7424	0.3104	5.7205	0.0168
VWI	1	-0.2116	0.1700	1.5497	0.2132
RRK90	1	0.5454	0.1700	10.2944	0.0013
FRUEHMI	1	0.3598	0.2015	3.1866	0.0742
LYSE	1	-0.3243	0.1743	3.4611	0.0628

Teilt man den Parameterschätzer für einen Regressionskoeffizienten durch seinen Standardfehler, so ist die entstehende Prüfgröße unter der Hypothese, daß der entsprechende Regressionskoeffizient den Wert Null hat, asymptotisch standardnormalverteilt. Das Quadrat dieses Ausdrucks ist dann χ^2-verteilt (Wald Chi-Square) für das in der letzten Spalte im obigen Ausdruck ein P-Wert angegeben wird. Offensichtlich spielen in der betrachteten Situation die Variablen Geschlecht und Infarktlokalisation keine Rolle. Es empfiehlt sich jedoch vielfach, nicht nur auf P-Werte zu schauen: wie das Beispiel Linksschenkelblock und Lyse gezeigt hat, kann auch ein "nichtsignifikanter" Effekt, wenn er sehr groß ist, Anlaß für weitere Untersuchungen sein. Mit der */risklimits*-Option erhält man wieder die Odds-Ratios und die zugehörigen Konfidenzintervalle:

Ergebnis (in Auszügen):

Analysis of Maximum Likelihood Estimates

Variable	Standardized Estimate	Odds Ratio	Conditional Odds Ratio and 95% Confidence Limits Lower	Upper
INTERCPT	.	1.241	0.780	1.974
GESCHL	-0.000737	0.997	0.683	1.455
ALTER75	0.118550	1.721	1.098	2.699
PHZS4	-0.127172	0.631	0.452	0.880
LSB	0.746301	2.109	1.149	3.873
VWI	-0.041130	0.861	0.621	1.195
RRK90	0.157731	1.774	1.277	2.466
FRUEHMI	0.082900	1.419	0.958	2.100
LYSE	-0.098833	0.698	0.498	0.979

Hier gibt es keinen Hinweis darauf, daß sich die Variablen Geschlecht und Infarktlokalisation durch dramatische Risikoerhöhungen auszeichnen (beide Gruppen sind auch relativ groß). Für das "beste Modell" (Option: */selection=backward*) erhält man:

Ergebnis (in Auszügen):

Analysis of Maximum Likelihood Estimates

Variable	Standardized Estimate	Odds Ratio	Conditional Odds Ratio and 95% Confidence Limits Lower	Upper

```
INTERCPT              .       1.247       0.895       1.736
ALTER75         0.125344      1.776       1.146       2.752
PHZS4          -0.134769      0.613       0.441       0.854
RRK90           0.160953      1.795       1.294       2.492
LYSE           -0.098121      0.700       0.500       0.980
```

Man bemerkt, daß obwohl der Linksschenkelblock im vollen Modell enthalten gewesen ist und inhaltlich etwas Bedeutendes passiert, er im Rahmen der Modellsuchstrategie eliminiert wird. Zu beachten ist, daß das Modellieren, das im wesentlichen ein nach mathematischen Kriterien ablaufender Prozess ist, aus inhaltlicher Sicht eine Kunst ist.

6. Modellanpassung

Üblicherweise sind Schätzungen (wie z.B. der Mittelwert) für einen Parameter (im einfachsten Fall der Erwartungswert einer Normalverteilung) Funktionen der Beobachtungen in einer Stichprobe. Man kann jedoch auch umgekehrt fragen, welcher Parameter der Verteilung die Auftretenswahrscheinlichkeit der aktuellen Serie von Beobachtungen maximiert. Die Funktion, die in Abhängigkeit vom Parameter einer Verteilung der aktuellen Serie von Beobachtungen eine Wahrscheinlichkeit zuordnet, bezeichnet man als Likelihoodfunktion. Häufig werden Parameterschätzungen dann so bestimmt, daß diese Funktion für die gegebenen Beobachtungen maximiert wird.

