Eine SAS-Anwendungsentwicklung für ein offenes Ein-Kompartment-Modell

Hans-Peter Altenburg

Deutsches Krebsforschungszentrum
Abt. Klinische Epidemiologie
Telefon: 06221 / 422389
eMail: hp.altenburg@dkfz-heidelberg.de

Abstract

Es wird eine SAS-Anwendungsentwicklung für ein sog. offenes heteroskedastisches Ein-Kompartment-Modell vorgestellt. Im Dialog kann der Nutzer dabei eine geeignete Varianzfunktion sowie aus verschiedenen Verfahren für die Schätzung der drei (interpretierbaren) Parameter das am besten Geeignete auswählen, um eine optimale Schätzung zu erreichen. Neben dem Standardschätzverfahren Kleinste-Quadrate-Methode sind noch folgende Verfahren möglich: verallgemeinerte Kleinste-Quadrate-Methode inkl. einer winsorisierten Variante, L_1- und L_p-Norm-Schätzung und ein adaptiertes L_p-Norm-Verfahren. Insbesondere lassen sich mit den in der Anwendung verwendeten Verfahren Ausreißer in den Daten erkennen und entsprechend in der Schätzprozedur berücksichtigen. Die Anwendung wurde realisiert im wesentlichen mit Hilfe SAS/AF und den SAS-Prozeduren NLIN sowie NLP.

1. Einleitung

In der SAS-Anwendung wird ein offenes Ein-Kompartment-Modell mit drei interpretierbaren, für eine praktische Anwendung relevanten Parametern betrachtet, dessen Regressionsbeziehung für den Erwartungswert $E[y_i] = f(t, \Theta) = E[$ Konzentration zur Zeit $t\ ]$ funktional z.B. in der Form

$$f_i(t,\Theta) = (\theta_1\theta_3 / (\theta_1-\theta_2)\ )\ (\ \exp(-\theta_2 t) - \exp(-\theta_1 t)\ )$$

beschrieben werden kann. Hierbei stellen

$f_i(t, \Theta)$ die erwartete Substanzkonzentration zur Zeit t dar,

sowie die Parameter

θ_1 die Absorptionsrate,

θ_2 die Eliminationsrate, und

θ_3 kann als die physikalische Dichte ([Masse/Volumen]) interpretiert werden (oder auch als der Quotient aus aD und Volumen, wobei a den vom Gastrointestinaltrakt absorbierten Anteil der Substanz und D die zu Beginn verabreichte Dosis bezeichnet).

Der für die praktische Nutzung wichtige Bioverfügbarkeitsparameter AUC (**a**rea **u**nder the concentration **c**urve) kann für dieses Modell dann z.B. durch den Quotienten θ_2 / θ_3 berechnet werden. Mit $\Theta=(\theta_1, \theta_2, \theta_3)^T$ bezeichnen wir den Vektor der zu schätzenden strukturellen Parameter. Die beobachtete Substanzkonzentration sei im folgenden mit $y_i(t)$ bezeichnet.

Viele Anwendungen zeigen hinsichtlich der Varianz keine Homogenität mehr, so daß bei der Datenanpassung u.U. ein heteroskedastisches Varianzmodell mit den Daten angepaßt werden muß. Ein sehr allgemeines Varianzmodell könnte etwa die Form

$$VAR[\ y\ ] = \Lambda(\ f(\ t, \theta\ ), t, \kappa\ ) = \sigma^2\ f(\ t, \theta\ )^{\kappa}$$

haben, d.h. die Varianz ist im wesentlichen die Potenz des Erwartungswertes mit einem zusätzlich zu schätzenden Parameter. Hierbei haben unterschiedliche Konstellationen der Varianzparameter bestimmte praktische Konsequenzen, etwa

$\kappa=0$: konstante Varianz,

$\kappa=1$: konstanter Variationskoeffzient,

σ=1 und κ=1: Standard-Poissonmodell oder
κ>1: eine Überstreuung.
Als alternative Varianzmodelle sind Funktionen der Form

$$\Lambda(f(t, \theta), t, \kappa) = \sigma^2 \alpha f(t, \theta)^{\kappa} = \sigma^2 (\alpha + \gamma f(t, \theta)^{\kappa})$$

vorstellbar, wobei α und γ weitere zusätzlich zu schätzende Parameter sind.

