Reinhold Hatzinger

Quasi-Likelihood Methoden zur Analyse von unabhängigen und abhängigen Beobachtungen

Working Paper

Quasi-Likelihood Methoden zur Analyse von unabhängigen und abhängigen Beobachtungen

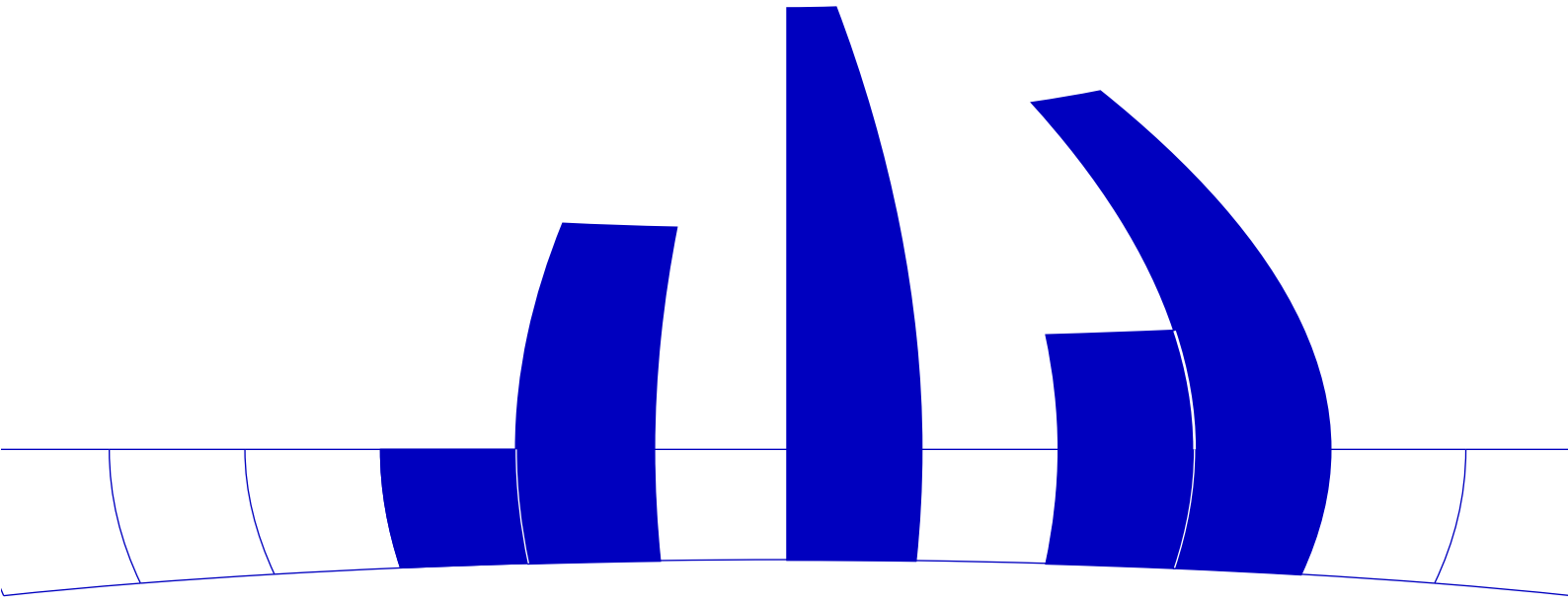
Reinhold Hatzinger

Institut für Statistik
Wirtschaftsuniversität Wien

Forschungsberichte

Bericht 13
1991

http://statmath.wu-wien.ac.at/



Quasi-Likelihood Methoden zur Analyse von unabhängigen und abhängigen Beobachtungen

Reinhold Hatzinger
Institut für Statistik, Wirtschaftsuniversität Wien
Augasse 2 – 6, A-1090 Wien

Zusammenfassung

Ausgehend vom klassischen linearen Modell werden Regressionsmethoden für Datenstrukturen dargestellt, bei denen die Standardannahmen (Unabhängigkeit, normalverteilte Fehler und konstante Varianz) nicht erfüllt sind. Läßt man die Responsevariable aus einer Exponentialfamilie zu, so erhält man die Klasse generalisierter linearer Modelle (GLM). Dies erlaubt, den Erwartungswert von verschiedensten stetigen und diskreten Responsevariablen (z.B. Anteile, Häufigkeiten, etc.) über eine fixe Kovariatenstruktur zu modellieren. Hebt man zusätzlich die Notwendigkeit auf, eine Verteilung aus Exponentialfamilien spezifizieren zu müssen, erhält man Quasi-Likelihood Modelle, bei denen nur mehr eine Beziehung zwischen Erwartungswert und Varianz festgelegt werden muß. Die Berücksichtigung einer Korrelationsstruktur führt zu verallgemeinerten Schätzgleichungen, d.h. es können auch Longitudinaldaten ohne besondere Verteilungsannahmen analysiert werden. Ziel der Arbeit ist es, diese Methoden und ihre statistischen Eigenschaften vorzustellen und anhand eines Beispiels (Überdispersion bei wiederholt gemessenen binomialen Anteilen) ihre Bedeutung in der biometrischen Praxis zu illustrieren.

Schlüsselworte: Regressionsmethoden; Generalisierte lineare Modelle; Quasi Likelihood; Überdispersion; Verallgemeinerte Schätzgleichungen; Longitudinaldaten

1 Einleitung

Ausgangspunkt für die in diesem Beitrag vorgestellten Methoden ist das klassische lineare Modell für einen Responsevariable Y

$$y = X\beta + \varepsilon \tag{1}$$

mit einer $n \times p$ Matrix erklärender Variablen X, einem $p \times 1$ Vektor unbekannter Parameter β sowie Störgrößen ε. NELDER und WEDDERBURN führten 1972 die generalisierten linearen

Modelle (GLM) als eine Erweiterung dieses klassischen linearen Modells ein, wobei eine Vielzahl von Regressionsmethoden für unterschiedliche Datentypen vereinheitlicht wurde. Die Anwendbarkeit von (1) wird hierbei durch Aufheben der Annahme additiver Fehler wesentlich erweitert. Kann im linearen Fall die Dichte von Y

$$f_Y(y) = f_\epsilon(y - x'\beta)$$

geschrieben werden, so ist die verallgemeinerte Form gegeben durch

$$f_Y(y) = f(y; x'\beta), \tag{2}$$

wobei x' und β den linearen Prädiktor $\eta = x'\beta$ konstituieren. Existiert der Erwartungswert $E(Y) = \mu$, dann wird μ bestimmt durch η, d.h. $g(\mu) = \eta$ und $g(\mu)$ wird Linkfunktion genannt. Die Dichte in (2) kann jede geeignete Dichte oder Wahrscheinlichkeitsfunktion sein, allerdings ist es aus verschiedenen noch zu erläuternden Gründen vorteilhaft Exponentialfamilien zu verwenden, die hier die gleiche Rolle spielen wie die Normalverteilung im klassischen linearen Modell. Verwendet man Likelihood-Methoden zur Schätzung der Parameter für eine geeignete lineare Exponentialfamilie, so haben diese Eigenschaften analog zu Kleinst-Quadrate Schätzern im linearen Modell. (Die in dieser Arbeit gebene Darstellung folgt im wesentlich FIRTH (1991), MCCULLAGH und NELDER (1989), sowie LIANG und ZEGER (1986).)

2 Generalisierte lineare Modelle

Im Unterschied zum klassischen linearen Modell, in dem $\mu = \eta$, d.h. daß die Funktion $E(\mu) = \mu(\beta) = \eta$ linear in den Parametern β ist, hat ein GLM die Form

$$\mu = g^{-1}(\sum_{j=1}^{p} x_j\beta_j) \quad .$$

$\beta_1, \ldots, \beta_p$ sind unbekannte Parameter, $x_1, \ldots, x_p$ sind bekannte Konstanten, die in Beziehung zur Responsevariable Y stehen. Die x_j können quantitative Variablen, wie etwa Blutdruck, oder Indikatorvariablen sein, die die Stufen einer qualitativen Variable repräsentieren. Verallgemeinerte lineare Modelle sind also selbst nicht linear, allerdings bestimmt die Linkfunktion $g(\cdot)$, die streng monoton sein muß, die Skala auf der Linearität angenommen wird. Überdies ist die Wahl von $g(\cdot)$ durch den Wertebereich von μ eingeengt. Sind $\beta_1, \ldots, \beta_p$ nicht beschränkt, kann $g(\cdot)$ jeden Wert im Intervall $(-\infty, \infty)$ annehmen. Sind z.B. Häufigkeiten als Response Y festgelegt, dann wird $g(\cdot)$ das Intervall $[0, \infty)$ auf die gesamte reelle Achse abbilden. Obwohl die Linkfunktion unter diesen milden Annahmen frei wählbar ist, ist es dennoch sinnvoll diese Klasse noch weiter einzuschränken. Darauf wird in Kap. 2.2 eingegangen.