Betrachtet man zwei logistische Regressionsmodelle M1 und M2, die sich nur dadurch unterscheiden, daß zu den Variablen in Modell M1 lediglich eine weitere Variable hinzugenommen wurde, so bezeichnet man als Deviance das Doppelte der Differenz zwischen den beiden Log-Likelihoodfunktionen. Unter der Hypothese, daß sich durch die Hinzunahme des weiteren Parameters die Modellanpassung nicht verbessert, ist die Deviance chiquadratverteilt mit einem Freiheitsgrad. Sie gibt die Möglichkeit, zwei in direkter Beziehung zueinander stehende Modelle miteinander zu vergleichen.

Betrachtet man wieder das Modell, in dem die Mortalität im Krankenhaus mit Hilfe der Variablen Alter und Lyse erklärt werden soll, so kann man sich die Frage stellen, ob es gerechtfertigt ist, einen Interaktionsterm zwischen den beiden erklärenden Variablen mit ins Modell aufzunehmen. Hier sind für das Modell mit und ohne Interaktionsterm die entsprechenden Sektionen des SAS-Ausdrucks zusammengestellt:

Ergebnis (in Auszügen): Untersuchung der Interaktion mit Alter:

Criteria for Assessing Model Fit ***(ohne Interaktionsterm)***

Criterion	Intercept Only	Intercept and Covariates	Chi-Square for Covariates
AIC	907.014	888.061	.
SC	911.505	901.533	.
-2 LOG L	905.014	882.061	22.953 with 2 DF (p=0.0001)
Score	.	.	22.450 with 2 DF (p=0.0001)

Criteria for Assessing Model Fit ***(mit Interaktionsterm)***

Criterion	Intercept Only	Intercept and Covariates	Chi-Square for Covariates
AIC	907.014	890.056	.
SC	911.505	908.019	.
-2 LOG L	905.014	882.056	22.958 with 3 DF (p=0.0001)
Score	.	.	22.453 with 3 DF (p=0.0001)

Die Deviance beträgt hier 0,005 (Differenz der beiden unterstrichenen Werte). Das ist für eine chiquadrat-verteilte Prüfgröße ein sehr kleiner Wert. Die Hinzunahme eines Interaktionsterms verbessert folglich die Modellanpassung nicht wesentlich. Die in derselben Zeile angegebenen P-Werte beziehen sich jeweils auf den Vergleich mit dem Modell, das nur den konstanten Term enthält und besagen folglich nur, daß im Vergleich zu diesem Modell eine deutliche Verbesserung erhalten wurde.

Ergebnis (in Auszügen): Interaktion mit LSB:

```
Criteria for Assessing Model Fit (ohne Interaktionsterm)

                     Intercept
           Intercept     and
Criterion    Only    Covariates Chi-Square for Covariates
AIC          891.560    871.328      .
SC           896.032    884.745      .
-2 LOG L     889.560    865.328    24.232 with 2 DF (p=0.0001)
Score           .          .       23.408 with 2 DF (p=0.0001)

Criteria for Assessing Model Fit (mit Interaktionsterm)

                     Intercept
           Intercept     and
Criterion    Only    Covariates Chi-Square for Covariates
AIC          891.560    864.984      .
SC           896.032    882.873      .
-2 LOG L     889.560    856.984    32.576 with 3 DF (p=0.0001)
Score           .          .       30.968 with 3 DF (p=0.0001)
```

Die Deviance hier beträgt: 8,344. (großer Wert). Die Modellanpassung ist also deutlich besser geworden. Für die Untersuchung solcher Fragestellungen hat *proc genmod* große Vorteile, da man hier für die Hinzunahme mehrerer Variablen schrittweise die Verbesserung der Modellanpassung überprüfen kann:

Programmcode:

```
;data eins; set a.herzmass; n=1; run
;proc genmod
;     class lyse lsb
;     model klintod/n=lyse lsb lyse*lsb / dist=bin link=logit
                                          type1
;run;
```

Ergebnis:

```
The GENMOD Procedure

              LR Statistics For Type 1 Analysis
             Source       DF   ChiSquare  Pr>Chi
             LYSE          1     12.9938  0.0003
             LSB           1     11.2382  0.0008
             LYSE*LSB      1      8.3441  0.0039
```

7. Dichotomisierung stetiger Variablen

Der Einfluß der Variablen Alter ist in den vorausgegangenen Abschnitten schon verschiedentlich untersucht worden. Die eigentlich stetige Variable ist hier nach klinischer Einschätzung dichotomisiert worden: man geht davon aus, daß Patienten, die zum Zeitpunkt des Infarkts älter als 75 Jahre sind, ein deutlich höheres Risiko haben, den

Infakrt nicht zu überleben. Deshalb ist der Einfluß der Variablen Alter über eine Indikatorvariable (ALTER75) modelliert worden.

