Missing Data: Dial M for ???

Vortrag am 27.01.2012

Leitung: Prof. Dr. Augustin
Betreuer: Dr. Marco Cattaneo

Referent: Matthias Faber

Übersicht

Gliederung:

- Einleitung
- Der Datensatz
- McKendrick's Methode
- Der Weg zum EM-Algorithmus
- Der EM-Algorithmus
- Ausblick
- Diskussion

Einleitung

- Für jegliche Arbeiten in der Statistik ist eine hohe Datenqualität der erhobenen Daten enorm wichtig

- Das Auftreten von fehlenden Werten in Daten kann die Qualität der Studienergebnisse auf verschiedene Arten beeinflussen und somit deren Aussagekraft stark einschränken

- In dieser Präsentation wird unter fehlenden Werten verstanden, dass der fehlende Wert theoretisch existiert, dieser aber nicht aus dem Datensatz erkennbar ist, also fehlt

Der Datensatz

Der Datensatz stammt von Anderson Gray McKendrick, der 1926 Informationen über eine Choleraepidemie in einem indischen Dorf sammelte und untersuchte

Table 1. Data and Fitted Values For McKendrick's Problem

x	0	1	2	3	4	≥ 5	Total
n_x	168	32	16	6	1	0	223

In dem Datensatz gibt x die Anzahl der Cholerafälle in einem Haushalt an und n_x Anzahl der tatsächlich aufgetretenen Fälle, die gezählt wurden.

McKendrick's Methode

Die Poisson-Verteilung:

McKendrick kannte den wahren Wert von λ nicht, deshalb versuchte er λ folgendermaßen zu schätzen

$$\hat{\lambda} = \sum_x x\, n_x \,/\, \sum_x n_x = 86/223 = 0.38565 \approx 0.386$$

Da $\hat{\lambda}$ das arithmetische Mittel, also den Poisson-Mittelwert von λ darstellt, muss die Poisson-Verteilung folgendermaßen aussehen

$$f(X=x) = (\sum_i n_i)\hat{\lambda}^x \exp(-\hat{\lambda})/x!$$ für x= 0,1,...

Die einfache Poisson-Verteilung wird mit dem Faktor $\sum_i n_i$ mulipliziert, weil λ in der Poisson-Verteilung den Erwartungswert darstellen muss.

Die Poisson-Verteilung

In der Tabelle sind die Werte für die Poisson-Verteilung berechnet

Table 1. Data and Fitted Values For McKendrick's Problem

x	0	1	2	3	4	≥ 5	Total
n_x	168	32	16	6	1	0	223
Direct Poisson fit	151.64	58.48	11.28	1.45	0.00	.01	223

Die einfache Poisson-Verteilung ist kein gutes Modell, um die Daten zu beschreiben.

McKendrick vermutete, dass in dem Datensatz zu viele Nullen enthalten sind, also zu viele Haushalte ohne Infektion vorliegen

Momentenschätzung

Zunächst berechnet McKendrick einen Schätzer für n:
Dazu sind folgende Hilfsgrößen notwendig:

$$s_1 = \sum_x x n_x = 86 \quad und \quad s_2 = \sum_x x^2 n_x = 166$$

$$\hat{n} = \frac{s_1^2}{s_2 - s_1} = 92.45 \approx 93$$

Im Folgenden legt man fest, dass n_x gleich der Zahl aus $\{i : x_i = x\}$ist. Damit entspricht $\sum_x x n_x = \sum_i x_i$ und der Erwartungswert kann berechnet werden:

$$E(s_1) = E(\sum x n_x) = E(\sum_i x_i) = n\lambda$$
$$E(s_2) = E(\sum_x x^2 n_x) = E(\sum_i x_i^2) \overset{Verschiebungssatz}{=} n\lambda + n\lambda^2$$
$$E(s_2 - s_1) = E(\sum_x x^2 n_x) - E(\sum_x x n_x) = E(\sum_i x^2{}_i) - E(\sum_i x_i) = n\lambda + n\lambda^2 - n\lambda = n\lambda^2$$