2.1 Exponentialfamilien

Einige der wichtigsten Familien statistischer Verteilungen haben eine Likelihoodfunktion für eine einzelne Beobachtung y_i

$$f(y_i;\theta_i,) = \exp\{(\theta_i y_i - b(\theta_i))/\phi + c(y_i,\phi)\}, \tag{3}$$

wobei die Funktionen $b(\cdot)$ und $c(\cdot)$ bekannt sind. Ist überdies ϕ, der sogenannte Dispersionsparameter, bekannt, so ist (3) eine lineare Exponentialfamilie, die durch den natürlichen oder kanonischen Parameter θ gesteuert wird. 'Linear' wird verwendet um anzudeuten, daß die minimal suffizienten Statistiken aus einer Stichprobe linear in Y sind. (Ist ϕ unbekannt, so spricht man von 'exponential dispersion models'.)

Lineare Exponentialfamilien beinhalten unter anderem folgende Verteilungen für Y:

Verteilung	Erwartungswert	Varianz	Bemerkung
Normal	θ	ϕ	—
Poisson	e^θ	e^θ	$\phi = 1$
Gamma	$-1/\theta$	ϕ/θ	ϕ ist Kehrwert des Gammaindex
Binomial	$\frac{e^\theta}{1+e^\theta}$	$\phi\frac{e^\theta}{(1+e^\theta)^2}$	ϕ ...Anzahl der Versuche y ...Anzahl der Erfolge

Einige elementare Eigenschaften von linearen Exponentialfamilien folgen aus den Identitäten:

$$E\Big(\frac{\partial l}{\partial \theta}\Big) = 0 \tag{4}$$

$$-E\Big(\frac{\partial^2 l}{\partial \theta^2}\Big) = Var\Big(\frac{\partial l}{\partial \theta}\Big) \tag{5}$$

mit l als der logarithmierten Likelihood. Angewandt auf (3) ergibt sich

$$E(Y) = b'(\theta) = \mu(\beta) \quad ,$$

sowie

$$\mathrm{Var}(Y) = \phi b''(\theta) = \phi V(\mu) \quad .$$

Durch $V(\mu)$, die sogenannte Varianzfunktion, werden lineare Exponentialfamilien charakterisiert und haben eine wesentliche Funktion bei der Schätzung der Parameter β. Einige Beispiele sind:

Verteilung	Varianzfunktion
Normal	$V(\mu) = 1$
Poisson	$V(\mu) = \mu$
Gamma	$V(\mu) = \mu^2$
Binomial	$V(\mu) = \mu(1-\mu)$

2.2 Suffizienz und die kanonische Linkfunktion

Seien $y_1, \ldots, y_n$ n unabhängige Realisationen von Zufallsvariablen $Y_1, \ldots, Y_n$ mit jedem Y_i aus einer Exponentialfamilie mit Parameter θ_i und ϕ_i, dann ist die logarithmierte Likelihood für die Stichprobe

$$l = \sum_{i=1}^{n} \{(\theta_i y_i - b(\theta_i))/\phi_i + c(y_i, \phi_i)\} \quad . \tag{6}$$

Spezifiziert man in (6) ein GLM durch

$$g(\mu_i) = g(b'(\theta_i)) = \sum_{j=1}^{p} x_{ij}\beta_j \qquad i = 1, \ldots, n,$$

dann kann die Likelihood für die Regressionsparameter $\beta_1, \ldots, \beta_p$ algebraisch relativ kompliziert werden. Eine wesentliche Vereinfachung ergibt sich aber im Spezialfall $g(\cdot) = 1/b'(\cdot)$, sodaß $g(\mu_i) = \theta_i$. Dann wird die logarithmierte Likelihood zu

$$l = \sum_{j=1}^{p} \beta_j \sum_{i=1}^{n} \frac{y_i x_{ij}}{\phi_i} - \sum_{i=1}^{n} \left\{ \frac{b(\theta_i)}{\phi_i} - c(y_i, \phi_i) \right\}.$$

Sind überdies die ϕ_i bekannt, leiten sich die minimal suffizienten Statistiken aus $\sum_{i=1}^{n} y_i x_{ij}/\phi_i$ für $j = 1, \ldots, n$ ab. Die spezielle Linkfunktion $g(\cdot) = 1/b'(\cdot)$, die diese Vereinfachung erlaubt, wird kanonische Linkfunktion genannt, wobei die kanonische Linkfunktion und die Varianzfunktion durch $V(\mu) = 1/g'(\mu)$ in Beziehung stehen. Einige Beispiele hierfür sind:

Verteilung	Linkfunktion
Normal	$g(\mu) = \mu$
Poisson	$g(\mu) = \ln \mu$
Gamma	$g(\mu) = -\mu^{-1}$
Binomial	$g(\mu) = \ln(\mu/(1-\mu))$

2.3 Schätzen in GLMs

Die interessierenden Parameter werden mittels Maximum Likelihood Methode (ML–Methode) geschätzt. Differenzieren der logarithmierten Likelihood nach β_j liefert die Likelihood Schätzgleichungen

$$\sum_{i=1}^{n} \frac{y_i - \mu_i}{\phi_i V(\mu)} \cdot \frac{\partial \mu_i}{\partial \beta_j} = 0, \quad j = 1, \ldots, p \tag{7}$$

die im Falle eines GLMs zu

$$\sum_{i=1}^{n} \frac{y_i - \mu_i}{\phi_i V(\mu)} \cdot \frac{x_{ij}}{g'(\mu_i)} = 0, \quad j = 1, \ldots, p \tag{8}$$

werden. Die Gleichungen (8) hängen von den unter Umständen unbekannten $\phi_1, \ldots, \phi_n$ ab. In vielen wichtigen Anwendungen ist aber $\phi_i = a_i \phi$, mit bekannten Konstanten a_i und einem einzelnen Dispersionsparameter ϕ. Dann werden die Schätzgleichungen zu

$$\sum_{i=1}^{n} \frac{y_i - \mu_i}{a_i V(\mu)} \cdot \frac{x_{ij}}{g'(\mu_i)} = 0, \quad j = 1, \ldots, p,$$

und sind nicht mehr vom (möglicherweise) unbekannten Dispersionsparameter ϕ abhängig.

Beispiel. Seien die Zufallsvariablen $Y_1, \ldots, Y_n$ binomialverteilt, $Y_i \sim B(m_i, \pi_i)$, sodaß $\mu_i = m_i \pi_i$ und $V(\mu_i) = m_i \pi_i (1 - \pi_i)$, dann ist die logarithmierte Likelihood

$$l(\pi_i; y_i) = \sum_{i=1}^{n} \Big(y_i \ln \big(\frac{\pi_i}{1 - \pi_i}\big) + m_i \ln(1 - \pi_i) \Big).$$

Die Schätzgleichungen erhält man aus

$$\frac{\partial l}{\partial \beta_j} = \frac{\partial l}{\partial \pi_i} \cdot \frac{d\pi_i}{d\eta} \cdot \frac{\partial \eta}{\partial \beta_j}.$$

Im konkreten Fall ist dies:

$$\frac{\partial l}{\partial \pi_i} = y_i \frac{1}{\pi_i} + \frac{y_i}{1 - \pi_i} - m_i \frac{1}{1 - \pi_i} = \frac{y_i - m_i \pi_i}{\pi_i (1 - \pi_i)}, \tag{9}$$

$$g(\pi_i) = \eta_i = \ln \frac{\pi_i}{1 - \pi_i} = \sum_{j=1}^{p} x_{ij} \beta_j, \tag{10}$$

$$\frac{d\eta_i}{d\pi_i} = \frac{d}{d\pi} \ln \frac{\pi_i}{1 - \pi_i} = \frac{1}{\pi_i (1 - \pi_i)}, \tag{11}$$

und

$$\frac{\partial \pi_i}{\partial \beta_j} = \frac{d\pi_i}{d\eta_i} \cdot \frac{\partial \eta_i}{\partial \beta_j} = \left(\frac{d\eta_i}{d\pi_i}\right)^{-1} x_{ij} = \pi_i(1-\pi_i)x_{ij}. \tag{12}$$

Setzt man (9) – (12) zusammen, erhält man

$$\sum_{i=1}^{n}(y_i - m_i\pi_i)x_{ij} = 0.$$

In diesem Beispiel sind also die $a_i = 1/m_i$ und $\phi = 1$.

Allgemein vereinfachen sich für die kanonische Linkfunktion $g(\cdot)$ die Schätzgleichungen zu

$$\sum_{i=1}^{n} \frac{y_i x_{ij}}{a_i} = \sum_{i=1}^{n} \frac{\mu_i x_{ij}}{a_i} \quad j = 1, \ldots, p,$$

d.h. die gemeinsam suffizienten Statistiken werden ihren Erwartungswerten gleichgesetzt.

