Auxiliary Mixture Sampling für Poisson Random Effects Modelle

Michaela Dvorzak
Johannes Kepler Universität Linz
Institut für Angewandte Statistik

Random Effects Modelle

Poisson Random Effects Modelle eignen sich für die Analyse von poissonverteilten Längsschnittdaten, die angesichts der wiederholten Messungen an den Versuchseinheiten mit dem Problem korrelierter Daten verbunden sind. Die vorhandene Abhängigkeit zwischen den zu verschiedenen Zeitpunkten beobachteten Zähldaten impliziert dabei eine unbeobachtete Heterogenität zwischen den einzelnen Individuen, die bei der Modellierung berücksichtigt werden muss.

Unter Verwendung eines Random Effects Modells wird durch die zusätzliche Aufnahme von Random Effects für jede Person eine individuelle Abweichung von einem durchschnittlichen, durch die Parameter der fixen Effekte spezifizierten, Niveau geschätzt. Mit diesem Random Intercept werden die nicht beobachtbaren, individuellen Unterschiede zwischen den Versuchseinheiten erfasst und die mit der Verteilungsannahme einer Poissonverteilung verbundene Overdispersion der Daten berücksichtigt.

Modellierung

Sei $\boldsymbol{Y}$ eine Zählvariable, die für verschiedene Individuen $i = 1, \ldots, n$ zu wiederholten Zeitpunkten bzw. Zeitintervallen $j = 1, \ldots, J$ beobachtet wird. Bei gegebenem Parametervektor der fixen Effekte $\boldsymbol{\beta}$ und dem Vektor der zufälligen Effekte $\boldsymbol{\alpha}$ sind die Zähldaten y_{ij} unabhängig poissonverteilt mit dem Parameter λ_{ij} (vgl. [1]):

$$y_{ij}|\boldsymbol{\beta}, \boldsymbol{\alpha} \sim \text{Poisson } (\lambda_{ij}),$$
$$\lambda_{ij} = E(y_{ij}|\boldsymbol{\beta}, \boldsymbol{\alpha}) = \exp(\boldsymbol{X}_{ij}\boldsymbol{\beta} + \boldsymbol{Z}_{ij}\boldsymbol{\alpha}).$$

$\boldsymbol{\beta}$ stellt den unbekannten, aus den Daten zu schätzenden Parametervektor der fixen Effekte dar und $\boldsymbol{X}_{ij} = \left(1, X_{ij2}, \ldots, X_{ijp}\right)$ den entsprechenden Kovariablenvektor für $\boldsymbol{\beta}$. $\boldsymbol{\alpha}$ ist der Vektor der individuellen Effekte und $\boldsymbol{Z}$ die zugehörige Kovariablenmatrix bestehend aus den Werten 0 und 1, die die Zuordnung der Effekte zu den entsprechenden Individuen vornimmt. Allen J Beobachtungen der i-ten Versuchseinheit wird folglich der gleiche Zufallseffekt α_i (mit $i = 1, \ldots, n$) zugeordnet. Mit diesem Effekt wird ein individuelles Verhalten geschätzt, das die an diesem Individuum wiederholten Beobachtungen beeinflusst.

Unter der Annahme der Unabhängigkeit zwischen den Effekten der einzelnen Individuen wird für den Vektor $\boldsymbol{\alpha}$ folgende Verteilungsannahme getroffen:

$$\boldsymbol{\alpha} \sim N\left(\mathbf{0}, \sigma_\alpha^2 \cdot \mathrm{I}_n\right), \quad \sigma_\alpha^2 \sim \Gamma^{-1}\left(c_0/2, C_0/2\right).$$

Die Varianz σ_α^2 der Random Effects deutet auf das Ausmaß der unbeobachteten Heterogenität in den Daten und somit auf die Notwendigkeit einer zusätzlichen Aufnahme der individuellen Effekte bei der Modellierung hin.

Bayesianische Schätzung

Die bayesianische Schätzung der Modellparameter $\boldsymbol{\beta}$, $\boldsymbol{\alpha}$ und σ_α^2 erfolgt mithilfe von MCMC-Methoden unter Anwendung des in [2] vorgeschlagenen Auxiliary Mixture Samplers. Dieser ermöglicht direktes Gibbs Sampling sämtlicher Parameter, das lediglich Ziehungen aus Standardverteilungen vorsieht und einen Metropolis Hastings-Algorithmus vermeidet. Der Auxiliary Mixture Sampler setzt allerdings eine künstliche Aufbereitung der Daten voraus, die den Datensatz um zwei Folgen latenter Variablen erweitert.

Durch die zusätzliche Einführung unbeobachteter Wartezeiten eines angenommenen Poisson-Prozesses wird im ersten Schritt dieser Datenerweiterung die Nichtlinearität des Modells beseitigt.