Häufig ist im Bereich der Medizin jedoch die Dichotomisierung einer stetigen Variablen das Ziel einer Untersuchung: so müssen diagnostische Tests, die auf der Bestimmung eines stetigen Laborwertes basieren, häufig dichotomisiert werden (Diagnostische Tests haben das Ziel, eine Krankheit "nachzuweisen" bzw. das Vorliegen einer Krankheit auszuschließen. Ein positiver diagnostischer Test löst weitere Untersuchungen oder gegebenenfalls eine Behandlung aus). Da schlußendlich eine zweiwertige Entscheidung getroffen werden muß (Patient krank/ gesund), muß ein Schwellenwert ausgewählt werden.

Im Zeitalter der Kostendämpfung im Gesundheitswesen geht es häufig auch darum, zu entscheiden, ob ein Patient einer Hochrisikogruppe angehört und folglich ein längerer Aufenthalt im Akutkrankenhaus gerechtfertigt werden kann, oder ob der Patient schon zu einem früheren Zeitpunkt in eine Rehabilitationseinrichtung verlegt werden kann. Hier hinkt das Beispiel, in dem eine Entscheidung über eine Verlegung, insbesondere bei Patienten nach Reanimationsbedingungen, an der Variablen Alter festgemacht werden soll. Das Beispiel ist also nur deshalb gewählt worden, damit am selben Datensatz auch dieser Aspekt erklärt werden kann. Risikobewertungen, allerdings komplexerer Natur, werden in der Praxis schon vorgenommen.

Das Problem hier soll also sein, daß Patienten mit hohem Risiko (an der stetigen Variablen Alter bewertet) auf der Intensivstation verbleiben sollen. Betrachtet man noch einmal die Dichotomisierung bei 75 Jahren, so ergibt sich das folgende Bild:

Ergebnis (in Auszügen): Ausgabe von proc freq

```
ALTER75(Alter>75)
          KLINTOD(in der Klinik verstorben)
Frequency|
Percent  |
Row Pct  |
Col Pct  |       0|       1|  Total
---------+--------+--------+
       0 |    252 |    274 |    526
         |  38.24 |  41.58 |  79.82
         |  47.91 |  52.09 |
         |  86.30 |  74.66 |
---------+--------+--------+
       1 |     40 |     93 |    133
         |   6.07 |  14.11 |  20.18
         |  30.08 |  69.92 |
         |  13.70 |  25.34 |
---------+--------+--------+
Total         292      367      659
            44.31    55.69   100.00
```

526 Patienten werden folglich als Niedrigrisikopatienten eingestuft. Insgesamt liegen jedoch 75% der Todesfälle (als Außenkriterium der Zuordnung) in der "falschen" Gruppe. Mit *proc logistic* kann man sich neben der Schätzung der Modellparameter auch eine Klassifikationstabelle ausgeben lassen:

Programmcode:

```
;proc logistic descending data=a.herzmass
;     model klintod=alter75 / ctable
;run
```

Dabei erhält man als Modellgleichung:

logit(p)=0,0837+0,76(Alter>75)

Ergebnis (in Auszügen):

Classification Table

Prob Level	Correct Event	Correct Non-Event	Incorrect Event	Incorrect Non-Event	Percentages Correct	Sensi-tivity	Speci-ficity	False POS	False NEG
0.500	367	0	292	0	55.7	100.0	0.0	44.3	.
0.520	93	0	292	274	14.1	25.3	0.0	75.8	100.0
0.540	93	252	40	274	52.4	25.3	86.3	30.1	52.1
0.700	0	252	40	367	38.2	0.0	86.3	100.0	59.3
0.720	0	292	0	367	44.3	0.0	100.0	.	55.7

