Exakte logistische Regression

Kaspar Rufibach

Uni Zürich, ISPMZ, Abt. Biostatistik

6. November 2008

Fragestellung

Beratungsfall Sauermann: Modelliere Zielgrösse

Harnleiterstein Abgang in 22 Tagen ja/nein

durch die erklärenden Variablen

- Medikament: Placebo vs. Tamsulosin,
- Grösse des Steins in cm,
- Lage des Steins: intramural vs. Prävesikal.

Univariate Analysen: Binäre Variablen

		Medikament		
		Tamsulosin	Placebo	Total
Abgang	kein Abgang	6	5	11
	Abgang	38	40	78
	Total	44	45	89

Tamsulosin:

- Kein Abgang: $5/11 = 0.45$,
- Abgang: $40/78 = 0.51$.

		Steinlage		
		praevesikal	intramural	Total
Abgang	kein Abgang	11	0	11
	Abgang	45	33	78
	Total	56	33	89

Intramural:

- Kein Abgang: $0/11 = 0.00$,
- Abgang: $33/78 = 0.42$.

Univariate Analysen: stetige Variable

Variable	**Levels**	n	Min	$\bar{x}$	Max	s
Steingrösse	kein Abgang	11	3.5	4.9	7.0	1.3
	Abgang	78	2.0	4.0	6.7	1.0
	all	89	2.0	4.1	7.0	1.1

95%-Konfidenzintervall für MW-Differenz: $[-1.71, 0.04]$.

Logistische Regression

Tests für Parameter: asymptotisch gültig $\Leftrightarrow$ lineare Regression.

Faustregel Skript: Pro stetige Einflussgrösse 10, pro binäre Einflussgrösse je 20 Ereignisse/nicht-Ereignisse.

Beispiel: 78 Ereignisse, 11 Nicht-Ereignisse $\Rightarrow$ wenige Beobachtungen, wenige Ereignisse und Nuller in Kovariable!

"Sparse", schiefe oder "heavy-tailed" Daten verursachen Probleme.

Was kann schiefgehen?

- Keine Konvergenz, lächerliche Odds Ratios (`SAS`: $>$999.99).
- "Complete separation".

Logistische Regression: ML - Schätzung

	odds ratio	CI low	CI up	*p*-value
mediPlacebo	1.07	0.26	4.43	0.93
locationintramural	58302764.65	0.00	Inf	0.99
size	0.57	0.29	1.10	0.09

Tabelle: Resultate Maximum Likelihood Schätzung.

p-Werte globale Tests:

- Likelihood-Ratio: 0.0024
- Score: 0.0150
- Wald: 0.3652

Unbedingte Likelihood

Beobachtungen $\mathbf{y}^\top = (y_1, \ldots, y_n) \in \{0,1\}^n$,
$\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^\top \in \mathbb{R}^{(p+q)\times n}$.

Unbedingte Likelihood-Funktion:

$$L(\boldsymbol{\beta}) \quad = \quad \frac{\exp(\mathbf{y}'\mathbf{X}\boldsymbol{\beta})}{\Pi_{i=1}^n (1 + \exp(\mathbf{x}_i\boldsymbol{\beta}))}$$

Maximiere $L(\boldsymbol{\beta})$ über $\mathbb{R}^{p+q}$, asymptotische Likelihood-Inferenz.

Beachte: $T_j = \sum_{i=1}^n y_i x_{ij}$ ist suffiziente Statistik für β_j.

Dichtefunktion der suffizienten Statistik

Dichtefunktion von $\mathbf{T} = (T_1, \ldots, T_{p+q})$:

$$P(\mathbf{T} = \mathbf{t}) \quad = \quad \frac{C(\mathbf{t}) \exp(\mathbf{t}'\boldsymbol{\beta})}{\Pi_{i=1}^{n}(1 + \exp(\mathbf{x}_i\boldsymbol{\beta}))},$$

wobei $C(\mathbf{t})$ die Vektoren $\mathbf{y}$ zählt, die $\mathbf{t}$ generieren.

Betrachte $\boldsymbol{\beta}_0 = (\beta_1, \ldots, \beta_p)$ als nuisance-Parameter, $\mathbf{X}_0, \mathbf{T}_0, \mathbf{t}_0$.