Unter der Annahme, daß das Modell korrekt ist, stellt sich das Schätzproblem als nichtlineares heteroskedastisches Regressionsmodell dar, mit dem Erwartungswert: $E[y_i] = f(x_i , \Theta)$ $i=1, \ldots, n$, als einer nichtlinearen Regressionsfunktion $f(.,.)$, den x_i als unabhängige Variable (z.B. Zeit, Dosis, ...), den y_i als abhängigen Responsen, die zu einer diagonalen Kovarianzmatrix führen:

$$VAR[y] = \sigma^2 \Lambda(f(x_i , \theta), x_i , \kappa), i=1, \ldots, n$$

und κ als dem zusätzlich zu schätzendem (Varianz-) Parameter.

2. Schätzverfahren

Folgende Schätzverfahren wurden in der Anwendung realisiert:

- Gewöhnliche Kleinste Quadrate Methode (OLS),
- Generalisierte Kleinste Quadrate Methode (GLS),
- Winsorisiertes GLS-Verfahren (WGLS),
- L_1-Norm-Verfahren (L1, LAV),
- Lp-Norm-Verfahren (Lp) und
- adaptiertes Lp-Norm-Verfahren (LpA).

Alle Verfahren wurden mit Hilfe von SAS realisiert:
Die gewöhnliche Kleinste-Quadrate-Methode (OLS) kann z.B. direkt über die Prozedur NLIN durchgeführt werden,

$$\sum_i (y_i - f(x_i , \Theta))^2 \rightarrow \min$$

ebenso eine gewichtete Schätzung bei bekannter Varianz

$$\sum_i w_i (y_i - f(x_i , \Theta))^2 \rightarrow \min$$

mit den Gewichten

$$w_i = 1 / \Lambda(f(x_i , \theta), x_i , \kappa).$$

Eine empfohlene Standardschätzprozedur bei unbekannten, noch zusätzlich zu schätzenden Varianzparametern wäre die verallgemeinerte Kleinste-Quadrate Schätzung (GLS), die iterativ abläuft und mit Hilfe der SAS-Prozedur NLIN und über ein SAS-Makro durchgeführt werden kann:
0. Suche Anfangswerte für die Parameter Θ und κ (etwa über ein OLS-Verfahren).

Iteriere danach folgende zwei Schritte solange bis :
1. Schätze den bzw. die Varianzparameter κ (Θ fix) über eine gewichtete OLS, und
2. schätze die Parameter Θ der Regressionsfunktion (κ fix) über eine gewichtete OLS.

Das winsorisierte GLS-Verfahren ist eine robuste Variante des GLS-Verfahrens. Die zugrundeliegende Idee ist Ausreißerdiagnostik mit robuster Schätzung zu kombinieren. Auch dieses Verfahren läßt sich im wesentlichen iterativ über die Prozedur NLIN, ein SAS-Makro und zwischengeschaltete Daten-Steps realisieren, wobei sich der Winsorisierungsteil grob mit folgender Verfahrensweise beschreiben läßt:
Winsorisierung (j bezeichnet dabei den Iterationsschritt):
ersetze $y_i^{(j)} \Rightarrow y_i^*$ falls $r_i^{(j-1)} > k_2 \sqrt{(m/n)}$

wobei
$y_i^* = y_i^{(j-1)} + \text{sign}(r_i^{(j)*})\, k_1\, s_{r,i}$
und
$r_i^{(j)}$ = approximiertes Deletionresiduum (=Kennzahl aus der Ausreißerdiagnostik)
$r_i^{(j)*}$ = normales Residuum
$s_{r,i}$ = zugehöriger Standardfehler des normalen Residuum.