Sensitivität und Spezifität sind Begriffe, die aus der Sprache der diagnostischen Tests entlehnt sind: so bezeichnet die Sensitivität die Wahrscheinlichkeit, daß im Kollektiv bekanntermaßen erkrankter Personen ein Patient vom diagnostischen Test wirklich als krank ausgewiesen wird. Umgekehrt bezeichnet die Spezifität die Wahrscheinlichkeit, daß im Kollektiv bekanntermaßen gesunder Probanden der Test auch wirklich negativ ist. In unserer Situation entsprechen die erste und die letzte Zeile dieser Tabelle den trivialen Entscheidungen (also: alle 367+ 292 Patienten werden als Hochrisikopatienten eingestuft, die Sensitivität beträgt 100%, 44,3% der Patienten werden fälschlich als Hochrisikopatienten ausgewiesen. Die umgekehrte Entscheidung wird gefällt, wenn alle Patienten als Niedrigrisikopatienten eingestuft werden, die Spezifität beträgt 100%). logit(p) und damit auch p können hier bekanntermaßen jeweils nur zwei Werte annehmen:

Tabelle 9: Logit und Wahrscheinlichkeit

	logit(p)	p
Alter<75	0,0837	0,520
Alter>75	0,8437	0,699

Für alle dazwischenliegenden Schnittpunkte erhält man dieselbe Klassifikation. Dichotomisiert man die Variable Alter bei 65 Jahren, so erhält man ein schon wesentlich besseres Ergebnis:

Ergebnis (in Auszügen):

Classification Table

Prob Level	Correct Event	Correct Non-Event	Incorrect Event	Incorrect Non-Event	Percentages Correct	Sensi-tivity	Speci-ficity	False POS	False NEG
0.440	367	0	292	0	55.7	100.0	0.0	44.3	.
0.460	221	181	111	146	61.0	60.2	62.0	33.4	44.6
0.660	221	181	111	146	61.0	60.2	62.0	33.4	44.6
0.680	0	292	0	367	44.3	0.0	100.0	.	55.7

Anstelle den Schnittpunkt in der Variablen Alter zu verschieben, die entsprechenden Vierfeldertafeln auszurechnen und die Anzahl der korrekt klassifizierten Beobachtungen zu prüfen (Anteil der Patienten in der Niedrigrisikogruppe, die nicht verstorben sind + Anteil der Patienten in der Hochrisikogruppe, die verstorben sind), geht man so vor, daß man in der Modellgleichung verschiedene Schwellenwerte für logit(p) ansetzt. Der Vorteil

dieses Verfahrens, bei dem man einen Patienten der Hochrisikogruppe zuordnet, wenn seine Sterbewahrscheinlichkeit einen bestimmten Wert überschreitet, ist, daß man dieses Verfahren auch durchführen kann, wenn mehrere (potentiell stetige) Variablen in der Modellgleichung vorhanden sind. Dies kann man nun benutzen, um einen geeigneten Schwellenwert der Variablen Alter zu suchen:

Programmcode:

```
;proc logistic descending data=a.herzmass
;     model klintod=alter / ctable pprob=(0.1 to 0.9 by 0.1)
;run
```

Man erhält als Modellgleichung:

logit(p)=-2,7230+0,0458 Alter

Ergebnis (in Auszügen):

```
Classification Table
        Correct    Incorrect             Percentages
      ----------- ----------- ---------------------------------
 Prob        Non-        Non-         Sensi- Speci- False False
Level Event Event Event Event Correct tivity ficity  POS   NEG
-------------------------------------------------------------
0.100   367     0   292     0    55.7  100.0    0.0  44.3    .
0.200   367     1   291     0    55.8  100.0    0.3  44.2   0.0
0.300   361    22   270     6    58.1   98.4    7.5  42.8  21.4
0.400   341    63   229    26    61.3   92.9   21.6  40.2  29.2

0.500   280   130   162    87    62.2   76.3   44.5  36.7  40.1

0.600   185   203    89   182    58.9   50.4   69.5  32.5  47.3
0.700    66   259    33   301    49.3   18.0   88.7  33.3  53.8
0.800     2   292     0   365    44.6    0.5  100.0   0.0  55.6
0.900     0   292     0   367    44.3    0.0  100.0     .  55.7
```

Bei gleicher Gewichtung von falsch positiven und falsch negativen Ergebnissen kann der Youden-Index Y=1+Se+Sp benutzt werden, um den besten Schwellenwert zu finden. Er liegt bei 0.50. Der zugehörige logit ist 0, womit man für das Alter=2,723/0.0450=60,5 Jahre erhält. Wie im Folgenden noch ausgeführt wird, besteht stets die Notwendigkeit, das gefundene Ergebnis anhand eines unabhängigen Datensatzes zu validieren.