Daraus ergibt sich:

$$\hat{n} = \frac{E(s_1)^2}{E(s_2 - s_1)} = \frac{n^2\lambda^2}{n\lambda^2} = n$$

Momentenschätzung

McKendrick benötigt nun noch eine neue Schätzung für λ:

$$\tilde{\lambda} = \frac{s_1}{\hat{n}} = \frac{86}{93} = 0.92473 \approx 0.93$$

Einsetzten in folgende Formel ergibt die Ergebnisse:

$$f(X=x) = \hat{n}\,\tilde{\lambda}^{x} \exp(-\tilde{\lambda})/x!, \; \textit{für } x = 0,1,\ldots,\geq 5$$

Table 1. Data and Fitted Values For McKendrick's Problem

x	0	1	2	3	4	≥5	Total
n_x	168	32	16	6	1	0	223
Direct Poisson fit	151.64	58.48	11.28	1.45	0.00	.01	223
McKendrick's fit	36.89	34.11	15.77	4.86	1.12	.24	93

McKendrick's Methode

McKendrick (1926, S.101) stellt fest: “This suggests that the disease was probably water borne, that there were a number of wells, and that inhabitants of 93 out of 223 houses drank from one well which was infected. On further local investigation it was found that there was one particular infected well from which a certain section of the community drank.“

Vorteile von McKendrick's Vorgehensweise:

- einfache Berechnung
- relativ gute Schätzung der Werte
- nicht Beachtung des Werts n_0 , da er wenig Information über den wahren Wert, also über die Haushalte, die dem Cholerabakterium ausgesetzt waren, aber sich nicht infizierten, besitzt

McKendrick's Berechnungen bilden die ursprüngliche Grundlage für den Expectation-Maximization Algorithmus

Der Weg zum EM-Algorithmus

Joseph Oscar Irwin verbesserte McKendrick's Formel 1963, indem er Iterationen einbaute.

$$n^{(t+1)} = n^{(t)} \exp(-\lambda^{(t)}) + n_{obs} \quad (1.1)$$

$$mit\ n_{obs} = \sum_{x \geq 1} n_x\ und\ (t)\ gleich\ den\ \ Indexiterationen.$$

Hier gibt lt. Meng (1997, S. 5) $n \exp(-\lambda^{(t)})$ die erwartete Anzahl an Nullen in einem Stichprobenumfang von n an.

In einem zweiten Schritt wird nun ein neues λ berechnet:

$$\lambda^{(t+1)} = \frac{s_1}{n^{t+1}} \quad (1.2)$$

Der Weg zum EM-Algorithmus

- Die Berechnungen werden abwechselnd zwischen (1.1) und (1.2) ausgeführt, bis der Wert für λ konvergiert. Der Wert konvergiert, sobald $|\lambda^{(t+1)} - \lambda^{(t)}| \leqslant 0.0001$ erreicht ist

- Irwin's Methode konvergiert nach 24 Iterationsschritten von einem Startwert λ = 0.93 gegen λ = 0.97218, vgl. Meng (1997, S.9)

- Irwin's Methode bildet die Grundlage des EM-Algorithmus

- Jedoch unterscheidet sich seine Methode gerade in der Iterations-geschwindigkeit deutlich vom EM-Algorithmus

Der EM-Algorithmus

- Der Expectation-Maximization (EM) Algorithmus, nach Dempster (1977, S.1), ist ein vielseitig anwendbarer Algorithmus zur Berechnung der Maximum-Likelihood Schätzung mit Hilfe von Iterationen

- Zwei Schritte: Der Expectation (E-Step) und der Maximization (M-Step) Schritt

- Die Idee ist, die fehlenden Werte durch geeignete Erwartungswerte zu ersetzen (E-Step), die Parameter neu zu schätzen (M-Step) mit diesen neuen Parametern E-Step und M-Step zu wiederholen bis sich die Schätzungen der Parameter im M-Step nicht mehr verändern