Mit Ausnahme des linearen Modells mit konstanter Varianz, $V(\mu) = 1$ und $g(\mu) = \mu$, wo ML für die Normalverteilungsfamilie der gewichteten Kleinst-Quadrate Schätzung entspricht, gibt es keine expliziten Lösungen für (8). Im Spezialfall des linearen Modells erhält man den Lösungsvektor durch

$$\hat{\beta} = (X'WX)^{-1}X'Wy,$$

mit X als Matrix erklärender Variablen und $W = diag\{1/a_i\}$ als Diagonalmatrix mit bekannten Gewichten. Die Existenz einer expliziten Lösung in diesem Spezialfall legt eine Lösungsmethode für den allgemeinen Fall nahe. Betrachtet man

$$z_i = \eta_i + (y_i - \mu_i)g'(\mu_i),$$

dann ist $E(Z_i) = \eta_i = \sum_{j=1}^{p} x_{ij}\beta_j$. Wären also die z_i bekannt, könnten die $\beta_1, \ldots, \beta_p$ mittels gewichteter Kleinst-Quadrate Methoden geschätzt werden, mit Gewichten als Kehrwert von

$$\mathrm{Var}(Z_i) = \{g'(\mu_i)\}^2 a_i V(\mu_i) \quad .$$

In der Praxis sind die $z_1, \ldots, z_n$ unbekannt, da die η_i bzw. die μ_i unbekannt sind. Es bietet sich aber folgende iterative Prozedur an.

1. Man beginne mit Startwerten $\hat{\mu}_i^{(0)} = y_i$ und $\hat{\eta}_i^{(0)} = g(\hat{\mu}_i^{(0)})$ für Erwartungswert und linearen Prädiktor. (Bei gewissen Linkfunktionen, z.B. $g(\mu) = \ln \mu$ muß darauf geachtet werden, daß $y_i > 0$. Dies erreicht man etwa durch die Adjustierung $\hat{\mu}_i^{(0)} = \max\{y_i, \varepsilon\}$, mit kleinem positiven ε.)

2. Gegeben $\hat{\mu}_i^{(t)}$ und $\hat{\eta}_i^{(t)}$, berechnet man die adjustierte abhängige Variable

$$\hat{z}_i^{(t)} = \hat{\eta}_i^{(t)} + (y_i - \hat{\mu}_i^{(t)})g'(\hat{\mu}_i^{(t)})$$

mit iterativem Gewicht

$$\hat{w}_i^{(t)} = \frac{1}{a_i V(\hat{\mu}_i^{(t)})\{g'(\hat{\mu}_i^{(t)})\}^2}, \quad i = 1, \ldots, n.$$

3. Im $t+1$-tem Schritt erhält man $\beta^{(t+1)}$ mittels gewichteter Kleinst–Quadrate Schätzung

$$\hat{\beta}^{(t+1)} = (X'W^{(t)}X)^{-1}X'W^{(t)}z^{(t)},$$

mit $W^{(t)} = diag\{w_i^{(t)}\}$. Danach definiert man $\hat{\eta}_i^{(t+1)} = X\hat{\beta}^{(t+1)}$ und $\hat{\mu}_i^{(t+1)} = g^{-1}(\hat{\mu}_i^{(t+1)})$.

4. Schritte 2) und 3) werden solange wiederholt, bis ein angemessenes Konvergenzkriterium erfüllt ist.

Diese Prozedur wird iterierte gewichtete Kleinst-Quadrate Schätzung (*iterative weighted least squares* - IWLS) genannt. Dieses Verfahren entspricht im Falle kanonischer Linkfunktion der Newton-Raphson Methode, allgemeiner ist es die Fisher Scoring Methode. Existenz und Eindeutigkeit der Lösungen des Gleichungssystems (7) diskutiert WEDDERBURN (1976). Für die Praxis empfiehlt sich das Programmpaket GLIM (PAYNE, 1986) das speziell zur Berechnung von GLMs konzipiert wurde.

Hat man einen Lösungsvektor gefunden, dann sind die Schätzer für β konsistent, asymptotisch normal und asymptotisch effizient mit einer approximativen Normalverteilung $N_p(\beta, i^{-1})$. $i = i_\beta$ ist die Informationsmatrix mit Elementen

$$\{i_\beta\}_{jk} = \sum_{i=1}^{n} \frac{x_{ij}x_{ik}}{\phi a_i V(\mu_i)\{g'(\mu)\}^2},$$

d.h. $i_\beta = \phi^{-1}X'WX$ mit $W = diag\{w_i\}$ und

$$w_i = \frac{1}{a_i V(\mu_i)\{g'(\mu)\}^2}.$$

Die geschätzten Standardfehler für $\hat{\beta}$ ergeben sich aus der Wurzel der Diagonalelemente von

$$\mathrm{Cov}(\hat{\beta}) = \phi(X'WX)^{-1},$$

wobei $(X'WX)^{-1}$ ein Nebenprodukt der letzten IWLS-Iteration ist.

Ist ϕ unbekannt, wird ein Schätzer $\hat{\phi}$ zu Berechnung der Standardfehler der $\hat{\beta}$ benötigt. Prinzipiell ist es möglich ϕ mittels ML zu schätzen. In der Praxis ist es aber meist einfacher, einen Momenten-Schätzer zu verwenden. Falls $\beta_1, \ldots, \beta_p$ bekannt sind, ist eine erwartungstreue Schätzfunktion für ϕ durch

$$\hat{\phi} = \frac{\mathrm{Var}(Y_i)}{a_i V(\mu)} = \frac{1}{n} \sum_{i=1}^{n} \frac{(y_i - \mu_i)^2}{a_i V(\mu)}$$

gegeben. Da $\beta_1, \ldots, \beta_p$ geschätzt werden, verwendet man in Analogie zum klassischen linearen Modell einen um die Freiheitsgrade korrigierten erwartungstreuen und konsistenten Schätzer

$$\tilde{\phi} = \frac{1}{n-p} \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{a_i V(\hat{\mu})}.$$

(Eine andere Methode basiert auf 'modified profile likelihoods', JØRGENSEN,1987).

2.4 Testen von Hypothesen

Eine spezielle Wahl der Matrix der erklärenden Variablen X, die meist aus einer größeren Menge von interessierenden Kovariaten getroffen wird, definiert die zu prüfenden Hypothesen, d.h. durch die Aufnahme gewisser Variablen in X wird ein bestimmtes Modell festgelegt. Hierbei geht es um die Balance zwischen Sparsamkeit und möglichst guter Modellanpassung. Zur Lösung dieses Problems werden üblicherweise Likelihood-Ratio Tests herangezogen.

Seien X_A und X_B zwei verschiedene Auswahlen von X, wobei diese zwei hierarchisch geordnete Modelle spezifizieren, $X_A < X_B$. Anders ausgedrückt: alle Spaltenvektoren von X_A sind im linearen Raum, der von X_B aufgespannt wird, enthalten. Dann muß Modell B mindestens so gut zu den Daten passen wie Modell A. Die Verbesserung der Anpassung kann relativ zur hinzugekommenen Komplexität von Modell B durch den Test der Nullhypothese: Modell A gegen die Alternativhypothese: Modell B geprüft werden. Sei der Rang $rg(X_B) = p_B$ und der Rang $rg(X_A) = p_A$, dann ist die verallgemeinerte LR–Statistik

$$\Lambda = 2\{l(y; \hat{\mu}^{(B)}, \phi) - l(y; \hat{\mu}^{(A)}, \phi)\} \tag{13}$$

unter Modell A approximativ χ^2–verteilt mit $df = p_B - p_A$. Ist diese Statistik signifikant, dann wird der zusätzliche Beitrag von Modell B als relevant erachtet.

Verallgemeinert spielt die Quantität

$$2\phi\{l(y; y, \phi) - l(y; \hat{\mu}, \phi)\} = D(y; \hat{\mu})$$

die gleiche Rolle, die im klassischen Modell von der Fehlerquadratsumme (RSS) gespielt wird. Im speziellen kann Λ in (13) als

$$\{D(y; \hat{\mu}^{(B)}) - D(y; \hat{\mu}^{(A)})\}/\phi$$

geschrieben werden. Die sogenannte Devianz $D(y;\hat{\mu})$ ist im Fall von linearen Exponentialfamilien durch

$$D(y;\hat{\mu}) = \sum_{i=1}^{n} d_i(y_i;\hat{\mu}) \quad ,$$

mit

$$d_i(y_i;\hat{\mu}) = -2\int_{y_i}^{\hat{\mu}} \frac{y_i - u}{a_i V(u)} du = 2\Big[y_i\{\theta(y_i) - \theta(\hat{\mu}_i)\} + b\{\theta(\hat{\mu}_i)\} - b\{\theta(\hat{y}_i)\}\Big]/a_i \quad ,$$

gegeben. Wie die RSS hängt $D(y;\hat{\mu})$ nur von den Daten, nicht aber von irgendwelchen Parametern ab.

Vorher wurde angenommen ϕ sei bekannt. Die Differenz der Devianzen muß aber mit $1/\phi$ skaliert werden, bevor sie auf eine χ^2–Verteilung mit $df = p_B - p_A$ bezogen werden kann. Im Falle der Poisson-, Binomial- und Exponentialverteilung ist ϕ bekannt und gleich 1, andernfalls muß ein Schätzer verwendet werden. In der Normalverteilungstheorie, speziell bei varianzanalytischen Modellen, wird ϕ durch $\tilde{\phi}$ aus der RSS des komplexesten Modells einer Reihe hierarchischer Modelle geschätzt. Das Verhältnis $(\mathrm{RSS}_A - \mathrm{RSS}_B)/\tilde{\phi}(p_B - p_A)$ kann dann mittels der F–Verteilung geprüft werden. Diese Vorgangsweise basierend auf der Differenz der Devianzen kann analog in einem allgemeineren Rahmen verwendet werden. Voraussetzung hiefür ist *i)* $\tilde{\phi}$ ist konsistent für ϕ und hat approximativ eine entsprechend skalierte χ^2–Verteilung, *ii)* $\tilde{\phi}$ und $\{D(y;\hat{\mu}^{(B)}) - D(y;\hat{\mu}^{(A)})\}$ sind approximativ unabhängig.

