Vorlesung: Lineare Modelle

Prof. Dr. Helmut Küchenhoff

Institut für Statistik, LMU München

SoSe 2015

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN | Institut für Statistik | STABLAB Statistisches Beratungs Labor

Modelle mit diskreten Einflussgrößen

Bei der ANOVA geht es um den Vergleich von Mittelwerten.

Die einfaktorielle Varianzanalyse hat die primäre Fragestellung:
Sind die Mittelwerte von verschiedenen Gruppen gleich?

Diese Frage lässt sich als lineares Modell darstellen. Wir verwenden eine diskrete Variable, die die Gruppenzugehörigkeit beschreibt.

Dummycodierung

Wir betrachten ein nominales Merkmal C mit K Ausprägungen.

a) Einfache Dummy-Kodierung

$$Z_k(C) = \begin{cases} 1 \text{ für } C = k; \\ 0 \text{ für } C \neq k; \end{cases} \quad k = 1, \ldots, K \tag{4.1}$$

b) Effekt-Kodierung

$$Z_k^e(C) = \begin{cases} 1 \text{ für } C = k; \\ 0 \text{ für } C \neq k; C \neq K \\ -1 \text{ für } C = K; \end{cases} \quad k = 1, \ldots, K-1; \tag{4.2}$$

Einfache Varianzanalyse

Gegeben sei eine nominale Einflussgröße C mit K Ausprägungen (Gruppen). Der Zielgrößenvektor Y wird in die K Gruppen mit jeweils n_k Beobachtungen aufgeteilt:

$$Y = (Y_{11}, \ldots Y_{1n_1}, Y_{21}, \ldots, Y_{Kn_K})'$$

a) Mittelwertsmodell:

$$Y_{kl} = \mu_k + \varepsilon_{kl} \qquad l = 1, \ldots, n_k;\ k = 1, \ldots, K$$

$$Y = (Z_1(C) \ldots Z_K(C)) \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_K \end{pmatrix} + \varepsilon \qquad (4.3)$$

Beispiel

Design-Matrix X für $K = 3$ Gruppen mit je $n_k = 2$ Beobachtungen pro Gruppe:

Die Regressionsgleichung lautet:

$$\begin{pmatrix} Y_{11} \\ Y_{12} \\ Y_{21} \\ Y_{22} \\ Y_{31} \\ Y_{32} \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{pmatrix} + \begin{pmatrix} \varepsilon_{11} \\ \varepsilon_{12} \\ \varepsilon_{21} \\ \varepsilon_{22} \\ \varepsilon_{31} \\ \varepsilon_{32} \end{pmatrix}$$

b) Effekt-Kodierung:

$Y_{kl} = \mu + \tau_k + \varepsilon_{kl}; \qquad \sum_{k=1}^{K} \tau_k = 0$

$$Y = (e \ \ Z_1^e(C) \ldots Z_{K-1}^e(C)) \begin{pmatrix} \mu \\ \tau_1 \\ \vdots \\ \tau_{K-1} \end{pmatrix} + \varepsilon \qquad (4.4)$$

Design-Matrix X für $K = 3$ Gruppen mit je $n_k = 2$ Beobachtungen pro Gruppe.

$$X = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & -1 & -1 \\ 1 & -1 & -1 \end{pmatrix}$$

c) Modell mit Referenzkategorie K:

$Y_{kl} = \mu_K + \tau_k + \varepsilon_{kl}, \quad \tau_K = 0;$

$$Y \quad = \quad (e \ \ Z_1(C) \ldots Z_{K-1}(C)) \begin{pmatrix} \mu_K \\ \tau_1 \\ \vdots \\ \tau_{K-1} \end{pmatrix} + \varepsilon \qquad (4.5)$$

Design-Matrix X für 3 Gruppen mit je 2 Beobachtungen pro Gruppe:

$$X = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix}$$

Nullhypothesen zum Test auf „Effekt von C“

a) $H_0 : \mu_1 = \mu_2 = \ldots = \mu_K$

b) $H_0 : \tau_1 = \tau_2 = \ldots = \tau_{K-1} = 0$

c) $H_0 : \tau_1 = \tau_2 = \ldots = \tau_{K-1} = 0$

Zusammenhang zwischen Kodierungen

Mittelwertsmodell:	$\mu_1, \ldots, \mu_K$	kein Intercept
Effektkodierung:	$\mu = \frac{1}{K} \sum_{k=1}^{K} \mu_k$	$\leftrightarrow$ Intercept
	$\tau_k = \mu_k - \mu$ $\Leftrightarrow \mu_k = \mu + \tau_k$	„Unterschied zum Gesamtmittel“
Referenzkodierung:	$\mu = \mu_K$	$\leftrightarrow$ Intercept
	$\tau_k = \mu_k - \mu_K$ $\Leftrightarrow \mu_k = \mu_K + \tau_k$	„Unterschied zur Referenz“

Bemerkungen

- Alle 3 Kodierungen führen zu gleicher Modellanpassung (R^2)
- Parameter haben unterschiedliche Interpretation
- Parameter und deren Schätzungen aus verschieden Varianten direkt ineinander überführbar
- Modelle erweiterbar mit zusätzlichen Einflussgrößen

Modell der zweifaktoriellen Varianzanalyse

Wir betrachten zwei diskrete Einflussgrößen C und D mit K_1 bzw. K_2 Ausprägungen. Man spricht dann von einer zweifaktoriellen Varianzanalyse mit einem K_1-stufigen und einem K_2-stufigen Faktor. Hier ist die Mittelwertsdarstellung nicht möglich.