Statistische Modellierung III

-Regression mit Zähldaten-

Dr. Martin Scharpenberg

MSc Medical Biometry/Biostatistics

WiSe 2019/2020

Setup

- Betrachten nun Zielvariablen Y, die die Anzahl von Ereignissen innerhalb einer festen Zeitperiode (z.B. die Anzahl der Migräneanfälle in drei Monaten) wiedergeben
- Y kann im Prinzip alle ganzzahligen Werte $0, 1, 2, \ldots$ annehmen
- Oft ist es praktikabel von einer unbeschränkten Zahl von Ereignissen auszugehen, auch wenn sie in der Praxis immer beschränkt ist, denn das Festlegen der Schranke kann schwierig sein
- Bei einer großen Zahl von Ereignissen ist Y oft gut durch die Normalverteilung (z.B. als Approximation der Binomialverteilung) zu beschreiben
- Bei kleineren Ereigniswahrscheinlichkeiten ist die Poisson-Verteilung oft das bessere Verteilungsmodell

Log-Lineares Poissonmodell

Verteilungsmodell

- Wir betrachten – wie bisher – den Erwartungswert $\lambda_i = E(Y_i|\mathbf{x}_i)$
- Wie bisher sei $\eta_i = \mathbf{x}_i\beta = x_{i1}\beta_1 + \ldots + x_{ik}\beta_k$; $x_{i1} = 1$ der lineare Prädiktor
- Nehmen nun $Y_i \sim Pois(\lambda_i)$ an (also auch $\lambda_i = E(Y_i|\mathbf{x}_i)$)
- Dies führt zum Log-Likelihood-Kern

$$l(\lambda_i) = Y_i \log(\lambda_i) - \lambda_i$$

- Unter der Annahmen einer Poissonverteilung ist die Varianzfunktion die Identität, denn $Var(Y_i) = \lambda_i$

Verteilungsmodell

- Die Dichte der Poissonverteilung wird oft um einen Dispersionsparameter und um Gewichte erweitert
- In diesem Fall ist der Log-Likelihood-Kern

$$l(\lambda_i, \phi) = \phi^{-1}\{Y_i \log \lambda_i - \lambda_i\}\,\omega_i$$

- Es folgt

$$E(Y_i) = \lambda_i \quad \text{und} \quad Var(Y_i) = \phi\,\lambda_i/\omega_i$$

Schätzung der Regressionskoeffizienten

- Bestimmen den MLE für β duch Maximierung von

$$l(\beta) = \sum_{i=1}^{n} l_i(\beta, \phi) = \sum_{i=1}^{n} \phi^{-1} \{ Y_i \mathbf{x}_i \beta - e^{\mathbf{x}_i \beta} \} \omega_i$$

- Wir bestimmen also mit dem Fisher-Scoring-Algorithmus die Lösung des Gleichungssystems

$$\phi \, \mathbf{s}(\beta) = \sum_{i=1}^{n} \mathbf{x}_i^T \{ Y_i - e^{\mathbf{x}_i \beta} \} \omega_i = \mathbf{0}$$

- Die Fisher-Information

$$\mathbf{F}(\beta, \phi) = \phi^{-1} \sum_{i=1}^{n} \mathbf{x}_i^T \mathbf{x}_i e^{\mathbf{x}_i \beta} \omega_i$$

ist invertierbar, falls $\text{Rang}(\mathbf{X}) = k$ und $\omega_i > 0$ für alle $i = 1, \ldots, n$

Überdispersion

- Meist wird $\phi = 1$ angenommen
- Bei gruppierten Daten kann die Varianz innerhalb einer Gruppe G_l durch die empirische Varianz $(n_l - 1)^{-1} \sum_{i \in G_l} (Y_i - \bar{Y}_l)^2$ abgeschätzt werden
- Oft beobachtet man, dass die empirische Varianz größer ist als die theoretische $v(\lambda_i)$ bei $\phi = 1$
- In diesem Fall ist ein Modell mit Dispersion $\phi \neq 1$ sinnvoll, wobei ϕ aus den Daten geschätzt wird
- Man unterscheidet zwei Fälle:
  - Überdispersion: $\phi > 1$
  - Unterdispersion: $\phi < 1$.