2.5 Goodness of fit

Die Devianzfunktion hat einige einfache Eigeschaften, die ihre Nützlichkeit zur Einschätzung der Güte der Anpassung anzeigen. Paßt ein Modell perfekt, $y = \hat{\mu}$, dann nimmt sie den Wert 0 an, sonst ist sie positiv. Da Maximieren der Likelihood für irgendein Modell dem Minimieren der Devianz entspricht, liefert die ML–Methode den besten Fit auch nach dem Devianzkriterium. Die Devianz kann selbst als Differenz $\{D(y;\hat{\mu}) - D(y;y)\}$ aufgefaßt werden, d.h. als Differenz der Devianzen des aktuell gefitteten Modells und dem saturierten Modell in dem $y = \hat{\mu}$. Trivialerweise sind diese beide Modelle in einer hierarchischen Ordnung und man ist versucht aufgrund der Ergebnisse des vorherigen Abschnitts zu schließen, daß die Devianz selbst auch approximativ $\phi\chi^2_{n-p}$–verteilt ist, wenn das gefittete Modell gültig ist. Standardtheorie, die zur $\chi^2_{p_B-p_A}$ Approximation für die Nullverteilung der LR–Statistik führt, basiert auf dem Grenzwert $n \to \infty$, mit fixierten p_A und p_B. Wenn B das saturierte Modell ist, dann ist $p_B = n$ und die Standardtheorie gilt nicht mehr. Daraus folgt, daß die Devianz nicht unter allgemeinen Bedingungen asymptotisch χ^2–verteilt ist, wenn die Anzahl der Beobachtungen wächst, d.h. die Devianz kann weit von einer χ^2–Verteilung entfernt sein, auch dann wenn n groß ist. Eine weitere Konsequenz besteht darin, daß die $\chi^2_{p_B-p_A}$ Approximation dann schlecht sein kann, wenn p_B im Verhältnis zu n groß ist. Allerdings ist die χ^2 Approximation der Verteilung der Devianz ohnehin

meistens gut, besonders wenn der Informationsgehalt für jede Beobachtung einzeln betrachtet groß ist. Dies ist vor allem bei Poissonmodellen mit großen μ_i, Binomialmodellen mit großen m_i und Gammamodellen mit kleinem ϕ der Fall. Man sollte sich aber davor hüten, exakte Wahrscheinlichkeitsaussagen zu treffen.

3 Quasi–Likelihood Modelle

Die Schätzung der interessierenden Parameter in verallgemeinerten Modellen beruht auf der ML Theorie. Um eine Likelihood Funktion konstruieren zu können ist es üblicherweise notwendig, einen probabilistischen Mechanismus anzugeben, der für einen Bereich von Parameterwerten, die Wahrscheinlichkeit für alle relevanten Stichproben spezifiziert, die möglicherweise hätten beobachtet werden können. Diese Spezifikation erfordert entweder Kenntnisse über den Mechanismus, durch den Daten generiert wurden oder substantielle Erfahrung mit ähnlichen Daten aus früheren Experimenten.

Oft gibt es keine Theorie über diesen Zufallsmechanismus, man kann aber eventuell den Wertebereich möglicher Responsewerte (diskret, kontinuierlich, positiv, ...) angeben, oder aufgrund früherer Erfahrung einige zusätzliche Charakteristika spezifizieren, etwa *i*) wie der Mittelwert oder Median von externen Stimuli oder Treatments beeinflußt wird, *ii*) wie die Variabilität der Response sich mit dem Erwartungswert der Response ändert, *iii*) ob die Beobachtungen statistisch unabhängig sind, *iv*) welche Schiefe die Responseverteilung unter fixen Treatment-Bedingungen hat.

Gibt es Vorinformationen, dann üblicherweise über die Art der Beziehung, wie die mittlere Reponse von Kovariaten beeinflußt wird, aber kaum über das Muster höherer Momente der Responsevariable. Die hier gegebene Darstellung soll Methoden vorstellen, wie man Inferenz betreiben kann, wenn zuwenig Information zur Konstruktion einer Likelihoodfunktion vorhanden ist.

Ausgangspunkt dieser Überlegungen sind die Scoregleichungen (7) , die unter der Voraussetzung, daß die Regressionsgleichung $E(Y_i) = \mu_i(\beta)$ korrekt ist, erwartungstreue Schätzgleichungen sind. Unter milden Bedingungen kann das Gleichungssystem gelöst werden und ergibt allgemein eine konsistente Schätzfunktion für β, auch wenn die Y_i nicht aus einer linearen Exponentialfamilie stammen. Setzt man Exponentialfamilien voraus, dann geht aufgrund dieser Annahme in (7) nur die Spezifikation der Varianzfunktion $V(\mu)$ ein, da in jeder dieser Familien gilt, daß

$$\frac{\partial l}{\partial \mu_i} = \frac{y_i - \mu_i}{\phi V(\mu_i)}.$$

Daher erscheint es interessant, das Verhalten der Schätzer, die sich aus (7) ergeben, nur unter Annahmen über die ersten beiden Momente,

$$E(Y_i) = \mu_i(\beta) \quad , \qquad \mathrm{Var}(Y_i) = \phi_i V(\mu_i) \tag{14}$$

zu untersuchen, anstatt die strengeren Annahmen einer Exponentialfamilie vorauszusetzen. Das wesentlichste hierbei ist, daß die Score- bzw. Informationsidentitäten

$$E\Big(\frac{\partial l}{\partial \mu_i}\Big) = E\Big(\frac{Y_i - \mu_i}{\phi V(\mu_i)}\Big) = 0$$

$$E\Big(-\frac{\partial^2 l}{\partial \mu_i^2}\Big) = E\Big(\frac{V(\mu_i) + (Y_i - \mu_i)V(\mu_i)}{\phi\{V(\mu_i)\}^2}\Big) = \frac{1}{\phi V(\mu_i)} = Var\Big(\frac{\partial l}{\partial \mu_i}\Big)$$

auch unter (14) gelten. Da diese Identitäten die Basis für die asymptotische Theorie der ML–Schätzung bilden, gelten deren Resultate auch hier. Im speziellen sind die $\hat{\beta}$ ebenso asymptotisch normalverteilt wie im Abschnitt 2.3 beschrieben. Man verwendet also Ergebnisse der Theorie über Inferenz in linearen Exponentialfamilien. Trifft man dabei nur Annahmen nur über die ersten beiden Momente wird dies Quasi–Likelihood (QL) Schätzung genannt (WEDDERBURN, 1974). Ein Modell der Form (14) heißt QL–Modell und soll sinnvolle Inferenz auch dann ermöglichen, wenn eine auf der Likelihood basierende Analyse unter gegebenen Annahmen nur sehr schwierig oder gar nicht erfolgen kann. Die Eigenschaft, die eine QL von direkter Anwendung in Schätzgleichungen unterscheidet, ist die Existenz (in vielen Fällen) einer Quasilikelihood, d.h. einer skalaren Funktion, deren Gradientenvektor die Schätzgleichungen gibt. Existiert eine solche Funktion, kann sie zur Konstruktion von Konfidenzbereichen für Parameter verwendet werden, so wie bei üblichen Likelihoods in voller parametrischer Inferenz, und ist daher besser als Methoden, die direkt auf Schätzgleichungen bzw. auf Schätzern beruhen.

Die eben gegebene Formulierung ist sehr allgemein, von primärer praktischer Bedeutung sind folgende Anwendungsfälle, auf die im weiteren (abgesehen vom ersten Punkt, der den Fall konstanter Varianz behandelt) detaillierter eingegangen werden soll.

1. *Konstante Varianz*: In diesem Fall ist QL–Schätzung mit dem Kleinst–Quadrate Verfahren (wobei unter Umständen noch die bekannten Konstanten $1/a_i$ als Gewichte dienen) ident.

2. *Konstanter Variationskoeffizient*: $V(\mu) = \mu^2$. Diese Annahme ist dann nützlich, wenn eine multiplikative Fehlerstruktur vermutet wird, $Y_i = \mu_i(\beta)\varepsilon_i$, aber die Verteilung der ε_i unbekannt ist. Der QL–Ansatz ist in diesem Fall äquivalent zum ML–Ansatz mit der Annahme, daß die ε_i einer Gammaverteilung folgen.

3. *Überdispersion*: Dies betrifft besonders die Poisson-, Binomial- und Exponentialverteilung. Bei diesen drei Verteilungen, die die Standardannahmen bei Häufigkeitsdaten, Anteilswerten und Wartezeiten sind, ist $\phi = 1$ bekannt. In der Praxis tritt aber öfters der Fall ein, daß die Streuung der Daten gegenüber den Standardannahmen zu groß ist., d.h. $\phi > 1$. Die Formulierung eines QL–Modells ist eine mögliche Lösung dieses Problems.