Die Verteilung der nicht-normalverteilten Fehler im entstandenen Modell wird wie in [3] durch eine Mischung von Normalverteilungen angenähert. Durch die Erweiterung der Daten um den Indikator der Mischungskomponente erhält man ein normalverteiltes, gemischtes Regressionsmodell, aus dem die Parameter direkt mittels Gibbs Sampling geschätzt werden können.

Datenerweiterung

Im ersten Schritt der Datenerweiterung wird für jede Anzahl an Ereignissen y_{ij} ein unbeobachteter Poisson-Prozess mit der Intensität λ_{ij} im Intervall $[0, 1]$ angenommen und die Daten um die Wartezeiten τ_{ijk} mit $k = 1, \ldots, (y_{ij} + 1)$ erweitert. Den Eigenschaften eines Poisson-Prozesses entsprechend sind die Wartezeiten unabhängig exponentialverteilt mit dem Parameter λ_{ij}:

$$\tau_{ijk}|\boldsymbol{\beta}, \boldsymbol{\alpha} \sim \text{Ex}\left(\lambda_{ij}\right) = \frac{\xi_{ijk}}{\lambda_{ij}}, \quad \xi_{ijk} \sim \text{Ex}(1).$$

Die Verteilung des Fehlerterms $\varepsilon_{ijk} \sim -\log\left(\text{Ex}\left(1\right)\right)$ im entstandenen Modell

$$-\log \tau_{ijk}|\boldsymbol{\beta}, \boldsymbol{\alpha} = \boldsymbol{X}_{ij}\boldsymbol{\beta} + \boldsymbol{Z}_{ij}\boldsymbol{\alpha} + \varepsilon_{ijk}$$

wird wie in [3] durch eine Mischung von 10 Normalverteilungen approximiert

$$p_{\boldsymbol{\varepsilon}}(\boldsymbol{\varepsilon}) = \exp\left(-\boldsymbol{\varepsilon} - e^{-\boldsymbol{\varepsilon}}\right) \approx \sum_{r=1}^{10} w_r \cdot f_{\mathrm{N}}\left(\boldsymbol{\varepsilon}; m_r, s_r^2\right).$$

Durch unabhängiges Ziehen aus $p(r|\varepsilon_{ijk})$ wird für jedes ε_{ijk} der Indikator r_{ijk} der Mischungskomponente ermittelt, um den die Daten im Zuge dieses zweiten Datenaufbereitungsschrittes erweitert werden.

Bei gegebenen Wartezeiten $\boldsymbol{\tau}$ und Mischungsindikatoren $\boldsymbol{\mathcal{R}} = \{r_{ijk} : i = 1, \ldots n; j = 1, \ldots J; k = 1, \ldots (y_{ij} + 1)\}$ ergibt sich ein normalverteiltes, lineares Regressionsmodell für den Vektor $\boldsymbol{y}_{ij}$:

$$\boldsymbol{y}_{ij}|\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\tau}, \boldsymbol{\mathcal{R}} = \mathbf{1}\boldsymbol{X}_{ij}\boldsymbol{\beta} + \mathbf{1}\boldsymbol{Z}_{ij}\boldsymbol{\alpha} + \boldsymbol{\varepsilon}_{ij}$$
$$\text{mit } \boldsymbol{y}_{ij} = -\log \boldsymbol{\tau}_{ij} - \boldsymbol{m}_{\boldsymbol{r}_{ij}}$$
$$\text{und } \boldsymbol{\varepsilon}_{ij} \sim N\left(\mathbf{0}, \boldsymbol{\Sigma}_{ij} = \text{diag}(\boldsymbol{s}^2_{\boldsymbol{r}_{ij}})\right).$$

Auf der Grundlage des hergeleiteten Regressionsmodells erfolgt die Schätzung der Modellparameter mithilfe des folgenden 3-stufigen Gibbs Samplers.

Auxiliary Mixture Sampler

Nach der Wahl der Startwerte für $\boldsymbol{\beta}^{(0)}, \boldsymbol{\alpha}^{(0)}$ und $\sigma_\alpha^{2\,(0)}$ sind die Schritte (a) bis (c) des Auxiliary Mixture Samplers für $m = 1, \ldots, M$ zu wiederholen:

(a) Bestimme die Wartezeiten $\boldsymbol{\tau}$ sowie die Indikatoren $\boldsymbol{\mathcal{R}}$ bei gegebenen Parametern $\boldsymbol{\beta}, \boldsymbol{\alpha}, \sigma_\alpha^2$ und $\boldsymbol{y}$:

(a1) Ziehe $\boldsymbol{\tau}^{(m)}$ aus $p(\boldsymbol{\tau}|\boldsymbol{\beta}^{(m-1)}, \boldsymbol{\alpha}^{(m-1)}, \sigma_\alpha^{2\,(m-1)}, \boldsymbol{y})$:

- ziehe $u_{ij(1)}, \ldots, u_{ij(y_{ij})} \sim U[0, 1]$,
- berechne $\tau_{ijk} = u_{ij(k)} - u_{ij(k-1)}, \quad k = 1, \ldots y_{ij}$ (mit $u_{ij(0)} := 0$),
- $\tau_{ij(y_{ij}+1)} = 1 - \sum_{k=1}^{y_{ij}} \tau_{ijk} + \psi_{ij}, \quad \psi_{ij} \sim \text{Ex}(\lambda_{ij})$.

(a2) Ziehe die Indikatoren $\boldsymbol{\mathcal{R}}^{(m)}$ aus der diskreten Verteilung $p(\boldsymbol{\mathcal{R}}|\boldsymbol{\tau}^{(m)}, \boldsymbol{\beta}^{(m-1)}, \boldsymbol{\alpha}^{(m-1)}, \sigma_\alpha^{2\,(m-1)})$.

(b) Bestimme den Vektor der fixen Effekte $\boldsymbol{\beta}$ und den Vektor der individuellen Effekte $\boldsymbol{\alpha}$:

(b1) Ziehe $\boldsymbol{\beta}^{(m)}$ aus $p(\boldsymbol{\beta}|\boldsymbol{\tau}^{(m)}, \boldsymbol{\mathcal{R}}^{(m)}, \sigma_\alpha^{2\,(m-1)})$:

$$\boldsymbol{\beta}|\boldsymbol{\tau}, \boldsymbol{\mathcal{R}}, \sigma_\alpha^2 \sim N_p\left(\boldsymbol{b}_N, \boldsymbol{B}_N\right)$$
$$\text{mit } \boldsymbol{b}_N = \boldsymbol{B}_N \cdot \left(\boldsymbol{B}_0^{-1}\boldsymbol{b}_0 + \sum_{i=1}^n \boldsymbol{X}_i'\mathbf{1}'\boldsymbol{V}_i^{-1}\boldsymbol{y}_i\right)$$
$$\text{und } \boldsymbol{B}_N = \left(\boldsymbol{B}_0^{-1} + \sum_{i=1}^n \boldsymbol{X}_i'\mathbf{1}'\boldsymbol{V}_i^{-1}\mathbf{1}\boldsymbol{X}_i\right)^{-1}$$

(b2) Ziehe $\boldsymbol{\alpha}^{(m)}$ aus $p(\boldsymbol{\alpha}|\boldsymbol{\tau}^{(m)}, \boldsymbol{\mathcal{R}}^{(m)}, \boldsymbol{\beta}^{(m)}, \sigma_\alpha^{2\,(m-1)})$:

$$\boldsymbol{\alpha}|\boldsymbol{\beta}, \boldsymbol{\tau}, \boldsymbol{\mathcal{R}}, \sigma_\alpha^2 \sim N_n\left(\boldsymbol{a}_N, \boldsymbol{A}_N\right)$$
$$\text{mit } \boldsymbol{a}_N = \boldsymbol{A}_N \cdot \left(\sum_{i=1}^n \sum_{j=1}^J \boldsymbol{Z}_{ij}'\mathbf{1}'\boldsymbol{\Sigma}_{ij}^{-1}(\boldsymbol{y}_{ij} - \mathbf{1}\boldsymbol{X}_{ij}\boldsymbol{\beta})\right)$$
$$\text{und } \boldsymbol{A}_N^{-1} = (\sigma_\alpha^2 \cdot \boldsymbol{I}_{\mathrm{n}})^{-1} + \sum_{i=1}^n \sum_{j=1}^J \boldsymbol{Z}_{ij}'\mathbf{1}'\boldsymbol{\Sigma}_{ij}^{-1}\mathbf{1}\boldsymbol{Z}_{ij}$$

(c) Bestimme die Varianz σ_α^2 der individuellen Effekte aus $p(\sigma_\alpha^2|\boldsymbol{\tau}^{(m)}, \boldsymbol{\mathcal{R}}^{(m)}, \boldsymbol{\beta}^{(m)}, \boldsymbol{\alpha}^{(m)})$:

$$\sigma_\alpha^2|\boldsymbol{\alpha} \sim \Gamma^{-1}\left((c_0 + n)/2, (C_0 + \sum_{i=1}^n \alpha_i^2)/2\right)$$