8. Allgemeine Bemerkungen zur Modellierung von Binärdaten mit logistischer Regression

Wird im Rahmen einer Untersuchung pro Proband (Untersuchungseinheit) eine Variable erhoben, die nur zwei verschiedene Werte annehmen kann, so spricht man von einer dichotomen Variablen oder einfach von Binärdaten. Beispiele gibt es (zumindest im Bereich der Medizin) zuhauf: Unter einer Therapie wird bei einem Patienten eine Heilung erzielt, bzw. der Patient kann nicht geheilt werden; es gelingt, einen chronisch kranken Patienten in einem stabilen Zustand zu halten, bzw. der Zustand des Patienten verschlechtert sich; ein Herzinfarkt-Patient überlebt den Infarkt, bzw. verstirbt.

Wie in den vorausgegangenen Abschnitten aufgezeigt, bietet die logistische Regression die Möglichkeit, den gemeinsamen Einfluß unabhängiger Variablen auf eine dichotome Zielvariable zu untersuchen. Mehrere Probleme bestehen, wenn man das Verfahren in einer realen Datensituation anwenden möchte. Sie sollen hier kurz angesprochen werden.

Die angeführten Literaturstellen sind als Einstieg in die weiterführende Beschäftigung mit dieser Thematik gedacht.

Zumindest in der Medizin besteht häufig das Problem, daß bei der Untersuchung von prognostischen Faktoren für den Eintritt eines Zielereignisses eine große Anzahl möglicher Kandidaten in Betracht gezogen werden muß. Die oben beschriebene Grenzsituation, daß ungefähr genausoviele prognostische Faktoren von Interesse sind, wie Patienten zur Untersuchung zur Verfügung stehen, wird nicht selten erreicht. Verschiedene Strategien werden durchgeführt, um ein optimales Modell zu finden[1]:

1. Alle Variablen werden gemeinsam in ein logistisches Regressionsmodell aufgenommen und dann mit Hilfe von Modellsuchstrategien untersucht.
2. Der Einfluß der unabhängigen Variablen auf die Zielvariable wird zunächst mit Hilfe von z.B. Chiquadrat-Tests in der Vierfeldertafel untersucht und nur die "signifikanten" Variablen werden zu einer gemeinsamen Bewertung in ein logistisches Regressionsmodell aufgenommen.

Beide Strategien sind nicht optimal: umso mehr Variablen in ein logistisches Regressionsmodell aufgenommen werden, desto unklarer ist, wie die Modellsuchverfahren reagieren werden. In der zweiten Situation wird unter Umständen sehr viel "auf den Daten herumgetestet", hat aber wenigstens den Vorteil, daß man sich intensiver mit seinen Daten auseinandersetzt. Man sollte jedoch dann bei der Vorauswahl diejenigen Variablen in Betracht ziehen, die einen P-Wert kleiner als 0,20 aufweisen.

Nach meiner Meinung empfieht es sich, die Anzahl der Variablen vor Beginn der Auswertung zu reduzieren, indem man versucht, die verschiedenen Variablen zu gruppieren und dann aus den einzelnen Gruppen die wichtigste Variable auszuwählen (z.B. aus verschiedenen Variablen, die auf eine schlechte Funktion des Herz-Kreislauf-Systems hindeuten, wird der Blutdruck ausgewählt). Alternativ können manchmal auch verschiedene Variablen (z.B. Reanimationsbedingung, niedriger Blutdruck bei Aufnahme, der aktuelle Infarkt ist bereits ein Reinfarkt) zu einer Variablen (z.B. Hochrisikopatient bei Aufnahme) zusammengefaßt werden. Dies bietet zusätzlich die Möglichkeit, im Rahmen von Sensitivitätsanalysen, die den Einfluß der getroffenen Annahmen auf das Ergebnis der Untersuchung darstellen sollen, weitere inhaltlich gut begründete Modelle zu untersuchen.

