Aufgabe 23 (Bayesianisches GLM; Metropolis-(Hastings)-Algorithmus)

Im Datenarchiv des Statistikinstituts findet man unter der Adresse http://www.stat.uni-muenchen.de /service/datenarchiv/sprache/sprache.asc einen Datensatz zu den Auswirkungen des Lächelns auf Sprache mit der folgenden Beschreibung:

„Der Datensatz basiert auf einem Wahrnehmungsexperiment, bei dem entschieden werden sollte, ob beim Sprechen eines Wortes gelächelt wurde oder nicht. Dazu wurden zwölf verschiedene Wörter jeweils acht Mal von einem Sprecher gesprochen und verschiedene physikalische Messgrößen hierzu ermittelt. [...] Zusätzlich war bekannt, ob der Sprecher nach eigener Aussage beim Sprechen gelächelt hatte. “

Von Interesse ist dabei aus statistischer Sicht, ob durch die physikalischen Messgrößen vorhergesagt werden kann, ob der Sprecher beim Sprechen gelächelt hat oder nicht. Da die Response-Variable *Lächeln* binär ist, bietet sich ein Logit-Modell an. An Hand dieses Datensatzes wird im Folgenden die Bayes-Inferenz mit der Likelihood-Inferenz für das Logit-Modell verglichen.

Sei $y_i \in \{0, 1\}$ für $i = 1, \dots, n$ der Indikator dafür, ob der Sprecher bei Wort i gelächelt hat und $\boldsymbol{x_i}$ der Vektor der physikalischen Messgrößen zur Aufnahme von Wort i. Die Annahmen des Logit-Modells sind:

- Verteilungsannahme: $y_i|\boldsymbol{x_i} \overset{unabh.}{\sim} \text{Bin}(1, \pi_i)$
- Linearer Prädiktor: $\eta_i = \boldsymbol{x_i}^\top \boldsymbol{\beta}$
- Linkfunktion: $\text{E}(y_i|\boldsymbol{x_i}) = \pi_i = h(\eta_i) = \frac{\exp(\eta_i)}{1+\exp(\eta_i)} \left(\Leftrightarrow g(\pi_i) = \eta_i = \log\left(\frac{\pi_i}{1-\pi_i}\right)\right)$

(a) Berechnen Sie zunächst den ML-Schätzer für $\boldsymbol{\beta}$ mit Hilfe der Funktion `glm` in R. Welche Kovariablen zeigen auf einem Signifikanzniveau von $\alpha = 0.05$ einen signifikanten Zusammenhang mit der Response-Variablen *Lächeln*?

Für die bayesianische Inferenz ist die Spezifizierung einer Priori-Verteilung für $\boldsymbol{\beta}$ nötig. Wir nehmen im Folgenden eine multivariate Normalverteilung an

$$\boldsymbol{\beta} \sim \text{N}_p(\boldsymbol{b}_0, \boldsymbol{B}_0).$$

(b) Berechnen Sie die Dichte der gemeinsamen Posteriori-Verteilung von $\boldsymbol{\beta}|\boldsymbol{y}, \boldsymbol{X}$ bis auf Proportionalität.

Da dieser Ausdruck nicht dem Kern einer bekannten Verteilung entspricht, können die Momente und Quantile der Posteriori-Verteilung nicht direkt analytisch bestimmt werden, auch ein Gibbs-Sampling-Algorithmus kann nicht angewendet werden. Eine Möglichkeit zur Erzeugung von Zufallszahlen aus der Verteilung von $\boldsymbol{\beta}|\boldsymbol{y}, \boldsymbol{X}$ ist der Metropolis(-Hastings)-Algorithmus.

(c) Beschreiben Sie in eigenen Worten oder Pseudocode einen geeigneten Algorithmus und verwenden Sie dabei als Vorschlagsdichte für $\boldsymbol{\beta}_{neu}|\boldsymbol{\beta}_{alt}$ eine multivariate Normalverteilung, die um den Wert der vorherigen Iteration $\boldsymbol{\beta}_{alt}$ zentriert ist und als Kovarianzmatrix die geschätzte Kovarianzmatrix des ML-Schätzers hat:

$$\boldsymbol{\beta}_{neu}|\boldsymbol{\beta}_{alt} \sim \mathrm{N}_p\left(\boldsymbol{\beta}_{alt}, \widehat{\mathrm{Cov}}\left(\widehat{\boldsymbol{\beta}}_{ML}\right)\right).$$

Hinweis: $\widehat{\mathrm{Cov}}\left(\widehat{\boldsymbol{\beta}}_{ML}\right)$ wird bei `summary.glm` mit berechnet.