Anwendung: Epilepsie-Daten

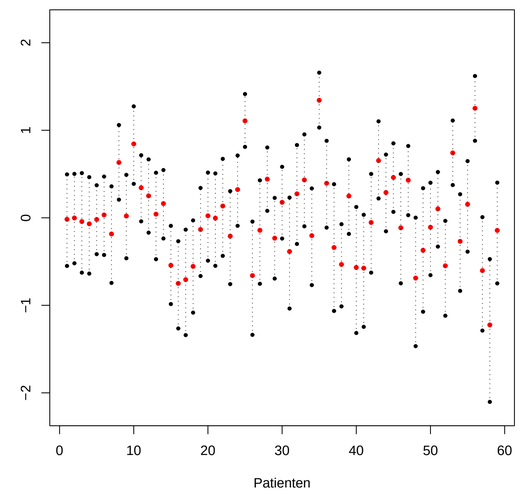
Für die praktische Umsetzung des Auxiliary Mixture Samplers wurde der Algorithmus im Rahmen dieser Arbeit in der Programmumgebung R implementiert.

Die Anwendung des Auxiliary Mixture Samplers erfolgt auf einen Paneldatensatz einer klinischen Studie, die an Epilepsie-Patienten durchgeführt wurde (vgl. [4]). Dieser Datensatz enthält für insgesamt $n = 59$ Patienten die Anzahl an epileptischen Anfällen y_{ij} innerhalb von $j = 1, \ldots, 4$ je zweiwöchigen Intervallen. Um den Effekt eines neuen Antiepileptikums namens *Progabide* auf die Frequenz der Anfälle zu untersuchen, wurde den Studienteilnehmern je nach Behandlungsgruppe dieses neue Medikament oder ein Placebo als Zusatz zur Standardbehandlung verabreicht (Indikatorvariable *Treatment*).

Durch die wiederholten Beobachtungen an den Patienten sind die Daten zu den verschiedenen Zeitpunkten korreliert ($\rho_{y_{i1}, y_{ij}} = 0.871, 0.738, 0.893$ für $j = 2, 3, 4$), während die Zahl der Krampfanfälle zwischen den Epileptikern stark streut (*within-subject dependence* vs. *between-subject variation*). Mithilfe eines Random Effects Modells wird aus diesem Grund für jeden Studienteilnehmer i ein individueller Patienteneffekt α_i (Random Intercept) geschätzt.

Auf der Grundlage des Auxiliary Mixture Samplers wird die Abhängigkeit der Anfallsfrequenz der Patienten vom Parametervektor $\boldsymbol{\beta}$ und den Zufallseffekten $\boldsymbol{\alpha}$ mittels Gibbs Sampling bestimmt. Der Algorithmus wird dazu $M = 12000$ mal durchgeführt (M_0 (burn-in) $= 2000$) und die Schätzer für die Modellparameter über die entsprechenden a posteriori-Erwartungswerte der Ziehungen ermittelt.

Die in der folgenden Abbildung dargestellten, geschätzten Random Effects α_i lassen deutliche Unterschiede zwischen den Patienten hinsichtlich ihres Anfallverhaltens erkennen.



EPILEPSIE-DATEN: A posteriori-Erwartungswerte sowie 2.5- bzw. 97.5-Perzentile der individuellen Patienteneffekte α_i $(i = 1, \ldots, 59)$

So weisen Patienten mit einer geringen Anfallsfrequenz Effekte nahe bei Null auf, während für jene mit einem auffällig starken Anfallverhalten vergleichsweise hohe, positive Patienteneffekte ermittelt werden.

Die geschätzte Varianz $\hat{\sigma}_\alpha^2 = 0.330$ (a posteriori-Intervall: [0.195;0.490]) der Zufallseffekte deutet darauf hin, dass die konkrete Datensituation die Aufnahme der individuellen Effekte bei der Modellierung erfordert (bestes Modell im bayesianischen Modellvergleich mittels DIC).

Literatur

[1] Chib, S., Winkelmann, R. (2001). Markov Chain Monte Carlo Analysis of Correlated Count Data. *Journal of Business & Economic Statistics*, Vol. 19, No. 4: 428-435.

[2] Frühwirth-Schnatter, S. und Wagner, H. (2006). Auxiliary Mixture Sampling for Parameter-driven Models of Time Series of Small Counts with Applications to State Space Modelling. *Biometrika*, Vol. 93: 827-841.

[3] Frühwirth-Schnatter, S. und Wagner, H. (2006). Data Augmentation and Gibbs Sampling for Regression Models of Time Series of Small Counts. *Student*, Vol. 5: 221-234.

[4] Leppik, I., Dreifuss, F., Porter, R., et al. (1987). A controlled study of progabide in partial seizures: Methodology and results. *Neurology*, Vol.37: 963-968.