Überdispersion

- Überdispersion häufig, Unterdispersion eher selten
- Mögliche Gründe für Überdispersion:
  - Unbeobachtete Heterogenität: λ_i variiert innerhalb der Gruppe durch unbeobachtete Kovariablen
  - Positive Korrelation: Beobachtungseinheiten gehören (unbeobachteten) Clustern an (z.B. Familien), innerhalb denen sie sich ähnlicher sind als zwischen den Clustern. Das bewirkt eine positive Korrelation zwischen Beobachtungen Y_i des selben Clusters

Schätzung des Dispersionsparameters

- Im Fall von Einzelbeobachtungen:

$$\hat{\phi} = \frac{1}{n-k} \sum_{i=1}^{n} \frac{(Y_i - \hat{\lambda}_i)^2}{\hat{\lambda}_i}.$$

- Bei gruppierten Daten ($\omega_i = n_l$):

$$\hat{\phi} = \frac{1}{m-k} \sum_{l=1}^{m} \frac{(\bar{Y}_l - \hat{\lambda}_l)^2}{\hat{\lambda}_l} n_l.$$

Datenbeispiel

- Anzahl Zitate von Patenten und welche Faktoren diese Beeinflussen
- Daten von Insgesamt $n = 4866$ Patenten
- Zielvariable: Zahl der Zitate eines Patents
- Kovariablen (unter Anderem):
  - Jahr der Patenterteilung
  - Anzahl der Länder für die der Patentschutz gilt
  - etc.

Datenbeispiel

Variable	Beschreibung	Mittelwert/Häufigkeit in %	Std-abw.	Min/Max
	Einspruch gegen das Patent			
einspruch	1 = Ja	41.49		
	0 = Nein	58.51		
	Patent aus der Biotech-/ Pharma-Branche			
biopharm	1 = Ja	44.31		
	0 = Nein	55.69		
	US Zwillingspatent			
uszw	1 = Ja	60.85		
	0 = Nein	39.15		
	Patentinhaber aus USA			
patus	1 = Ja	33.74		
	0 = Nein	66.26		
	Patentinhaber aus GER, CH, UK			
patdsg	1 = Ja	23.49		
	0 = Nein	76.51		
azit	Anzahl der Zitationen	1.64	2.74	0/40
aland	Anzahl der Länder für Patentschutz	7.8	4.12	1/17
ansp	Anzahl Patentansprüche	13.13	12.09	1/355

Datenbeispiel

- Beobachtungen mit ansp>60 ausgeschlossen, da sie das Resultat stark beeinflussen
- Es bleiben $n = 4832$ Patente
- Betrachten Poisson-Modell mit $\phi = 1$ und mit $\phi \neq 1$ (Ergebnisse auf kommenden Folien)
- Dispersionsparameter im zweiten Modell wird geschätzt durch

$$\hat{\phi} = \frac{1}{n-k} \sum_{i=1}^{n} (Y_i - \hat{\lambda}_i)^2 / \hat{\lambda}_i = 3.616$$

Ergebnisse im Modell mit $\phi = 1$

Variable	Koeffizient	Std-abw.	t-Wert	p-Wert
Konstante	0.176	0.032	5.53	$<$0.001
jahrc	-0.075	0.003	-25.86	$<$0.001
alandc	-0.025	0.004	-6.12	$<$0.001
anspc	0.020	0.001	17.12	$<$0.001
biopharm	0.294	0.031	9.41	$<$0.001
uszw	-0.043	0.025	-1.77	0.077
patus	0.018	0.026	0.69	0.491
patdsg	-0.230	0.031	-7.31	$<$0.001
einspruch	0.404	0.024	16.55	$<$0.001

Ergebnisse im Modell mit $\phi \neq 1$

Variable	Koeffizient	Std-abw.	t-Wert	p-Wert
Konstante	0.176	0.060	2.91	0.004
jahrc	-0.075	0.005	-13.60	$<$0.001
alandc	-0.025	0.008	-3.22	0.001
anspc	0.020	0.002	9.01	$<$0.001
biopharm	0.294	0.059	4.95	$<$0.001
uszw	-0.043	0.047	-0.93	0.353
patus	0.018	0.050	0.36	0.717
patdsg	-0.230	0.060	-3.84	$<$0.001
einspruch	0.404	0.046	8.70	$<$0.001

Bemerkung zu den Ergebnissen

- Regressionskoeffizienten ändern sich zwischen den Modellen nicht
- Grund: Hatten gesehen dass MLE nicht von ϕ abhängen
- Standardabweichungen ändern sich, da die Fisher Matrix invers-proportional zu ϕ ist
- Änderung um Faktor $\sqrt{3.616} = 1.9$
- Teststatistiken und p-Werte ändern sich entsprechend mit