-Man unterscheidet die unvollständigen (aber beobachteten) Daten y von den vollständigen (aber teilweise unbeobachteten) Daten x

-Ziel ist es, die Likelihood $L(\theta, y)$ bezüglich θ zu maximieren, während man nur $L(\theta,x)$, bzw. $l(\theta,x)$ verwendet

Der EM-Algorithmus

Theoretische Vorüberlegungen:

1. Schritt: Überlegung eines Startwertes $\theta^{(0)}$

2. Schritt: E-Step: Berechnung des bedingten Erwartungswerts:

$$Q(\theta)=Q(\theta|\theta^{(t)})=E(l(\theta,x)\,|\,y,\theta^{(t)})$$

3. Schritt: M-Step: Maximieren $Q(\theta|\theta^{(t)})$ von Funktion von θ, um den Wert $\theta^{(t+1)}$ zu erhalten. Hierbei muss gelten, dass $Q(\theta^{(t+1)}|\theta^{(t)})\geq Q(\theta|\theta^{(t)})$

4. Schritt: Iterationen solange ausführen, bis die Schätzung konvergiert

Der EM-Algorithmus

Der EM-Algorithmus für McKendrick's Daten:

Zunächst stellt Y_{obs} die beobachteten Daten (wie z.B. $\{n_x, x \geq 1\}$) und Y_{mis} die fehlenden Daten (wie z.B. n_0) dar. Somit sind die vollständigen Daten definiert als $Y = (Y_{obs}, Y_{mis})$

Im E-Step des Algorithmus muss nun $Q(\lambda) = Q(\lambda|\lambda^{(t)}) = E(l(\lambda, x) \mid y, \lambda^{(t)})$ bestimmt werden. Davon ausgehend, dass n_0 beobachtet worden ist, ergibt sich für die vollständige Likelihood-Funktion:

$$L(\lambda|Y) = \prod_x (\lambda^x \exp(-\lambda))^{n_x}$$

Die Konstante x! fällt bei der Betrachtung der Likelihood heraus, da sie für die Schätzung der ML-Schätzers keine Rolle spielt. Daraus ergibt sich die Log-Likelihoodschätzung:

$$l(\lambda|Y) = (\sum_x x n_x) \log \lambda - n\lambda \qquad (2.1)$$

Der EM-Algorithmus

Der E-Step:

$$Q(\lambda|\lambda^{(t)}) = E(l(\lambda, Y) \mid Y_{obs}, \lambda^{(t)}) = E(l(\lambda|Y) \mid n_1, n_2, \ldots, \lambda^{(t)}) = (\sum_x x n_x) \log\lambda - n_{obs}\lambda - E(n_0|\lambda^{(t)}, Y_{obs})$$

Es muss nur noch der bedingte Erwartungswert für n_0 berechnet werden:

$$n_0^{(t+1)} = E(n_0|\lambda^{(t)}; Y_{obs}) = \frac{W'keit\ nicht\ angesteckt\ zu\ sein}{W'keit\ angesteckt\ zu\ sein} n_{obs}$$

$$n_0^{(t+1)} = \frac{\exp(-\lambda^{(t)})}{1-\exp(-\lambda^{(t)})} n_{obs} \qquad (2.2)$$

Der EM-Algorithmus

Der M-Step: Ableiten der Log-Likelihood (2.1) und gleich Null setzen:

$$l(\lambda|Y)' = \frac{(\sum_x xn_x)}{\lambda} - n = 0$$

$$\hat{\lambda} = \frac{\sum_x xn_x}{n} = \frac{\sum_x xn_x}{n_0 + n_{obs}}$$

Umschreiben der Schätzung in:

$$\hat{\lambda} = \frac{\sum_x xn_x}{n_0^{(t+1)} + n_{obs}} \qquad (2.3)$$

Kombination der Formeln von (2.2) und (2.3):

$$\hat{\lambda}^{(t+1)} = \frac{\sum_x xn_x(1 - \exp(-\lambda^{(t)}))}{n_{obs}} \qquad (2.4)$$