Es ist schwierig, wenigstens eine Daumenregel für das Verhältnis der Anzahl der Beobachtungen zur Anzahl der Einflußgrößen anzugeben, die maximal in ein logistisches Regressionsmodell aufgenommen werden sollten. Im Zusammenhang mit Cox-Regressionsmodellen zur Untersuchung von Einflußfaktoren auf die Überlebenszeit schlagen Harrel et al.[4] ein Verhältnis von 10/1 vor, wobei sich der Zähler auf die Anzahl der beobachteten Ereignisse bezieht. Altman[1] schlägt im Zusammenhang mit multipler linearer Regression vor, bei n Beobachtungen nicht mehr als $\sqrt{n}$ oder auch nur $n/10$ Einflußfaktoren in das initiale Modell aufzunehmen.

Generell besteht das Problem, daß die Ergebnisse von Regressionsanalysen mit Hilfe von Modellsuchverfahren schwierig zu interpretieren oder sogar irreführend sein können: da dieselben Daten für das Modellsuchverfahren und für die Schätzung der Parameter verwendet werden, sind diese Schätzungen verzerrt und die P-Werte sind im strengen Sinne nicht zulässig[9]. Dies führt zu der Empfehlung, die Aussagen von Regressionsmodellen stets auf unabhängigen Datensätzen zu validieren bzw. den Untersuchungsdatensatz von vorneherein in eine Test- und eine Validierungsstichprobe zu unterteilen.

Darüber hinaus kann man bei stetigen oder ordinalen Variablen eine Dichotomisierung vornehmen: wenn man mit einem Landwirt über die Milchleistung einer bestimmten Kuh sprechen will, wird der Bauer vermutlich nicht eine genaue Angabe zur Milchleistung im

letzten Monat angeben, sondern eher zum Ausdruck bringen, daß eine Kuh, seiner Ansicht nach eine "gute Kuh" und andere eben eine "schlechte Kuh" ist.

Eine Dichotomisierung stetiger Variablen sieht primär, da man ja Information verwirft, wie ein großer Informationsverlust aus. Es gibt eine Reihe von Situationen, in denen der reale Informationsverlust gering ist, die Methodik ist vielfach sehr einfach und in jedem Fall sind dichotome Variablen besser interpretierbar und kommunizierbar.

Dies zieht natürlich die Frage nach sich, wie ein Schwellenwert zur Einteilung der Einflußvariable gewählt werden soll. Vielfach ist darauf hingewiesen worden, daß die Suche nach einem optimalen Schwellenwert (z.B. indem man alle möglichen Schwellenwerte ausprobiert und dann denjenigen auswählt, bei dem der größte Einfluß auf die Zielvariable festzustellen ist) und dann die Benutzung dieses Schwellenwertes in Regressionsmodellen anhand derselben Daten, nicht korrekt ist und ebenfalls zu einer überoptimistischen Bewertung der entsprechenden Einflußvariable führt[2]. Solange nur eine einzige Einflußvariable betrachtet wird, ist es möglich, den P-Wert zu korrigieren[8], in einer praktischen Situation mit vielen Einflußgrößen ist dies jedoch von geringer Bedeutung.

Ein korrektes Vorgehen besteht darin, vor Untersuchungsbeginn den Schwellenwert z.B. aus inhaltlichen Überlegungen festzulegen. Gelegentlich wir auch empfohlen, den Median der entsprechenden Einflußgröße als Schwellenwert zu benutzen. Soll jedoch die Einteilung in zwei Kategorien inhaltlich so etwas wie die Auszeichnung einer Hochrisikogruppe bedeuten, so ist nicht einzusehen, weshalb gerade fünfzig Prozent der Patienten zur "Hochrisikogruppe" gehören sollen[9].

Prinzipiell ist es auch möglich, die Variable in ihrer ursprünglichen Form in das Modell einzubringen; dann macht man jedoch die Annahme, daß der Logit der Erfolgswahrscheinlichkeit linear von dieser Einflußgröße abhängt. Dies macht deutlich, daß man die Ergebnisse eines Regressionsmodells in jeder Stufe überprüfen muß (indem man z.B. die stetige Variable in mehrere Kategorien einteilt und dann überprüft, ob die Linearität der Logits wenigstens annähernd gegeben ist)[3].

