

## 6.2 Die Varianzanalyse und das lineare Modell

Man kann die Varianzanalyse auch in einem linearen Modell darstellen. Im univariaten einfaktoriellen Fall lautet die Gleichung des linearen Modells in Komponentenschreibweise:

$$(*) \quad Y_{ij} = \beta_0 + \beta_j + E_{ij}, \quad \text{mit } i = 1, \dots, n_j \text{ und } j = 1, \dots, k.$$

In unserem Beispiel ist  $k = 3$  und  $n_j = 5$ .

Es folgt für die oben eingeführten Parameter  $\beta_0$  und  $\beta_j$ :

$$\beta_0 = \mu = \frac{1}{k} \sum_{j=1}^k \mu_j \quad \text{und} \quad \beta_j = \mu_j - \mu.$$

Hieraus ergibt sich die so genannte Reparametrisierungsbedingung:

$$\sum_{j=1}^k \beta_j = 0$$

Die Hypothesen der Varianzanalyse lauten:

$$(1) \quad H_0: \mu_1 = \mu_2 = \dots = \mu_k = \mu$$

gegen

$$H_1: \text{Es existiert ein } j \in \{1, 2, \dots, k\} \text{ mit } \mu_j \neq \mu$$

Bezogen auf das lineare Modell lauten die Hypothesen:

$$(2) \quad H_0: \beta_j = 0 \text{ für } j \in \{1, \dots, k\}$$

gegen

$$H_1: \beta_j \neq 0 \text{ für mindestens ein } j \in \{1, \dots, k\} .$$

Bei der Verwendung eines linearen Modells ist folgendes zu beachten: Man kann sowohl ein Modell unter Einbeziehung von  $\beta_0$  (mit „Achsenabschnitt“ bzw. mit Konstante), als auch ein Modell ohne diesen Achsenabschnitt verwenden. Bei einem Modell mit Achsenabschnitt sind die Hypothesen (1) und (2) äquivalent. Rechnet man aber mit einem Modell ohne Achsenabschnitt (Modellgleichung:  $Y_{ij} = \mu_j + E_{ij} = \beta_j + E_{ij}$ ), so gilt:  $\mu_j = \beta_j$ , womit die beiden Hypothesen nicht mehr äquivalent sind. Die Nullhypothese (2) wäre dann äquivalent zur Hypothese, dass alle Erwartungswerte  $\mu_j$  gleich Null sind, gegen die Alternativhypothese, dass mindestens ein Erwartungswert ungleich Null ist.

Zusammenfassend gilt: In einem Modell mit Achsenabschnitt ist die Hypothese (2) äquivalent zur Hypothese (1).

In Matrix Vektor Schreibweise lautet das lineare Modell allgemein:

$$\bar{Y} = X\bar{\beta} + \bar{E}$$

Der hierin auftretende Vektor  $\bar{Y}$  ergibt sich dadurch, dass die zufälligen Größen  $Y_{ij}$  derart untereinander angeordnet werden, dass sie folgenden Spaltenvektor bilden:

$$\bar{Y} = (Y_{11}, Y_{21}, \dots, Y_{n_1 1}, Y_{12}, Y_{22}, \dots, Y_{n_2 2}, \dots, Y_{n_k k})^T$$

In unserem Beispiel hat die Designmatrix die folgende Gestalt, wie man mit der Gleichung (\*) erkennen kann:

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

Der Vektor  $\bar{\beta}$  hat vier Komponenten:

$$\bar{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix}$$

Die erste Spalte der Designmatrix X enthält aufgrund des verwendeten Achsenabschnitts  $\beta_0$  nur Einsen. Die zweite Spalte enthält jeweils eine Eins in der Zeile, in der die Komponente des Vektors  $\bar{y}$  eine Beobachtung der ersten Gruppe enthält. Es stehen somit  $n_1 = 5$  Einsen oben in der zweiten Spalte. Danach folgen Nullen. Analog enthält die dritte Spalte in den Zeilen Einsen, in denen die Komponente des Vektors  $\bar{y}$  eine Beobachtung der zweiten Gruppe enthält und sonst nur Nullen u.s.w..

