12 Ergänzungen zu kategoriellen Regressionsmodellen

Notizen und Stichworte.

12.1 Korrespondenz-Analyse

a Die Korrespondenz-Analyse ist ein grafisches Mittel zur Veranschaulichung von Datenmatrizen. Sie ist mit dem Biplot und damit mit der Hauptkomonenten-Analyse eng verwandt. Sie wurde ursprünglich zur Darstellung von Kreuztabellen entwickelt, ist seither aber erweitert worden für andere Arten von Datenmatrizen.

Hier soll nur die Interpretation des Resultats diskutiert werden. Auf den Biplot werden wir im Block Mu-2a zurückkommen.

b Tabelle 12.1.b zeigt für eine Umfrage bei 193 Angestellten einer grossen Firma den Zusammenhang zwischen Angestellten-Kategorien und Rauch- resp. Trinkgewohnheiten.

	smoking				drinking		
Anzahlen	no.sm	light.sm	medium.sm	heavy.sm	no.alc	yes.alc	total
sen.man	4	2	3	2	0	11	11
jun.man	4	3	7	4	1	17	18
sen.emp	25	10	12	4	5	46	51
jun.emp	18	24	33	13	10	78	88
secr	10	6	7	2	7	18	25
total	61	45	62	25	23	170	193

Zeilen-%	no.sm	light.sm	medium.sm	heavy.sm	no.alc	yes.alc	total
sen.man	36	18	27	18	0	100	100
jun.man	22	17	39	22	6	94	100
sen.emp	49	20	24	8	10	90	100
jun.emp	20	27	38	15	11	89	100
secr	40	24	28	8	28	72	100
total	32	23	32	13	12	88	100

Tabelle 12.1.b: Rauch- und Trink-Gewohnheiten von Angestellten in einer grossen Firma: Anzahlen und Zeilenprozentwerte

Abbildung 12.1.b: Korrespondenz-Analyse-Diagramm für das Beispiel der Rauch- und Trink-Gewohnheiten

c Wir führen „Distanzen" zwischen standardisierten Zeilen $\underline{\widetilde{x}}_h = [x_{hk}/x_{h+}]$ $(x_{h+} = \sum_k x_{hk})$ ein:

$$d\langle h, h' \rangle = \sum_k (\widetilde{x}_{hk} - \widetilde{x}_{h'k})^2 / x_{+k} \; .$$

Sie sollen möglichst gut dargestellt werden in zwei Dimensionen. Wie bei der Hauptkomponenten-Analyse müssen diese als Linearkombinationen der Zeilen von X zustandekommen. – Analog für Kolonnen. Beide Darstellungen werden kombiniert. So bedeuten grafische Beziehungen zwischen den Punkten der Darstellung näherungsweise folgendes:

- Zwei Zeilen- (Spalten-) Punkte, die nahe beieinander liegen, bedeuten ähnliche Proportionen der Zeilen (Spalten).
- Zeilen- (Spalten-) Punkte, die nahe beim Nullpunkt liegen, bedeuten „durchschnittliche" Proportionen.
- Liegen ein Zeilenpunkt und ein Spaltenpunkt in ähnlicher Richtung vom Nullpunkt weg, so ist die entsprechende Kombination übermässig häufig in den Daten.

Die gesamte Darstellung zeigt also die Abweichungen der Kreuztabelle von der Unabhängigkeit.

d Im Beispiel sind eigentlich zwei Kreuztabellen dargestellt. Die Rauch-Gewohnheiten wurden zur Berechnung der Darstellungs-Achsen verwendet. Die Trink-Gewohnheiten sind Spalten, die auf die gleiche Art wie die anderen Spalten dargestellt werden. Ebenso werden als zusätzliche Zeile die durchschnittlichen Rauch-Gewohnheiten über die ganzen USA dargestellt.

12.2 Kombination unabhängiger Tests, Meta-Analyse

a Beispiel Herzinfarkte und Verhütungsmittel: Studie an mehreren Spitälern = Multicenter-Studie.
Ergibt mehrere Vierfeldertafeln mit verschiedenen Randsummen.
Kombinierte Evidenz?

„Beispiel":

Teststatistik	1.50	2.30	5.10	0.90	3.20
P.Wert	0.22	0.13	0.02	0.34	0.07

Ist damit der Zusammenhang bewiesen?

b Möglichkeiten für einen Gesamttest:

- Teststatistiken zusammenzählen: $T = \sum T_\ell \sim \chi^2_m$.
- „Mantel-Haenszel-Statistik": $U = \sum_\ell N_{11,\ell}$. Erwartungswert und Varianz bestimmen. $(U - \mathcal{E}\langle U\rangle)/\sqrt{\mathrm{var}\langle U\rangle}$ mit $\mathcal{N}\langle 0,1\rangle$ oder das Quadrat mit χ^2_1 vergleichen.
- P-Werte mitteln (arithmetisch, geometrisch, ...). Verteilung bestimmen ist nicht so schwierig.

Die dritte Möglichkeit und Varianten davon können auch eingesetzt werden, wenn die P-Werte von ganz verschiedenartigen Tests stammen.

12.3 Discrete Choice Models

a Antwortgrösse: Wahl einer Kategorie
aus einer Auswahl von Möglichkeiten.
Beispiel: Verkehrsmittel.

Erklärende Variable I: charakterisieren die Möglichkeiten.
Beispiel: Fahrzeiten, Umsteigen, ...

Erklärende Variable II: Gewöhnliche erklärende Variable,
unabhängig von der Kategorie der Antwortgrösse.
Beispiel: Alter, Geschlecht, Einkommen, ...

führt zu Multinomialer Regression (multinomial logit models).

b Literatur: Agresti, Kap. 9, Fahrmeir and Tutz (2001), Kap. 3.2.