9. SAS-Prozeduren

Logistische Regression kann mit verschiedenen SAS-Prozeduren, die ursprünglich auch aus verschiedenen methodischen Ansätzen heraus entstanden sind, durchgeführt werden. Der folgende Abschnitt soll abschließend einige Hinweise auf die Unterschiede zwischen diesen Prozeduren geben.

9.1. Datenorganisation

SAS gestattet die Verarbeitung von Datensätzen unterschiedlicher Struktur. Im einfachsten Fall liegen sie in einer sogenannten Urliste vor, in der jede Zeile den aktuell beobachteten Werten eines Patienten entspricht. Dies ist auch die Form, in der der Datensatz in den vorausgegangenen Beispielen organisiert war. Dann hat der SAS-Aufruf die folgende Form:

```
;data eins
;input pat lyse klintod
;      cards
;
      Daten (wie nebenstehend)

;run
;proc logistic descending
;     model klintod = lyse
;run;
```

Pat	Lyse	Klintod
1	1	1
2	1	1
3	1	0
4	0	1

Auch aggregierte Daten in Kontingenztafeln können verarbeitet werden. Dazu müssen Indikatoren für die Zeilen und Spaltenposition der Zellen generiert werden:

```
;data eins
;input lyse klintod anz
;      cards
;
0 0 135
0 1 223
1 0 157
1 1 144
;run
;proc logistic descending
;     model klintod = lyse
;     weight anz
;run;
```

Tabelle 10: Lyse und Mortalität

		in KH verstorben	
		nein	ja
Lyse	nein	135	223
	ja	157	144

In Publikationen wird häufig alternativ die Anzahl der Erfolge pro Therapiegruppe in der Gesamtzahl der Probanden angegeben. Dann muß der folgenden Aufruf benutzt werden:

```
;data eins
;input lyse klintod trials
;      cards
;
0 223 358
1 144 301
;run
;proc logistic descending
;   model klintod / trials = lyse
          /
;run;
```

Falls die Daten in der Urliste vorliegen, kann man sich mit dem folgenden Trick behelfen: man definiert eine Variable trials=1 im data-step und verfährt wie angegeben.

9.2. Unterschiede zwischen den Prozeduren

In diesem Abschnitt wird stichwortartig auf die Unterschiede zwischen den verschiedenen SAS-Prozeduren hingewiesen. Die Ausführungen erheben keinen Anspruch auf Vollständigkeit und für den praktischen Einsatz der verschiedenen Prozeduren wird der Leser sicherlich auf die Handbücher zurückgreifen müssen.

proc logistic: keine direkte Modellierung von Interaktionstermen, keine Deviance für den Vergleich mehrerer Modelle, kann Konfidenzintervalle für das Odds-Ratio ausrechnen, Modellsuchverfahren sind vorhanden.

proc probit:

```
;proc probit
;     class lyse alter75
;     model klintod = lyse alter75 lyse*alter75 / d=logistic
;run;
```

Bei *proc probit* bestehen dieselben Aufrufmöglichkeiten wie bei proc logistic, jedoch können Interaktionsterme direkt angegeben werden, ohne daß zunächst in einem Datenschritt die entsprechende Indikatorvariable definiert werden muß. Odds-Ratios und Konfidenzintervalle werden von proc probit nicht angegeben. Ebenso sind keine Modellsuchverfahren verfügbar.

proc genmod:

```
;data eins; set a.herzmass; n=1; run
;proc genmod
;     class lyse lsb
;     model klintod/n=lyse lsb lyse*lsb / dist=bin link=logit
```

```
                                        type1 waldci
;run;
```

Hier sind ebenfalls keine Modellsuchverfahren verfügbar, sonst ist *proc genmod* die flexibelste Prozedur. Zudem besteht die Möglichkeit, alle Ausgaben von *proc genmod* in SAS-Dateien zu schreiben, sodaß man sich z.B. Odds-Ratios und Konfidenzintervalle zu den Modellparametern ausrechnen und in der Form ausdrucken kann, wie man sie "am liebsten" in seine Auswertungsberichte übernehmen möchte.