Verteilung (mit Überdispersion)	Varianzfunktion
Poisson	$V(\mu) = \phi\mu$
Binomial	$V(\mu) = \phi\mu(1-\mu)$
Exponential	$V(\mu) = \phi\mu^2$

Da die Schätzgleichungen für $\beta_1, \ldots, \beta_p$ nicht von ϕ abhängen, ergeben sich die gleichen $\hat{\beta}$, wie im Fall $\phi = 1$. Allerdings ist die $\text{Cov}(\hat{\beta})$ proportional zu ϕ, und daher werden alle Standardfehler mit $\phi^{1/2}$ oder einem Schätzer davon multipliziert. Das Problem der Unterdispersion tritt in der Praxis weniger häufig auf, kann aber in gleicher Weise behandelt werden.

4. *Abhängige Beobachtungen:* Eine Erweiterung von Quasi-Likelihoodmodellen, in denen man auch Korrelationen zwischen Beobachtungen zuläßt, führt zu verallgemeinerten Schätzgleichungen für die β. Die Kovarianzmatrix $V(\mu)$ ist dann nicht mehr diagonal sondern üblicherweise blockdiagonal. Es können hiebei verschiedene Arten von Korrelationsstrukturen festgelegt werden, die zu unterschiedlichen Modellklassen führen.

3.1 Unabhängige Beobachtungen

3.1.1 Konstruktion der QL Funktion

Das erste Kapitel beschäftigte sich unter anderem damit, die statistischen Eigenschaften der Lösung der Gleichung $U(\hat{\beta}; y) = 0$ zu besprechen. Hier interessiert nun die Frage, unter welchen Bedingungen eine Funktion $Q(\beta; y)$ mit Gradientenvektor $U(\beta; y)$ existiert. Existiert $Q(\beta; y)$ und erfüllt sie die Bedingungen einer logarithmierten Likelihood, dann nennt man $Q(\beta; y)$ eine Quasi-Likelihood Funktion. (Die Bedingungen für ihre Existenz diskutieren McCullagh und Nelder (1989), Kap. 9.)

Ausgegangen wird wieder von einem Responsevektor Y mit voneinander unabhängigen Komponenten. Dieser hat den Erwartungswert μ, und eine Kovarianzstruktur $\phi V(\mu)$. $V(\mu)$ ist eine Matrix bekannter Funktionen, ϕ kann unbekannt sein. Wieder ist das Ziel die Parameter β zu schätzen, die die Art des Zusammenhangs zwischen μ und den erklärenden Variablen x festlegen, d.h. $\mu = \mu(\beta)$. Wichtig ist, daß ϕ konstant ist und nicht von β abhängt.

Da die Komponenten von Y unabhängig sind, ist $V(\mu)$ eine Diagonalmatrix, deren Elemente nur von der i-ten Komponente von μ abhängen,

$$V(\mu) = diag\{V_1(\mu_1), \ldots, V_n(\mu_n)\}$$

In den meisten Anwendungen werden die Funktionen $V_1(\mu_1), \ldots, V_n(\mu_n)$ identisch sein, obwohl die Werte, die sie annehmen, unterschiedlich sind, da sie ja von den μ_i abhängen. Diese Annahme ist aber nicht notwendig. Es haben unter diesen Bedingungen die Größen

$$U(\beta; y) = \sum_i \frac{y_i - \mu_i}{V_i(\mu_i)} \frac{\partial \mu_i}{\partial \beta}$$

die gleichen Eigenschaften, wie wenn sie die ersten Ableitungen einer logarithmierten Likelihood wären, nämlich

$$E(U) = 0$$

$$\text{Var}(U_i) = \frac{1}{\phi V(\mu_i)}$$

Da ein Großteil der asymptotischen Theorie bezüglich Likelihoodfunktion auf diesen Eigenschaften beruht, ist es nicht überraschend, daß sich

$$Q(\beta; y) = \sum_i \int_{y_i}^{\mu_i} \frac{y_i - t}{\phi V_i(t)} dt$$

wie eine logarithmierte Likelihoodfunktion verhält, deren Gradientenvektor bezüglich β $U(\beta; y)$ ist.

Die folgenden beiden Beispiele sollen diese Idee illustrieren.

1.Normalverteilung: $V(t) = 1$ und $\phi = \sigma^2$

$$\begin{aligned} Q(\mu; y) &= \int_y^\mu \frac{y - t}{\phi V(t)} dt = \frac{1}{\phi} \int_y^\mu (y - t) dt \\ &= \frac{1}{\phi}(yt \mid_y^\mu - \frac{t^2}{2} \mid_y^\mu) = \frac{1}{\phi}(y\mu - y^2 - 1/2(\mu^2 - y^2)) \\ &= -1/2(\frac{y - \mu}{\sigma})^2 \end{aligned}$$

Bis auf Konstanten, die beim Differenzieren wegfallen, bleibt also ein Term über, der der logarithmierten Likelihood der Normalverteilung

$$l = \ln\{\frac{1}{\sqrt{2\pi\sigma}} \exp(-1/2(\frac{y - \mu}{\sigma})^2\}$$

entspricht.

2. Bernoulliverteilung $V(t) = \mu(1 - \mu)$ und $\phi = 1$

$$\begin{aligned} Q(\mu; y) &= \int_y^\mu \frac{y - t}{t(1 - t)} dt = y \int_y^\mu \frac{1}{t(1 - t)} dt - \int_y^\mu \frac{1}{1 - t} dt \\ &= y\Big[\int_y^\mu \frac{1}{t} dt + \int_y^\mu \frac{1}{1 - t} dt\Big] - \int_y^\mu \frac{1}{1 - t} dt \\ &= y \ln \frac{t}{1 - t} \mid_y^\mu + \ln(1 - t) \mid_y^\mu = \\ &= y \ln \frac{\mu}{1 - \mu} + \ln(1 - \mu) + c \end{aligned}$$

Wieder entspricht dies bis auf Konstanten einer logarithmierten Likelihood, in diesem Beispiel der der Bernoulliverteilung:

$$l = y \ln(\frac{\mu}{1-\mu}) + \ln(1-\mu).$$

Weitere Beispiele sind in der folgenden Tabelle dargestellt, wobei viele aber nicht alle Quasi-Likelihoods wirklichen logarithmierten Likelihoods bekannter Verteilungen entsprechen.

Verteilung	Varianzfunktion $V(\mu)$	Quasi-Likelihood $Q(\mu;y)$	Kanonischer Parameter θ
Normal	1	$-(y-\mu)^2/2$	μ
Poisson	μ	$y\ln\mu - \mu$	$\ln\mu$
Gamma	μ^2	$-y/\mu - \ln\mu$	$-1/\mu$
Inverse Gauss	μ^3	$-y/(2\mu^2) + 1/\mu$	$-1/(2\mu^2)$
—	μ^ζ	$\mu^{-\zeta}\left(\frac{\mu y}{1-\zeta} - \frac{\mu^2}{2-\zeta}\right)$	$\frac{1}{(1-\zeta)\mu^{\zeta-1}}$
Binomial	$\mu(1-\mu)$	$y\ln\left(\frac{\mu}{1-\mu}\right) + \ln(1-\mu)$	$\ln\left(\frac{\mu}{1-\mu}\right)$
—	$\mu^2(1-\mu)^2$	$(2y-1)\ln\left(\frac{\mu}{1-\mu}\right) - \frac{y}{\mu} - \frac{1-y}{1-\mu}$	—
Negative Binomial	$\mu + \mu^2/k$	$y\ln\left(\frac{\mu}{k+\mu}\right) + k\ln\left(\frac{k}{k+\mu}\right)$	$\ln\left(\frac{\mu}{k+\mu}\right)$

3.1.2 Inferenz in QL–Modellen

Die Quasi-Likelihood Schätzgleichung für β erhält man durch Differenzieren von $Q(\mu;y)$. Dies liefert das Gleichungssystem

$$U(\beta) = D'V^{-1}\frac{(Y-\mu)}{\phi}$$

$U(\hat{\beta}) = 0$ wird auch Quasi-Score Funktion genannt. Dabei ist D eine $n \times p$–Matrix mit Elementen $D_{ir} = \partial\mu_i/\partial\beta_r$, d.h. den Ableitung von $\mu(\beta)$ nach den Parametern β. Sowohl D als auch V hängen von β ab.