Die L_1- Norm bzw. L_p-Norm-Schätz-Verfahren werden mit Hilfe der SAS-Prozedur NLP durchgeführt. Diese Prozedur ist Bestandteil des SAS-Produktes SAS/OR und kann nichtlineare Optimierungsaufgaben lösen helfen. Es läßt sich zeigen, daß die für die L_1- bzw. L_p-Norm-Verfahren erforderliche Minimierungsaufgabe

$$\sum_i | y_i - f(x_i, \theta) |^p \rightarrow \min$$

als nichtlineares Programmierungsproblem formuliert werden kann, was dann über die SAS-Prozedur NLP gelöst wird. In dieser Prozedur werden in der SAS-Anwendung spezielle Varianten und Algorithmen verwendet, wie z.B. die Nelder-Mead-Simplex Variante von Powell ("Constrained Optimization by linear Approximation"), welche etwa den Vorteil besitzt, daß keine Ableitungen der Funktion mehr bestimmt werden müssen. Ein genereller Nachteil der L_p-Norm-Verfahren ist, einen geeigneten Wert für den Normparameter p zu finden. Als Ausweg bietet sich das adaptierte Lp-Norm-Verfahren, was ebenfalls iterativ nach dem optimalen Normparameter p sucht. Konkret wird die Pearson-Wölbung b_2 der Residuen aus dem letzten Iterationsschritt mit Hilfe von Momenten aus der Prozedur UNIVARIATE bestimmt und daraus der neue Normparameter (nach Gonin and Money (1989) über die Beziehung

$$p = 9/\, b_2 + 1$$

berechnet.

3. Vergleich der Verfahren

Wenn keine Ausreißer vorliegen liefern alle Verfahren vergleichbare Schätzwerte. Liegt nur eine moderate Anzahl von Ausreißern vor, so sind das winsorisierte GLS- bzw. L_1- oder L_p-Norm-Verfahren zu bevorzugen, wobei bei dem L_p-Norm-Verfahren das Problem der Wahl eines geeigneten Normparameters auftaucht, u.U. kommt manchmal auch ein Normparameter kleiner eins in Frage. Beim adaptierten L_p-Norm-Verfahren ist die verwendete Beziehung Wölbung $\leftrightarrow$ Normparameter p manchmal ungeeignet und kann deshalb zu Problemen führen.

4. Weitere Möglichkeiten der SAS-Anwendung

Neben den bereits erwähnten Verfahren für die Parameterschätzung erlaubt die Anwendung auch potentielle Ausreißer zu identifizieren. Hierfür stehen drei Hauptgruppen mit entsprechend grafischen Möglichkeiten zur Verfügung:

- (i) *Leverage-Maße* (Hebelpunkt-Maße), die helfen Ausreißer in Richtung der unabhängigen Variablen zu erkennen.
- (ii) *Residual-Maße*, die zur Identifizierung von Ausreißern in Richtung der abhängigen Variablen oder zum Erkennen von systematischen Abweichungen von den Modellannahmen dienen, wie z.B. mögliche Trends erkennen helfen, die Darstellung der Verteilung der Fehler (Normalverteilung), Änderung in der Variabilität, etc.
- (iii) *Einfluß-Maße* (Influence-Measures), welche die Residual- und Leverage-Maße kombinieren.

Zur ersten Gruppe gehören etwa die Leverage-Maße, welche auf den Elementen der Hat-Matrix beruhen,

$$h_{ii} \text{ bzw. } h_{ii}/(1- h_{ii})$$

welche aber in vielen nichtlinearen Anwendungen eher als ungeeignet anzusehen sind, da Ausreißer in der Regel diese Werte auch stark beeinflussen. Benötigt werden hier mehr robuste Varianten! Für diese Kennzahlen bietet SAS in der Prozedur NLIN eine Option im OUTPUT-Statement (Option: H=*name*).

