§5 Gemischte Verallgemeinerte Lineare Modelle

- Wir betrachten zunächst einige allgemeine Aussagen für Gemischte Verallgemeinerte Lineare Modelle.
- Sei y der beobachtbare Zufallsvektor und u der Vektor der nicht-beobachtbaren zufälligen Effekte.
- Es wird typischerweise angenommen, dass der Zufallsvektor y aus bedingt unabhängigen Elementen besteht, die jeweils aus einer Verteilungen stammen, deren Dichte zur Exponentialfamilie gehört, z.B. Normalverteilung, Binomialverteilung oder Poisson-Verteilung.

Modellstruktur:

$$y_i|u \sim f_{Y_i|u}(y_i|u) \text{ unabhängig }, i = 1, \ldots, n$$

$$f_{Y_i|u}(y_i|u) = \exp\left\{ \left[y_i\gamma_i - b(\gamma_i)\right] / \tau^2 - c(y_i, \tau) \right\}$$

Der bedingte Erwartungswert von y_i steht in Beziehung zu $b(\gamma_i)$ durch $\mu_i = \partial b(\gamma_i)/\partial\gamma_i$.

- Wir benutzen dann eine Transformation dieses Erwartungswerts, so dass wir ein lineares Modell in festen und zufälligen Effekten erhalten:

$$\mathsf{E}[y_i|u] = \mu_i$$

$$g(\mu_i) = x_i^T \beta + z_i^T u.$$

- Die Funktion $g(\cdot)$ ist bekannt und heißt **link** Funktion.
- Für die Normalverteilung ist die identische Abbildung die $link$ Funktion, für die Binomial-Verteilung wird häufig die logit-Funktion als $link$-Funktion verwendet (logistische Regression), für Poisson- und Negative Binomial-Verteilung ist der Logarithmus die $link$-Funktion.
- x_i^T ist der i-te Zeilenvektor der Modellmatrix für die festen Effekte und β der Parametervektor der festen Effekte. z_i^T ist der i-te Zeilenvektor der Modellmatrix für die zufälligen Effekte und u der Vektor der zufälligen Effekte.
- Um das Modell vollständig zu spezifizieren, benötigen wir eine Verteilung der zufälligen Effekte:

$$u \sim f_U(u).$$

- Wir haben bisher die bedingte Verteilung von y spezifiziert. Wie sieht nun die marginale Verteilung von y aus?
- Für den Erwartungswert gilt

$$\mathsf{E}[y_i] = \mathsf{E}[\mathsf{E}[y_i|u]] = \mathsf{E}[\mu_i] = \mathsf{E}[g^{-1}(x_i^T\beta + z_i^T u)]$$

  Dieser Erwartungswert kann im Allgemeinen nicht vereinfacht werden.
- Zur Illustration betrachten wir den Logarithmus als *link* Funktion, d. h. $g(\mu) = \log(\mu)$ und $g^{-1}(x) = \exp(x)$. Dann gilt

$$\mathsf{E}[y_i] = \mathsf{E}[\exp(x_i^T\beta + z_i^T u)] = \exp(x_i^T\beta)\mathsf{E}[\exp(z_i^T u))] = \exp(x_i^T\beta)M_u(z_i).$$

  Dabei ist $M_u(z_i)$ die momentenerzeugende Funktion von u an der Stelle z_i.
- Nehmen wir weiterhin an, dass $u_i \sim \mathcal{N}(0, \sigma_u^2)$ und jede Zeile von Z hat einen einzigen Eintrag 1 und der Rest sind Nullen. Dann gilt $M_u(z_i) = \exp(\sigma_u^2)$ und

$$\mathsf{E}[y_i] = \exp(x_i^T\beta)\exp(\sigma_u^2/2)$$

  bzw.

$$\log \mathsf{E}[y_i] = x_i^T\beta + \sigma_u^2/2.$$

- Für Varianzen, Kovarianzen und Korrelationen, siehe McCulloch und Searle (2001).

- Die Likelihood-Funktion ist dann gegeben durch

$$L = \int \prod_i f_{y_i|u}(y_i|u) f_U(u) d_u,$$

  wobei die zufälligen Effekte 'rausintegriert' werden.
- Die ML-Schätzer müssen in der Regel nummerisch bestimmt werden.
- Alternativ können auch 'conditional' ML-Schätzer berechnet werden.
- Die Nutzung von marginalen und bedingten Modell diskutieren wir noch später.

- Eine Alternative zur ML-Schätzung bietet der Ansatz der **generalized estimating equations** (GEEs). Dabei wird der Erwartungswert im marginalen verallgemeinerten linearen Modell als Funktion der festen Effekte dargestellt.
- Zum Beispiel für die logistische Regression:

$$\text{logit}(\mathsf{E}[y]) = X\beta.$$

- Wenn wir die Arbeitshypothese der Unabhängigkeit aller Elemente in y haben, dann lautet die ML-Schätzgleichung für β

$$X^T y = X^T \mathsf{E}[y].$$

- Dies ist eine unverzerrte Schätzgleichung, denn $\mathsf{E}(X^T y - X^T \mathsf{E}(y)) = 0$.
- Unter gewissen Regularitätsbedingungen ergeben die Lösungen der Schätzgleichung konsistente Schätzer.