Der EM-Algorithmus

Die Schritte (2.2) und (2.4) müssen bis zur Konvergenz ausgeführt werden:

$$|\lambda^{(t+1)} - \lambda^{(t)}| \leqslant 0.0001$$

Aus der Berechnung mit R erhält man:

n_t	$\lambda^{(t+1)}$	t
35.8420795	0.9466978	1
34.8722499	0.9569138	2
34.2967864	0.9630806	3
33.9557653	0.9667726	4
33.7538327	0.9689722	5
33.6343161	0.9702788	6
33.5635983	0.9710536	7
33.5217615	0.9715125	8
33.4970133	0.9717842	9
33.482374	0.971945	10
33.4737158	0.9720401	11

Der EM-Algorithmus

Die Werte sind nun in folgender Tabelle zusammengefasst:

Table 1. Data and Fitted Values For McKendrick's Problem

x	0	1	2	3	4	≥ 5	Total
n_x	168	32	16	6	1	0	223
Direct Poisson fit	151.64	58.48	11.28	1.45	0.00	.01	223
McKendrick's fit	36.89	34.11	15.77	4.86	1.12	.24	93
MLE fit	33.46	32.53	15.81	5.12	1.25	.29	88.46

- Irwin's Methode konvergiert gegen dasselbe λ, nämlich 0.972

- Der EM-Algorithmus bietet die deutlich schnellere Methode, da Irwin's Methode 24 Iterationsschritte benötigte, vgl. Meng(1997, S. 9)

- Die Schätzung des EM-Algorithmus liefert noch bessere Ergebnisse als das Modell von McKendrick

Ausblick auf eine andere Methode

- Bei zu häufigen Vorkommen der Null wird ein Mischmodell aus Binomial- und Poissonverteilung gefittet

- Für McKendricks Daten ist das ein binomialer Indikator für den Status „Cholera-Erreger ausgesetzt“ ja/nein und unter der Bedingung, dass man mit dem Bakterium in Kontakt war, eine Possionverteilung

- Dieses Mischmodel wird in der Literatur auch oft als „Zero-Inflated Poissonmodell“ (ZIP) bezeichnet, siehe Bohning et al. (1999, S. 198)

- In R gibt es die Möglichkeit ein ZIP-Modell für die Daten zu fitten. Dazu benötigt man das Package „VGAM“, da die Daten von McKendrick als Vektor übergeben werden

Ausblick auf eine andere Methode

Berechnet man nun die Daten in R, so ergibt sich folgende Schätzung:

	Original	ZIP_Modell
[1,]	168	168.000000
[2,]	32	32.530218
[3,]	16	15.812579
[4,]	6	5.124213
[5,]	1	1.245412

- Das ZIP-Modell schätzt die richtige Anzahl für x = 0

- Die anderen Werte für $x \geq 1$ im Vergleich zu den Schätzungen des EM-Algorithmus ändern sich nicht mehr

Zusammenfassung

- McKendrick's Methode ist sehr einfach anzuweden und liefert eine erste Schätzung der Daten

- Der Momentenschätzer lässt sich aber nicht für jeden Datensatz reibungslos berechnen

- Irwin verbesserte McKendrick's Methode, indem er Iterationen in die Formel einbaute

- Die Iterationsgeschwindigkeit ist im Vergleich zu dem EM-Algorithmus jedoch deutlich langsamer

- Der EM-Algorithmus ist ein iterativer Algorithmus zur Berechnung der Maximum-Likelihood Schätzung mit Hilfe von Iterationen

- Obwohl der EM-Algorithmus sehr hilfreich ist, um den Maximum-Likelihood-Schätzer zu berechnen, ist dieser weder in der Lage Fehler in einer Modellannahme zu korrigieren, noch zusätzliche Information in die Studie einfließen zu lassen

Diskussion