Residuen gehören mit den entsprechenden graphischen Darstellungen zu den Standardmethoden für die Beurteilung der Güte einer Anpassung. Sowohl das Standard-Residuum „aktueller Wert minus geschätzter Wert" als auch Varianten davon können über die Prozedur NLIN im OUTPUT-Statement angesprochen werden und wurden auch in der Anwendung entsprechend berücksichtigt über die Option:

R=*name* / Residuum
STDR=*name* / Standardfehler
STUDENT=*name* / =R/STDR (studentisiertes Residuum).

Drei Haupttypen von Plots können in der Anwendung angefordert werden:

- Plot der Residuen gegen die Werte der unabhängigen Variablen, zum Erkennen von Abweichungen vom Modell oder von Ausreißern in Richtung der abhängigen Variablen,
- Plots von Residuen gegen beobachtete oder geschätzte Werte der abhängigen Variablen, helfen Veränderungen in der Varianz zu erkennen, und
- Plot der geordneten Residuen gegen die Perzentile einer vorgegebenen Wahrscheinlichkeitsverteilung (z.B. Normal-Probability-Plot), um das Vorliegen einer bestimmten Verteilung für die Residuen zu prüfen (etwa Normalverteilung).

Allerdings sind einige dieser Residuuen, wie etwa das normale (nicht standardisierte) Residuum, nicht sonderlich robust. Ähnlich auch die Standardisierung dieser Residuen mit der Standardabweichung aus dem Varianzterm σ^2_G.

Falls das Varianzmodell adäquat gewählt wurde, sollte ein Plot dieser standardisierten Residuen gegen die geschätzten Werte ein Band mit konstanter Breite zeigen. Unter Umständen besser geeignet sind Varianten der studentisierten Residuen, etwa in der Form $r_{i,\,stud}^{\,*} = r_i / \sqrt{1 - h_{ii}}$.

Ein Vorteil dieser studentisierten Residuen beim linearen Modell liegt darin, daß alle die gleiche Varianz haben, falls das Modell korrekt ist. Sie sind deshalb gut für Normal-Probability-Plots oder die Prüfung der Varianzhomogenität geeignet. Die besten Erfahrungen mit Ausreißerdiagnostik wurden bei den approximierten Deletion-Residuen und Cook-Statistiken gemacht.

Mit Blick auf das notwendige Vorgehen bei einem konkreten Problem wurde die SAS Anwendung so entwickelt, daß alle erforderlichen Schritte von einem mit SAS ungeübten Nutzer durchgeführt werden können:

- Graphische Darstellungen / Punktwolken für die beobachteten Werte,
- schätze die Modellparameter mit einem der implementieren Standardverfahren,
- Darstellung der Residuen / Deletion-Residuen zur Entdeckung von extremen Werten,
- Entscheidung für ein robustes Verfahren / Auswahl der Tuning-Konstanten bzw. Startwerte für die Normkonstante *p*.

Bei der Realisierung mit SAS wurden folgende Methoden verwendet:

- Oberfläche / Frames usw. mit SAS / AF,
- Schätzverfahren über SAS-Makros und die SAS/STAT-Prozeduren NLIN bzw. die OR-Prozedur NLP (etwa für die L_p-Norm-Verfahren),
- Graphische Darstellungen mit Hilfe der SAS-Prozeduren PLOT bzw. GPLOT,
- Die Kennzahlen der Extremwert-Diagnostiken über zwischengeschaltete DATA-Steps, in denen z. B. approximative Deletion-Residuen bestimmt werden.

Literatur:

1. Giltinan, D.M. and Ruppert, D.: Fitting heteroskedastic regression models to individual pharmacokinetic data using standard statistical software. *Journal of Pharmacokinetics and Biopharmaceutics* 17, 1989, 601-614
2. Gonin, R. and Money, A.H. (1989): *Nonlinear Lp-Norm estimation*. M. Dekker, New York
3. Ross, G.J.S.: *Nonlinear Estimation*. Springer, New York 1990
4. SAS Institute Inc. (1989): SAS/STAT User's Guide. Version 6 Fourth Edition Volume 1 und Volume 2
5. Seber, G.A.F. and Wild, C.J.: *Nonlinear Regression*. Wiley, New York 1989