- Für Longitudinaldaten mit m Individuen lautet die Schätzgleichung bei binären Daten

$$\sum_{i=1}^{m} X_i \mathbf{y_i} = \sum_{i=1}^{m} X_i \mathsf{E}[\mathbf{y_i}],$$

  mit X_i der Modellmatrix der festen Effekte für das i-te Individuum und $\mathbf{y_i}$ der Beobachtungsvektor des i-ten Individuums.

- Die asymptotische Varianz der Lösung $\hat{\beta}$ ist gegeben durch

$$\mathsf{Var}_\infty(\hat{\beta}) = \left(\sum_i X_i^T X_i\right)^{-1} \left(\sum_i X_i^T \mathsf{Cov}(y_i) X_i\right) \left(\sum_i X_i^T X_i\right)^{-1}$$

- Diese Kovarianzmatrix kann konsistent geschätzt werden durch

$$\widehat{\mathsf{Var}}_\infty(\hat{\beta}) = \left(\sum_i X_i^T X_i\right)^{-1} \left(\sum_i X_i^T (y_i - \widehat{\mathsf{E}}[\mathbf{y_i}])(y_i - \widehat{\mathsf{E}}[\mathbf{y_i}])^T X_i\right) \left(\sum_i X_i^T X_i\right)^{-1}$$

- Die Annahme der Unabhängigkeit kann zu ineffizienten Schätzern führen und andere 'working' Kovarianzmatrizen können eingebracht werden.
- Die GEEs lauten dann

$$\sum_i X_i W_i y_i = \sum_i X_i W_i \mathsf{E}[\mathbf{y_i}]$$

  mit $W_i^{-1} = \mathsf{Cov}(\mathbf{y_i})$ der 'working' Kovarianzmatrix für $\mathbf{y_i}$.
- Die Varianzformeln müssen dann entsprechend angepasst werden, siehe Diggle et al. (1994).

Für die 'working' Kovarianzmatrix stehen in der Regel folgende Typen zur Verfügung:

1. **Independence**: Die 'working' Kovarianzmatrix ist die Einheitsmatrix. Die wiederholten Messungen werden als unabhängig angenommen.
2. **Exchangeable**: Die 'working' Kovarianzmatrix wird auch als $compound\ symmetry$ bezeichnet. Hier ist die Korrelation zwischen zwei Messungen an einem Individuum immer gleich, d. h. Kor$(y_{ij}, y_{ik}) = \rho$.
3. **AR(1)**: Autoregressive Korrelationsmatrix mit Kor$(y_{ij}, y_{ik}) = \rho^{|k-j|}, j \neq k$. Zeitlich näher beieinanderliegende Messungen sind stärker korreliert als zeitlich weiter auseinanderliegende Messungen.
4. **Unstructured**: Korrelationsmatrix mit $k(k-1)/2$ Parameter, wobei k die Anzahl der wiederholten Messungen ist und Kor$(y_{ij}, y_{ik}) = \rho_{jk}$. Es ist keine Struktur vorgegeben. Die Messungen können beliebig korreliert sein.

- Um den Unterschied zwischen der Modellierung im marginalen und im bedingten Modell darzustellen, nutzen wir ein Beispiel aus McCulloch und Searle (2001).
- Sei $y_{ij} = 1$, falls das j-te Kind einer Frau i frühgeboren wurde und $y_{ij} = 0$ sonst, und nehmen wir an, dass es eine erklärende Variable x_{ij} = 'Anzahl der alkoholischen Drinks pro Tag' gibt.
- Im marginalen Modell wird der marginale Erwartungswert von y_{ij} direkt, z.B. durch logistische Regression, an die Daten angepasst:

$$\text{logit}(\text{E}[y_{ij}]) = \text{logit}(\text{P}(y_{ij} = 1)) = \alpha + \beta x_{ij} \ .$$

- Wir modellieren hierbei den $Logit$ für die Wahrscheinlichkeit einer Frühgeburt für eine Population von Frauen. Wenn wir die Korrelation bei den Frauen berücksichtigen müssen, können wir einen GEE Ansatz nutzen.

- Der bedingte Ansatz dagegen berücksichtigt einen Zufallseffekt für die Frauen und spezifiziert ein bedingtes Modell derart, dass gilt

$$\text{logit}(\text{E}[y_{ij}|u]) = \text{logit}(\text{P}(y_{ij} = 1)) = \alpha + \beta x_{ij} + u_i$$

  mit u_i dem zufälligen Fraueneffekt.
- Dies entspricht der Modellierung einer bedingten Wahrscheinlichkeit einer Frühgeburt für jede Frau separat.
- Wenn die Frage in dem Beispiel ist, inwieweit die Inzidenz einer Frühgeburt verringert werden kann, wenn der durchschnittliche Alkoholkonsum einer Frau gesenkt wird, so ist das marginale Modell das adäquate Modell.
- Wenn man jedoch an der Frage interessiert ist, wie der Alkoholkonsum die individuelle Physiologie der Frauen beeinflusst, so ist das bedingte Modell das geeignete Modell.
-
- Hinweis auf R Programm: BtheBgee.R