Die Kovaranzmatrix von $U(\beta)$ als negativer Erwartungswert von $\partial V(\beta)/\partial\beta$ ist

$$i_\beta = D'V^{-1}D\phi^{-1}.$$

Sie entspricht der Fisher–Information bei gewöhnlichen Likelihood Funktionen. Unter gewissen Voraussetzungen über die Eigenwerte von of i_β ist die asymptotische Kovarianzmatrix von $\hat{\beta}$ durch

$$\mathrm{Cov}(\hat{\beta}) \simeq i_\beta^{-1} = \phi(D'V^{-1}D)^{-1}$$

gegeben. Die statistische Eigenschaften der Lösungen $\hat{\beta}$ können aus den Eigenschaften der Quasi–Scorefunktion in der Nähe des wahren Parameterwertes deduziert werden. Der große Vorteil dieser indirekten Technik ist es, daß $U(\beta;y)$ eine lineare Funktion des Responsevektors ist, während $\hat{\beta}$ in y nichtlinear ist. Die exakten Momente sind leicht berechenbar und der zentrale Grenzwertsatz garantiert approximative Normalität der Quasi–Scorefunktion unter recht allgemeinen Bedingungen. Die Newton Approximation erster Ordnung für die Lösung $\hat{\beta}$ beginnt beim wahren aber unbekannten Parameter β und ist

$$\beta^{(1)} = \beta + (D'V^{-1}D)^{-1}U(\beta;y)$$

Vorausgesetzt, daß weitere Schritte vernachlässigbar sind, folgt daraus

$$E(\hat{\beta}) = \beta$$

und

$$\begin{aligned}\mathrm{Cov}(\hat{\beta}) &= E\Big\{(D'V^{-1}D)^{-1}U(\beta;y)((D'V^{-1}D)^{-1}U(\beta;y))'\Big\} = \\ &= (D'V^{-1}D)^{-1}\mathrm{Cov}(U(\beta;y))(D'V^{-1}D)^{-1} = \\ &= (D'V^{-1}D)^{-1} = i_\beta^{-1} \qquad (15)\end{aligned}$$

Die Quasi-score Funktion ist also eine optimale, erwartungstreue Schätzgleichung für β.

Besteht β nur aus einem einzigen Parameter und wenn $U(\beta;y)$ monoton fallend in β ist, dann kann die exakte Verteilung von $\hat{\beta}$ aus der Äquivalenz der Ereignisse $\hat{\beta} \leq \beta^*$ und $U(\beta^*;y) \leq 0$ direkt gewonnen werden.

Ist z.B. $U(\beta^*;y)$ normal verteilt, dann hat letzteres Ereignis die Wahrscheinlichkeit $\Phi(z)$ mit

$$Z = D'^*V^{*-1}(\mu^* - \mu)/(D^*V^{*-1}VV^{*-1}D^*)^{1/2}.$$

(D^*, V^* und μ^* sind D, V und μ berechnet an der Stelle β^*.) Unter all diesen Aspekten verhält sich eine Quasi-Likelihood wie eine gewöhnliche Likelihood.

Zur Schätzung von ϕ verhält sich $Q(\mu; y)$ allerdings nicht wie eine logarithmierte Likelihood. Üblicherweise verwendet man einen Momenten–Schätzer für $\hat{\phi}$. Dieser beruht auf dem Residuenvektor $y - \mu$, nämlich

$$\hat{\phi} = \frac{1}{n-p} \sum_i \frac{(Y_i - \mu_i)^2}{V_i(\mu_i)} = \frac{X^2}{n-p}$$

mit X^2 als der verallgemeinerte Pearson Statistik.

Die numerische Berechnung der $\hat{\beta}$ erfolgt wie bei generalisierten linearen Modellen mittels der IWLS Methode

$$\hat{\beta}^{(t+1)} = \{D'V^{-1}D\}^{-1}D'V^{-1}Z$$

mit $Z = D\hat{\beta}^{(t)} - S$. Der aktuelle Schätzer $\hat{\beta}^{(t)}$ geht hierbei sowohl in D, V als auch in $S = (y - \hat{\mu})$ ein. Anzumerken ist noch, daß die Schätzung unabhängig von ϕ erfolgt.

Ebenso in Analogie zu den generalisierten linearen Modellen wird die Quasi-Devianz

$$D(y; \mu) = -2\phi Q(\beta; y) = \sum_i \int_{y_i}^{\mu_i} \frac{y_i - t}{\phi V_i(t)} dt$$

zum Testen von Hypothesen herangezogen.

BEISPIEL 1. Konstanter Variationskoeffizient

Seien die Zufallsvariablen $Y_i, \ldots, Y_n$ unabhängig mit existierenden Erwartungswerten $E(Y_i) = \mu_i$ und $\mathrm{Var}(Y_i) = \phi\mu_i^2$. Dann ist nicht die Varianz, sondern der Variationskoeffizient $\phi^{1/2}$ konstant für alle Beobachtungen i. Sei ferner ein Regressionsmodell

$$\ln \mu_i = \alpha + \beta(x_i - \bar{x}), \qquad i = 1, \ldots, n.$$

mit bekannten Konstanten $x_1, \ldots, x_n$, wobei α und β geschätzt werden sollen. Dann ist *i*) die Beziehung zwischen $\mu = E(Y)$ und $\beta = (\alpha, \beta)$ nicht linear in β, *ii*) die Kovarianzmatrix der Y gleich $\mathrm{Cov}(Y) = \phi V(\mu)$ mit den bekannten Funktionen $V(\cdot)$ und *iii*) das Modell völlig durch die Beziehung zwischen den ersten beiden Momenten von Y spezifiziert. Die Quasi-Likelihoodfunktion ist $-\sum_i (y_i/\mu_i + \ln \mu_i)$ und die Schätzgleichungen werden zu

$$\sum_{i=1}^{n} \frac{y_i - \hat{\mu}_i}{\hat{\mu}_i} = 0$$

und

$$\sum_{i=1}^{n} \frac{(x_i - \bar{x})(y_i - \hat{\mu}_i)}{\hat{\mu}_i} = 0.$$

BEISPIEL 2. Überdispersion

Im Rahmen einer klinischen Studie sollte die Wirksamkeit eines bestimmten Medikamentes M zur Behandlung einer Zahnfleischerkrankung mit einer Placebotherapie verglichen werden. Das Ausmaß der Erkrankung wurde folgendermaßen operationalisiert: Nach Stimulation am Zahnfleisch eines Zahnes wird geprüft, ob eine Blutung eingetreten ist. Dies weist daraufhin, daß das Zahnfleisch an dieser Stelle krank ist. (Man könnte also vermuten, daß Y_i binomialverteilt ist, $Y_i \sim B(1, \pi_i)$, mit π_i als der Wahrscheinlichkeit für an dieser Stelle erkranktes Zahnfleisch.) Durch Messung an mehreren Zähnen m_i, erhält man als Responsevariable y_i die Anzahl der Stellen, an denen Blutungen aufgetreten sind., d.h. der Grad der Erkrankung könnte unter der Nullhypothese keiner Wirksamkeit von M gegenüber Placebo, durch $Y_i \sim B(m_i, \pi_i)$ repräsentiert sein. Nach Zufallsaufteilung erhielten 18 Probanden das Medikament M und 17 Placebo, wobei jeweils 9 Messungen in regelmäßigen Abständen $t_0, \ldots, t_8$ durchgeführt wurden. t_0 indiziert hierbei die Messung vor Beginn der Therapie. Die Rohdaten sind im Anhang dargestellt.

Unter der Annahme, daß die Messungen voneinander unabhängig sind, läßt sich die folgende (logistische) Regressionsstruktur erstellen, wenn die qualitative Variable M und die Zeit T $(t_0, \ldots, t_8)$ als quantitative Variable im Modell berücksichtigt werden:

Modell	linearer Prädiktor (η)	Interpretation
1	π_0	Y hängt weder von M noch T ab
M	$\pi_0 + \pi_M$	Y hängt von M ab, unabhängig von T
T	$\pi_0 + \pi_T$	Y hängt von T ab, unabhängig von M
M+T	$\pi_0 + \pi_M + \pi_T$	Y hängt von M und T ab, ohne Wechselwirkung zwischen M und T
M*T	$\pi_0 + \pi_M + \pi_T + \pi_{MT}$	Y hängt von M und T ab, mit Wechselwirkung zwischen M und T

Mittels GLIM wurde die folgende Modellsequenz errechnet:

Modell	Devianz	df	Differenz der Devianz zu Modell 1
1	1727.3	314	—
M	1722.9	313	4.4
T	852.5	313	874.8
M+T	819.1	312	908.2
M*T	810.6	311	916.7

Betrachtet man diese Werte, so fällt auf, daß das Medikament alleine nur relativ wenig zur Erklärung beiträgt, während die Zeit T alleine einen sehr starken Beitrag liefert. Nimmt man zu T aber das Medikament additiv in das Modell auf, so ist die Differenz zu Modell T mit $\chi^2 = 33.4$ bei $df = 1$ beträchtlich. Fügt man gegenüber diesem Modell (M + T) noch die Wechselwirkung hinzu, ergibt sich eine weitere Verbesserung um $\chi^2 = 8.5$ bei $df = 1$. Es ist also auch die Berücksichtigung dieses Terms im Modell notwendig. Die Parameterschätzer für Modell (M*T) sind:

Parameterschätzer	Standardfehler	Parameter
1.833	0.231	π_0
-0.665	0.062	π_T
-0.236	0.319	π_M
0.177	0.081	π_{MT}

Ein Vergleich der Parameterwerte aus dem Quasi-Likelihoodmodell mit jenen aus dem logistischen Modell zeigt nur geringfügige Unterschiede. Die Interpretation bleibt gleich: Es erfolgt generell eine Verbesserung über die Zeit, die aber in der mit dem Medikament behandelten Gruppe stärker ist.