9.3. A note of caution

Es ist wirklich am besten, wenn man seine Variablen ohne Schnörkel mit 0 / 1 für nein / ja codiert und immer mal wieder einen einfachen Fall von Hand nachrechnet. Bedauerlicherweise besteht innerhalb der SAS-Prozeduren keine offensichtliche Konsistenz , wie die Odds-Ratios gebildet werden. Dies stellt kein Problem dar, so lange man sich in einer Situation befindet, in der man sozusagen nur Textaufgaben nachrechnet, bei denen die Antwort im Anhang gegeben ist: meist genügt es dann, wenn ein Odds-Ratio größer als 1 herauskommt, den entsprechenden Kehrwert zu berechnen. Sobald man jedoch reale Probleme bearbeitet und nicht sicher ist, daß man immer genau weiß, ob die untersuchte "Therapie" auch wirklich profitabel ist, können die nachfolgend aufgeführten Inkonsistenzen zu ernsten Problemen führen. In der Situation unseres Beispiels (lyse=1, wenn die Therapie gegeben wurde, klintod=1, wenn der Patient im Krankenhaus verstorben ist) würde man zum Beispiel folgendes beobachten:

proc freq mit der *cmh*-Option berechnet dann ein Odds-Ratio kleiner 1, wenn die Therapie profitabel ist. *proc logistic* ohne die *descending*-Option und Aufruf klintod = lyse modelliert das Ereignis klintod=0, also die Wahrscheinlichkeit zu überleben. Das Odds-Ratio ist dann folglich größer als eins. Fügt man die *descending*-Option ein, so sind die Ergebnisse von *proc logistic* und *proc freq* konsistent. Mit der klintod/trials-Notation kommt stets ein Odds-Ratio kleiner als eins heraus.

proc probit modelliert klintod=1, also ist das Odds-Ratio kleiner als 1. Vergißt man jedoch lyse im class-statement, so wird ein Odds-Ratio größer als eins berechnet (SAS "denkt" dann, daß ein stetiger Confounder vorliegt). *proc genmod* berechnet nur die logits und beim Aufruf mit klintod/trials muß man OR=exp(-parameter) benutzen.

Manche dieser Unterschiede werden sicherlich verständlich, wenn man sich genauer mit der Philosophie beschäftigt, die den einzelnen Prozeduren zugrunde liegt. Sicherlich geben die angeführten Beispiele Anlaß, die oben gegebenen Empfehlungen ernst zu nehmen.

Danksagung

Der Vortrag, der dieser Ausarbeitung zugrunde gelegen hat, hat eine rege Diskussion erfahren, für die ich allen Diskutanten und insbesondere denen, die mich auf weitere Literatur aufmerksam gemacht haben, noch einmal sehr herzlich danken möchte.

Literatur

1. Altman DG. *Practical statistics for medical research*. London: Chapman and Hall, 1991;1-611
2. Altman DG, Lausen B, Sauerbrei W, Schumacher M. Commentary: Dangers of using "optimal" cutpoints in the evaluation of prognostic factors. *Journal of the National Cancer Institute* 1994;**86**:829-835.
3. Collet D. *Modelling binary data*. London: Chapman and Hall, 1991;1-369

4. Harrell FE, Lee KL, Matchar DB, Reichert TA. Regression models for prognostic prediction: advantages, problems and suggested solutions. *Cancer Treatment Reports* 1985;**69**:1071-1077.
5. Hosmer DW, Lemeshow S. *Applied logistic regression*. New York: John Wiley & Sons, 1989;1-307
6. Koch, A. and Dinkel, H. Das 60-Minuten-Herzinfarktprojekt: Biometrischer Abschlussbericht. 26, 1-93. 1996. Heidelberg, Abteilung Medizinische Biometrie.
7. Rustige J, Schiele R, Burczyk U, et al. The 60 Minutes Myocardial Infarction Project: treatment and clinical outcome of patients with acute myocardial infarction in Germany. *European Heart Journal* 1997;**18**:1438-1446.
8. Schulgen G, Lausen B, Olsen JH, Schumacher M. Outcome-orientetd cutpoints in analysis of quantitative exposures. *Amer.J.Epidemiol.* 1994;**140**:172-184.
9. Simon R, Altman DG. Statistical aspects of prognostic factor studies in oncology. *Brit.J.Cancer* 1994;**69**:979-985.