(d) Implementieren Sie den Algorithmus in R und simulieren Sie mit $\boldsymbol{b}_0 = \boldsymbol{0}, \boldsymbol{B}_0 = 1000 \cdot \boldsymbol{I}_p$ und Startwert $\boldsymbol{\beta}_0 = \boldsymbol{0}$ auf diese Weise $D = 5000$ Zufallszahlen, plotten Sie diese und entscheiden Sie, wie viele Zahlen als Burnin verworfen werden müssen.

(e) Bestimmen Sie aus den verbleibenden Zufallszahlen Punktschätzer sowie 95% - Kredibilitätsintervalle für die Komponenten von $\boldsymbol{\beta}$. Vergleichen Sie die Ergebnisse mit den Ergebnissen der ML-Schätzung.

(f) Führen Sie die bayesianische Analyse jetzt mit Hilfe der Funktion `MCMClogit` aus dem R-Paket `MCMCpack` durch. Sind die Ergebnisse vergleichbar?

(g) Das R-Paket `coda` stellt eine große Auswahl an Diagnostik-Funktionen für ein `mcmc`-Objekt (z.B. Output von `MCMClogit`) bereit. Machen Sie sich mit den Funktionen vertraut. Welchen Nutzen könnte ein Ausdünnen der Samplingpfade haben?

(h) Geben Sie einen Punktschätzer sowie ein 95% - Kredibilitätsintervall für $\exp(\beta_1)$ und für $\beta_1 + \beta_2$ an. Wie lässt sich die Schätzung für $\exp(\beta_1)$ interpretieren?

(i) Beschreiben Sie eine Möglichkeit, im Likelihood-Kontext ein Konfidenzintervall für $\exp(\beta_1)$ und für $\beta_1 + \beta_2$ zu bestimmen.

*__Aufgabe 13__ (Bayesianisches GLM)

Gegeben seien Wartezeiten $y_1, \ldots, y_n$ und Vektoren von p Kovariablen $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$, mit welchen die Wartezeiten erklärt werden sollen. Für diese Fragestellung bietet sich ein generalisiertes lineares Modell mit folgenden Annahmen an:

- Verteilungsannahme: $y_i|\boldsymbol{x}_i \overset{unabh.}{\sim} \text{Exp}(\lambda_i)$
- Linearer Prädiktor: $\eta_i = \boldsymbol{x}_i^\top \boldsymbol{\beta}$
- Linkfunktion: $\text{E}(y_i|\boldsymbol{x}_i) = \mu_i = h(\eta_i) = \exp(\eta_i)\,(\Leftrightarrow g(\mu_i) = \eta_i = \log(\mu_i))\,.$

Für die bayesianische Inferenz ist ausserdem die Spezifizierung einer Priori-Verteilung für $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\top$ nötig. Wir nehmen im Folgenden eine multivariate Normalverteilung an

$$\boldsymbol{\beta} \sim \text{N}_p(\boldsymbol{b}_0, \boldsymbol{B}_0).$$

Als Posteriori-Dichte für $\boldsymbol{\beta}|\boldsymbol{y}, \boldsymbol{X}$ ergibt sich

$$p(\boldsymbol{\beta}|\boldsymbol{y}, \boldsymbol{X}) \propto \exp\left\{ -\left(\sum_{i=1}^{n} \boldsymbol{x}_i^\top \boldsymbol{\beta} + y_i \exp(-\boldsymbol{x}_i^\top \boldsymbol{\beta}) \right) - \frac{1}{2} (\boldsymbol{\beta} - \boldsymbol{b}_0)^\top \boldsymbol{B}_0^{-1} (\boldsymbol{\beta} - \boldsymbol{b}_0) \right\}.$$

(a) Begründen Sie, wieso in diesem Fall der Metropolis-Algorithmus geeignet ist, um $D = 1000$ Zufallszahlen aus der Posteriori-Verteilung von $\boldsymbol{\beta}|\boldsymbol{y}, \boldsymbol{X}$ zu ziehen. Beschreiben Sie den Algorithmus in eigenen Worten oder Pseudocode. Verwenden Sie dabei als Vorschlagsdichte für $\boldsymbol{\beta}_{neu}|\boldsymbol{\beta}_{alt}$ eine multivariate Normalverteilung, die um den Wert der vorherigen Iteration $\boldsymbol{\beta}_{alt}$ zentriert ist und als Kovarianzmatrix die geschätzte Kovarianzmatrix des ML-Schätzers hat:

$$\boldsymbol{\beta}_{neu}|\boldsymbol{\beta}_{alt} \sim \text{N}_p\left(\boldsymbol{\beta}_{alt}, \widehat{\text{Cov}}\left(\widehat{\boldsymbol{\beta}}_{ML}\right)\right).$$

(b) Geben Sie eine Formel an, um aus diesen Zufallszahlen einen Punktschätzer für $\mathbb{E}(\boldsymbol{\beta})$ zu berechnen.

(c) Beschreiben Sie, wie aus diesen Zufallszahlen ein $(1-\alpha)$ – Kredibilitätsintervall für β_1 berechnet werden kann.