3.2 Abhängige Beobachtungen

Bisher wurden Methoden behandelt, die eine Erweiterung der GLM insoferne erlauben, als die Schätzgleichungen nicht unbedingt aufgrund bestimmter Verteilungsannahmen abgeleitet werden, sondern nur über die Beziehung der ersten beiden Momente definiert sind. Eine zusätzliche Verallgemeinerung besteht nun darin, die Kovarianzmatrix der Y nicht als Diagonalmatrix d.h. $V(\mu) = diag\{V_1(\mu_1), \ldots, V_n(\mu_n)\}$ aufzufassen, sondern zuzulassen, daß $\mathrm{Cov}(Y) = \phi V(\mu)$ ist, mit $V(\mu)$ als einer symmetrischen, positiv definiten $n \times n$ Matrix bekannter Funktionen $V_{ij}(\mu)$. Diese Verallgemeinerung ist speziell in Modellen für Longitudinaldaten sinnvoll. In solchen Fällen hat $V(\mu)$ eine Blockdiagonalstruktur unter der Voraussetzung, daß die Beobachtungseinheiten, $i = 1, \ldots, n$, voneinander unabhängig sind, aber (positive) Korrelationen zwischen wiederholten Messungen innerhalb einer Beobachtungseinheit zu vermuten sind. Seien also Y_i die Responses bei einer Beobachtungen i zu Zeitpunkten $t = 1, \ldots, T_i$ $Y_i = (y_{i1}, \ldots, y_{iT_i})'$ und $X_i = (x_{i1}, \ldots, x_{iT_i})'$ eine $T_i \times p$ Matrix der Kovariatenwerte, wobei die Randverteilung der Y_{it} aus einer linearen Exponentialfamile

$$f(y_{it}; \theta_{it}) = \exp\{(\theta_{it} y_{it} - b(\theta_{it}))/\phi + c(y_{it}, \phi)\}, \tag{16}$$

stammt. Der Unterschied zu (3) ist nur die zusätzliche Indizierung für die Wiederholungen t. Erwartungwert und Varianz sind nun $E(y_{it}) = b'\theta_{it}$ und $\mathrm{Var}(y_{it}) = \phi b''(\theta_{it})$. Zur Vereinfachung der Notation wird im folgenden $T_i = T$ für alle i geschrieben, wodurch aber die allgemeinere Formulierung nicht eingeschränkt wird.

Um die hier vorgestellten Methoden zu illustrieren und den Zusammenhang zum vorherigen Abschnitt herzustellen sei einmal die Annahme getroffen, daß wiederholte Beobachtungen voneinander unabhängig sind: Dann haben die Scoregleichungen die Form

$$U(\beta_U) = \sum_{i=1}^{n} X_i' \Delta_i S_i = 0 \tag{17}$$

mit der $T \times T$ Matrix $\Delta_i = diag(d\theta_{it}/d\eta_{it})$ und der $T \times 1$ Vektor $S_i = Y_i - \mu_i$. Die Lösung von (17) liefert unter der Unabhängigkeitannahme den Schätzer $\hat{\beta}_U$ für β.

Dann kann man zeigen (LIANG und ZEGER (1986)), daß $\hat{\beta}_U$ konsistent und asymptotisch normalverteilt ist, mit einer asymptotischen Kovarianzmatrix

$$V_U = \lim_{n\to\infty} n\{H_1^{(U)}(\beta)\}^{-1}\{H_2^{(U)}(\beta)\}\{H_1^{(U)}(\beta)\}^{-1}, \tag{18}$$

wobei

$$H_1^{(U)}(\beta) = \sum_{i=1}^{n} X_i'\Delta_i A_i \Delta_i X_i,$$

$$H_2^{(U)}(\beta) = \sum_{i=1}^{n} X_i'\Delta_i \mathrm{Cov}(Y_i)\Delta_i X_i.$$

A_i ist eine $n \times n$ Matrix, $A_i = diag\{b''(\theta_i)\}$. Die Varianz der $\hat{\beta}_U$ läßt sich konsistent schätzen, wenn in (18) $\hat{\beta}_U$ und $\mathrm{Cov}(Y_i) = S_i S_i'$ eingesetzt wird. Der Schätzer $\hat{\beta}_U$ ist leicht zu berechnen, z.B. mittels GLIM. Sowohl $\hat{\beta}_U$ als auch die zugehörigen geschätzten Standardfehler sind konsistent, wenn das spezifizierte Modell gilt. Der Hauptnachteil von $\hat{\beta}_U$ ist die geringe Effizienz, wenn die Korrelation unter den y_{it} hoch ist. Allerdings läßt sich (17) so verallgemeinern, daß solche Korrelationen berücksichtigt werden können.

3.2.1 Verallgemeinerte Schätzgleichungen

Sei $R(\alpha)$ eine symmetrische $T \times T$ Matrix, die als Korrelationsmatrix aufgefaßt werden kann, und α sei eine Menge von s Korrelationsparametern, die $R(\alpha)$ völlig charakterisieren. Dann ist

$$V_i = \phi A^{1/2} R(\alpha) A^{1/2} \tag{19}$$

gleich der Kovarianzmatrix der Y_i, wenn $R(\alpha)$ wirklich die Korrelationsmatrix der Y_i ist. Die verallgemeinerten Schätzgleichungen sind

$$U(\beta_G) = \sum_{i=1}^{n} D_i' V_i^{-1} S_i = 0 \tag{20}$$

wobei $D_i = \{\partial b'(\theta)/\partial\beta\} = A_i\Delta_i X_i$. Ist $R(\alpha)$ die Einheitsmatrix, reduziert sich (20) auf den Fall der Unabhängigkeit. Das Gleichungssystem (20) ist eine Erweiterung des QL-Ansatzes in dem Sinn, als die Matrix V_i nicht nur von β, sondern auch von den α abhängt. Außer bei spezieller Wahl von R und α ist der Dispersionsparameter ϕ in (20) enthalten. Sei $\hat{\beta}_G$ die Lösung von (20), dann gelten die gleichen Ergebnisse bezüglich Konsistenz und Normalität wie im vorigen Abschnitt für $\hat{\beta}_U$. Die Kovarianzmatrix der β_G hat die Form

$$V_G = \lim_{n\to\infty} n\{H_1^{(G)}(\beta)\}^{-1}\{H_2^{(G)}(\beta)\}\{H_1^{(G)}(\beta)\}^{-1}, \tag{21}$$

mit

$$H_1^{(G)} = \sum_{i=1}^{n} D_i' V_i^{-1} D_i,$$

$$H_2^{(G)} = \sum_{i=1}^{n} D' V^{-1} \mathrm{Cov}(Y_i) V_i^{-1} D_i.$$

3.2.2 Inferenz

Die Schätzung der Standardfehler von $\hat{\beta}_G$ erfolgt ebenso durch Einsetzen von $S_i S_i'$ für $\mathrm{Cov}(Y_i)$ in (21) und durch Ersetzen von β, ϕ, α durch ihre Schätzer in V_G. Wie im Falle der Unabhängigkeit hängt die Konsistenz von $\hat{\beta}_G$ und $\hat{V}_G$ nur davon ab, ob das Modell korrekt spezifiziert ist, nicht aber von der korrekten Wahl von R. Wie im QL-Ansatz hängt die asymptotische Varianz des $\hat{\beta}_G$ nicht von ϕ ab. Die Resultate erhält man im hier behandelten Fall, in dem die Likelihood nicht zur Gänze spezifiziert ist, aus der Wahl von Schätzgleichungen für β in (20), wo der individuelle Beitrag einer Beobachtungseinheit aus dem Produkt von Termen besteht, d.h. daß V_i von α aber nicht von den Daten abhängig ist und S_i unabhängig von α ist, mit $E(S_i) = 0$.

Zur Schätzung von β_G wird wieder die IWLS-Methode verwendet.

$$\hat{\beta}_G^{(t+1)} = \{D_i' V_i^{-1} D_i\}^{-1} D_i' V_i^{-1} Z_i$$

mit $Z_i = D_i \hat{\beta}_G^{(t)} - S_i$.

Nach einer gegebenen Iteration können die α und ϕ aus den Pearson Residuen

$$\hat{r}^{it} = \frac{\{y_{it} - b'(\hat{\theta}_{it})\}}{\{b''(\hat{\theta}_{it})\}^{1/2}}$$

berechnet werden, $\hat{\theta}_{it}$ hängt von den $\hat{\beta}_G^{(t+1)}$ ab. In Analogie zu Kapitel 2 erhält man einen Schätzer für ϕ aus

$$\tilde{\phi} = \sum_{i=1}^{n} \sum_{i=1}^{T} \frac{\hat{r}_{it}^2}{(N-p)}$$

mit $N = \sum T_i$.

