Lehrstuhl für Statistik und Ökonometrie
Übung zur Datenanalyse

Aufgabenserie 8: Log-lineare Modelle

Aufgabe 17

Im Internet existieren zahlreiche soziale Netzwerke wie facebook.com. Hier können Nutzer ihre privaten aber auch ihre beruflichen Kontakte pflegen. Es soll nun untersucht werden, ob ein Zusammenhang zwischen der "Art" der Nutzung (Merkmal A) und der "Anzahl der registrierten Nutzer" (Merkmal B) eines sozialen Netzwerks besteht. Zu diesem Zweck wurden von je 20 sozialen Netzwerken, die von den Nutzern überwiegend für private bzw. berufliche Zwecke verwendet werden, die Nutzerzahlen erhoben. Hierbei ergab sich folgende Kontingenztabelle:

	Anzahl der registrierten Nutzer (B)		
Art (A)	bis 5 Mio.	über 5 Mio.	
beruflich	15	5	20
privat	7	13	20
	22	18	40

1. Welches Erhebungsschema wurde verwendet?
2. Nennen und erläutern Sie weitere Erhebungsschemata für das log-lineare Modell.
3. Schätzen Sie das Kreuzproduktverhältnis α, und interpretieren Sie den Wert.
4. Stellen Sie das Unabhängigkeitsmodell inkl. der Normierungsbedingungen auf, und schätzen Sie die Parameter des Modells.
5. Betrachten Sie nun das saturierte Modell.
   (a) Stellen Sie das Modell inkl. der Normierungsbedingungen auf, und schätzen Sie die Parameter.
   (b) Geben Sie ausgehend von der Darstellung $\ln \boldsymbol{m} = \boldsymbol{X}\boldsymbol{\mu}$ die Designmatrix $\boldsymbol{X}$ und den Schätzer des Parametervektors $\boldsymbol{\mu}$ an.
6. Mithilfe welchen Verfahrens könnten Sie einen möglichen Zusammenhang zwischen Art und Nutzerzahl eines sozialen Netzwerks untersuchen, wenn Ihnen für jedes der 40 sozialen Netzwerke die exakte Anzahl der Nutzer bekannt wäre?
7. Testen Sie bei einer Irrtumswahrscheinlichkeit von 5% das saturierte Modell gegen das Unabhängigkeitsmodell.

8. Für den Parameter $\mu_{AB(11)}$ des saturierten Modells soll ein Signifikanztest auf einem Signifikanzniveau von $\alpha = 0.05$ durchgeführt werden.

   (a) Welches Testergebnis würden Sie auf Basis des Ergebnisses von Teilgabe 7 erwarten?

   (b) Führen Sie den entsprechenden Test durch. Für das Poisson-Erhebungsschema sei Ihnen die geschätzte Varianz-Kovarianz-Matrix von $\widehat{\boldsymbol{\mu}}$ gegeben als:

$$\widehat{\boldsymbol{V}} = \begin{pmatrix} 0.0304 & -0.0042 & 0.0029 & -0.0125 \\ -0.0042 & 0.03040 & -0.0125 & 0.0029 \\ 0.0029 & -0.0125 & 0.0304 & -0.0042 \\ -0.0125 & 0.0029 & -0.0042 & 0.0304 \end{pmatrix}.$$

Zusätzlich ist bekannt, ob die Internetseiten eines sozialen Netzwerkes in einer oder in mehreren Sprachen verfügbar sind, d.h. betrachtet wird die weitere qualitative Variable C "verfügbare Sprachen" mit den Ausprägungen C_1:= "eine" und C_2:= "mehrere".

9. Zur Erklärung der Zellhäufigkeiten m_{ijk} wird folgendes Modell unterstellt:

$$\ln m_{ijk} = \mu + \mu_{A(i)} + \mu_{B(j)} + \mu_{C(k)} + \mu_{AC(ik)} + \mu_{BC(jk)}, \quad i = 1, 2,\ j = 1, 2,\ k = 1, 2.$$

   Was impliziert dieses Modell im vorliegenden Sachverhalt, und was würde seine Gültigkeit für die Ergebnisse aus den Teilaufgaben 7 und 8(b) bedeuten?

10. Die zugehörigen Daten liegen Ihnen im Datensatz `net` vor. Schätzen Sie in `R`

    (a) ein saturiertes Modell,

    (b) ein Modell, bei dem die Assoziation zwischen der Art der Nutzung und der Anzahl der Nutzer nicht von den verfügbaren Sprachen abhängt,

    (c) das Modell aus Teilaufgabe 9,

    (d) das Modell der totalen Unabhängigkeit.

11. Wählen Sie mithilfe von Likelihood-Verhältnis-Tests aus der Folge der geschachtelten Modelle aus Teilaufgabe 10 ein geeignetes Modell aus.