Wir sind interessiert an: $\boldsymbol{\beta}_1, \mathbf{X}_1, \mathbf{T}_1, \mathbf{t}_1$.

Bedingte Likelihood

Entferne nuisance-Paramter durch Bedingen auf ihre suffizienten Statistiken:

$$\begin{aligned} P(\mathbf{T}_1 = \mathbf{t}_1 | \mathbf{T}_0 = \mathbf{t}_0) &= \frac{P(\mathbf{T} = \mathbf{t})}{P(\mathbf{T}_0 = \mathbf{t}_0)} \\ &= \frac{C(\mathbf{t}) \exp(\mathbf{t}_1' \boldsymbol{\beta})}{\sum_{\mathbf{u}} C(\mathbf{u}, \mathbf{t}_0) \exp(\mathbf{u}' \boldsymbol{\beta}_1)}. \end{aligned}$$

Hier zählt $C(\mathbf{u}, \mathbf{t}_0)$ die Vektoren mit

$$\mathbf{y}'\mathbf{X}_1 = \mathbf{u} \quad \text{und} \quad \mathbf{y}'\mathbf{X}_0 = \mathbf{t}_0.$$

Dies ist die bedingte Likelihood-Funktion.

Bedingte exakte Inferenz: erzeuge exakte bedingte Verteilung für interessierende Parameter.
$\Rightarrow$ Permutations-Verteilung.

Die exakte bedingte Verteilung

Wie wahrscheinlich ist beobachteter Response-Vektor $\mathbf{y} = (y_1, \ldots, y_n)$ unter allen 2^n möglichen Response-Vektoren?

- Erzeuge alle Vektoren $\mathbf{y}$ mit $\mathbf{y}'\mathbf{X}_0 = \mathbf{t}_0$.
- Zähle alle Vektoren für die für jedes eindeutige $\mathbf{t}_1$

$$\mathbf{y}'\mathbf{X}_1 \quad = \quad \mathbf{t}_1.$$

Konzeptionell einfach. Aber: Wir müssen immer 2^n verschiedene Vektoren $\mathbf{y}$ betrachten!

Lösung: multivariate shift algorithm (Hirji, Mehta, and Patel, 1987).

Vergleich der Resultate

	odds ratio	CI low	CI up	*p*-value
mediPlacebo	1.07	0.26	4.43	0.93
locationintramural	58302764.65	0.00	Inf	0.99
size	0.57	0.29	1.10	0.09

Tabelle: Resultate Maximum Likelihood Schätzung.

	odds ratio	CI low	CI up	*p*-value
mediPlacebo	1.03	0.45	2.35	1.00
locationintramural♠	2.84	1.11	Inf	0.03
size	0.58	0.29	1.11	0.10

Tabelle: Resultate exakte Logistische Regression.

Zeit bis Abgang als Ereigniszeit: $\Rightarrow$ HR für location $= 2.62$.

♠: Beobachtung $(\mathbf{T}_1 = \mathbf{t}_1, \mathbf{T}_0 = \mathbf{t}_0)$ liegt auf dem Rand von $(\mathbf{T}_1 = \mathbf{t}_1)$
$\Rightarrow$ auch hier CMLE $= \infty$. Berechne MUE.

Zusammenfassung

Betrachte exakte logistische Regression falls mind. eines der folgenden Kriterien erfüllt:

- kleine Anzahl Beobachtungen,
- kleine Anzahl (nicht-)Ereignisse,
- Verteilung von erklärenden Variablen schief oder heavy-tailed.

Zu beachten:

- Rechenaufwand explodiert mit wachsender Anzahl Beobachtungen.
- Im Allgemeinen liefert ELR konservative Tests/KIs.

Zusammenfassung

Alternativen:

- Penalisierte Methoden $\Rightarrow$ keine p-Werte und Konfidenzintervalle!
- Bayes?

Software:

- `SAS`: `proc logistic` mit `exact` statement. Gib Variablen an, für die exakte Inferenz gewünscht.
- `Stata`: `exlogistic`.
- `R`: Keine Implementation verfügbar. Dubioses package `elrm`. ELR via MCMC, auch für grosse Datensätze.

`SAS` und `Stata` verwenden multivariate shift algorithm.

Das wars. Danke für die Aufmerksamkeit.