Durch die Wahl einer bestimmten Korrelationsstruktur lassen sich unterschiedliche Modelle spezifizieren, wobei die α unterschiedlich geschätzt werden. Hiezu einige Beispiele:

1. Sei $R(\alpha) = R_0$ irgendeine Korrelationsmatrix. Ist R_0 die Einheitsmatrix, dann erhält man die Schätzgleichungen im Fall der Unabhängigkeit. Allerdings können für jede R_0 die Parameter $\hat{\beta}_G$ und $\hat{V}_G$ konsistent geschätzt werden. Je enger R_0 die wahre Korrelationsstruktur widerspiegelt,

um so höher wird die Effizienz sein. Zur Schätzung β und $\mathrm{Var}(\hat{\beta}_G)$ ist keine Kenntnis von ϕ nötig, wenn irgendein R_0 spezifiziert ist.

2. Sei $\alpha = (\alpha_1, \ldots, \alpha_{T-1})'$, mit $\alpha_t = \mathrm{Corr}(Y_{it}, Y_{i,t+1})$ für $t = 1, \ldots, T-1$. Dann ist ein natürlicher Schätzer für α_t, gegeben β und ϕ

$$\hat{\alpha}_t = \phi^{-1} \sum_{i=1}^{n} \frac{\hat{r}_{it}\hat{r}_{i,t+1}}{(n-p)}. \tag{22}$$

Wenn nun $R(\alpha)$ eine Bandmatrix mit Nebendiagonalelementen $\{R\}_{t,t+1} = \alpha_t$ ist, dann erhält man ein Modell, in dem jeweils 2 benachbarte Beobachtungen abhängig sind. Wieder ist es nicht notwendig ϕ zu schätzen, um $\hat{\beta}_G$ und $\hat{V}_G$ zu berechnen, da das ϕ in (22) sich bei der Berechnung von V_i wegkürzt. Als Spezialfall kann man ein gemeinsames $\alpha = \alpha_t$, $t = 1, \ldots, T-1$ festlegen. Die Schätzfunktion hierfür ist

$$\hat{\alpha} = \sum_{t=1}^{T-1} \frac{\hat{\alpha}_t}{(T-1)}$$

Ebenso lassen sich Abhängigkeiten höherer Ordnung berechnen.

3. Spezifiziert man nur einen Parameter α für alle Beobachtungen, d.h. $\mathrm{Corr}(y_{it}, y_{it'}) = \alpha$, für $t \neq t'$, dann entspricht dies einer 'austauschbaren' Korrelationsstruktur, wie man sie auch bei random–effect Modellen erhält, wo 'random–effect' Parameter über Beobachtungseinheiten hinweg variieren können (siehe z.B. LAIRD und WARE, 1982). Bei gegebenem ϕ wird α durch

$$\hat{\alpha} = \phi^{-1} \frac{\sum_{i=1}^{n} \sum_{t>t'} \hat{r}_{it}\hat{r}_{it'}}{\sum_{i=1}^{n} T_i(T_i-1)/2 - p}$$

geschätzt werden. Wieder ist es nicht notwendig ϕ zur Bestimmung von $\hat{\beta}_G$ und $\mathrm{Var}(\hat{\beta}_G)$ zu schätzen.

4. Bei Festlegung einer Korrelationsstruktur auf $\mathrm{Corr}(y_{it}, y_{it'}) = \alpha^{|t-t'|}$ entspricht dies im Falle der Normalverteilung einem autoregressiven Prozeß erster Ordnung, AR–1.Da unter diesem Modell $E(\hat{r}_{it}\hat{r}_{it'}) \simeq \alpha^{|t-t'|}$, kann α mittels des Regressionsansatzes $\ln(\hat{r}_{it}\hat{r}_{it'}) = \alpha(\ln \mid t - t' \mid)$ geschätzt werden. Hier ist es allerdings notwendig $\hat{\phi}$ zu bestimmen, damit β_G und $\mathrm{Var}(\hat{\beta}_G)$ geschätzt werden können.

5. Will man nicht a priori eine bestimmte Korrelationsstruktur voraussetzen, kann man $R(\alpha)$ unspezifiziert lassen, muß aber dann $s = T(T-1)/2$ Korrelationsparameter schätzen. $\hat{R}$ erhält man mittels

$$\phi^{-1} n^{-1} \sum_{i=1}^{n} A_i^{-1/2} S_i S_i' A_i^{-1/2}$$

In diesem Fall reduziert sich die asymptotische Kovarianz V_G zu

$$\lim_{n\to\infty} \left\{ \frac{1}{n} \sum_{i=1}^{n} D_i' \mathrm{Cov}(Y_i)^{-1} D_i \right\},$$

da R die tatsächliche Korrelationsmatrix ist. Aufgrund der möglicherweise hohen Zahl zu schätzender Parameter wird dieses Modell nur bei moderaten T sinnvoll sein.

Wendet man diese Methode auf das in Beispiel 2. (Kap. 3.1.2) dargestellte Problem an, erhält man folgende Parameterschätzer und Standardfehler:

Parameterschätzer	Standardfehler	Parameter
1.915	0.050	π_0
-0.639	0.013	π_T
-0.223	0.071	π_M
0.136	0.017	π_{MT}

Ein Vergleich mit den Werten aus dem logistischen Modell zeigt, daß die $\hat{\beta}$ nahezu ident sind, allerdings ist die Größe der Standardfehler wesentlich reduziert. Die geschätzen Korrelationen liegen zwischen -0.08 und 0.048.

Zitierte Literatur

FIRTH, D. (1991): Genealized linear models. In: HINKLEY, D.V., REID, N., SNELL, E.J.: *Statistical theory and modelling.* London: Chapman and Hall.

JØRGENSEN, B. (1987): Exponential dispersion models (with discussion). *J. R. Statist. Soc.* B **49**, 127 – 162.

LAIRD, N.M. UND WARE, J.H. (1982): Random–effects models for longitudinal data . *Biometrics* **38**, 963 – 974 .

LIANG, K.Y. UND ZEGER, S.L. (1986): Longitudinal data analysis using generalized linear models. *Biometrika* **73**, 13 – 22 .

MCCULLAGH, P. UND NELDER, J.A. (1989): *Generalized linear models. Second Edition.* London: Chapman and Hall.

PAYNE, C.D. (1986): *The GLIM Manual, Release 3.77* Oxford: NAG.

WEDDERBURN, R.W.M. (1974): Quasi–likelihood functions, generalized linear models, and the Gauss–Newton method. *Biometrika* **61**, 439 – 447.

WEDDERBURN, R.W.M. (1976): On the existence and uniqueness of the maximum likelihood estimates for certain generalized linear models. *Biometrika* **63**, 27 – 32.

WILLIAMS, D.A. (1982): Extra–binomial variation in logistic linear models. *Appl. Statist.* **31**, 144 – 148.

Anhang

Die folgende Tabelle enthält die Rohdaten zu Beispiel 2 in Kapitel 3.1.2 bzw. 3.2.2 mit m_i als Zahl der untersuchten Zähne y_0 bis y_8 als die zu den einzelnen Untersuchungszeitpunkten festgestellten Zahllen kranker Zähne.

Prob.Nr.	Behandlung	m_i	y_0	y_1	y_2	y_3	y_4	y_5	y_6	y_7	y_8
1	M	6	6	5	4	2	1	0	0	0	0
2	M	12	12	12	10	8	4	1	3	1	2
3	M	15	14	11	4	0	2	0	0	0	0
4	M	8	8	8	5	3	1	0	0	0	1
5	M	18	18	17	12	4	4	3	6	4	4
6	M	10	9	10	9	7	3	3	2	1	1
7	M	7	7	6	3	0	2	0	1	0	3
8	M	19	17	19	18	14	4	4	4	7	4
9	M	4	0	3	3	1	1	1	0	0	0
10	M	15	15	13	14	12	5	3	0	4	5
11	M	3	3	2	0	0	1	1	1	0	1
12	M	7	5	6	6	3	0	1	1	0	0
13	M	4	4	4	1	0	0	0	0	0	0
14	M	6	6	4	3	1	0	0	0	0	0
15	M	6	5	6	5	1	0	1	0	2	0
16	M	2	2	2	1	0	0	0	0	0	0
17	M	5	2	5	4	2	2	0	1	1	0
18	M	4	4	3	4	0	0	1	0	0	0
19	P	2	2	2	0	2	1	0	2	2	2
20	P	6	5	4	6	2	0	0	2	0	0
21	P	6	5	3	2	0	1	0	0	0	0
22	P	11	9	9	5	2	0	0	0	0	0
23	P	4	4	2	0	1	0	0	0	1	0
24	P	9	8	5	2	1	0	0	0	0	0
25	P	9	9	9	7	3	4	2	2	1	1
26	P	10	10	10	10	6	8	5	2	5	5
27	P	3	2	3	3	0	1	0	0	1	1
28	P	11	11	11	9	6	5	2	0	0	1
29	P	12	11	11	11	7	4	3	2	1	5
30	P	5	2	5	3	2	1	2	0	2	0
31	P	2	2	2	1	0	1	0	0	1	0
32	P	12	10	12	12	9	9	8	8	4	2
33	P	11	9	11	6	5	4	6	2	1	1
34	P	4	4	4	4	4	3	1	0	1	1
35	P	14	6	14	11	8	3	6	1	2	4