Hier tritt nun das Problem auf, dass die Designmatrix  $X$  von vorne herein nicht mehr, wie bei Regressionsanalyse, spaltenregulär ist. Wie Sie sehen, ergibt sich die erste Spalte als Summe der zweiten bis vierten Spalte. Wir können das Problem lösen, indem wir eine Spalte der Designmatrix, z.B. die erste, die nur aus Einsen besteht, streichen. Dies führt zu einem Modell ohne Achsenabschnitt, wie bereits beschrieben. Man könnte z.B. auch die letzte Spalte streichen. Dabei bleibt dann der Achsenabschnitt in der Modellgleichung erhalten. Je nachdem, wie man hier vorgeht, ist der Parametervektor  $\bar{\beta}$  (der dann natürlich eine Komponente weniger enthält) auf eine andere Art zu interpretieren. Streicht man die erste Spalte der Designmatrix, dann enthält der Schätzer für den unbekannt Parametervektor die jeweiligen Gruppenmittelwerte (als Schätzer für die entsprechenden Erwartungswerte). Im zweifaktoriellen Fall müssten entsprechend zwei Spalten gestrichen werden. Auf diese Möglichkeiten, eine spaltenreguläre Designmatrix zu erzeugen, gehen wir gleich noch genauer ein.

Wir gehen außerdem davon aus, dass die Werte in der Designmatrix voreingestellt (d.h. nicht stochastisch) sind. Es handelt sich also um eine Varianzanalyse mit festen Effekten. Dies ist in unserem Beispiel der Fall, da wir drei Gruppen von Personen gewählt haben und nicht zufällig drei Gruppen entstanden sind. Die einzige stochastische Größe auf der rechten Seite der Modellgleichung ist also der Fehler(zufalls)vektor  $\bar{E}$ , dessen Komponenten  $E_{ij}$ , wie bereits beschrieben, normalverteilt sind mit dem Erwartungswert  $\mu_j$  und der Varianz  $\sigma^2$ . Da die Komponenten von  $\bar{E}$  paarweise stochastisch unabhängig sind gilt:  $\text{Var}(\bar{E}) = \sigma^2 I$

Kommen wir nun zur Parameterschätzung. Den Parametervektor  $\bar{\beta}$  schätzen wir (wie bei der Regression) über die Methode der kleinsten Quadrate, d.h. wir verwenden denjenigen Vektor  $\hat{\bar{\beta}}$  als Schätzer, der die folgende Funktion  $Q$  minimiert:

$$Q(\bar{\beta}) = (\bar{y} - X\bar{\beta})^T (\bar{y} - X\bar{\beta})$$

Mit den Methoden der Analysis kann gezeigt werden, (wie bei der Regressionsanalyse) dass

$$\hat{\beta} = (X^T X)^{-1} X^T \bar{y}$$

die Funktion Q minimiert, falls  $X^T X$  positiv definit ist. Dies gilt immer, falls X spaltenregulär ist.

In unserem Beispiel sieht die Designmatrix X in einem Modell ohne Achsenabschnitt wie folgt aus:

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

Es folgt die Schätzung des unbekannt Parametervektors  $\bar{\beta}$  durch mit Hilfe der Methode der kleinsten Quadrate:

$$\hat{\beta} = (X^T X)^{-1} X^T \bar{y} \approx \begin{pmatrix} 11,8 \\ 10 \\ 6 \end{pmatrix}$$

Dies sind genau die Mittelwerte der Teilstichproben, die auch im Output zu finden sind.

Wie bereits beschrieben, gäbe es mehrere Möglichkeiten der Verwendung einer spaltenregulären Designmatrix.

Eine Möglichkeit (wir bezeichnen die Möglichkeit der Streichung der ersten Spalte der ursprünglich nicht spaltenregulären Designmatrix als die erste Methode), eine spaltenreguläre Designmatrix zu erhalten, besteht darin, die folgende Kodierung vorzunehmen:

$X_{i0} = 1$ , d.h. die erste Spalte enthält nur Einsen.

$$X_{ij} = \begin{cases} 1 & , \text{ falls die Beobachtung in der } i\text{-ten Zeile von } \bar{y} \text{ der } j\text{-ten } (j = 1, 2, \dots, k-1) \text{ Gruppe angehört} \\ 0 & , \text{ falls die Beobachtung in der } i\text{-ten Zeile von } \bar{y} \text{ nicht der } j\text{-ten } (j = 1, 2, \dots, k-1) \text{ Gruppe angehört} \\ -1 & , \text{ falls die Beobachtung in der } i\text{-ten Zeile von } \bar{y} \text{ der } k\text{-ten Gruppe angehört} \end{cases}$$

Diese Kodierung ergibt sich durch die Reparametrisierungsbedingung:

$$\sum_{j=1}^k \beta_j = 0 \Leftrightarrow -\sum_{j=1}^{k-1} \beta_j = \beta_k$$

Der Vorteil dieser Kodierung liegt darin, dass man auch bei zweifaktoriellen Modellen eine spaltenreguläre Designmatrix erhält, was beim Streichen der ursprünglich ersten Spalte nicht der Fall ist.

Bei der Kodierung der zweiten Faktorvariablen kann dann analog vorgegangen werden.

In unserem Beispiel würde sich mit der oberen Kodierung die folgende Designmatrix ergeben:

$$(3) \quad X = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & -1 & -1 \\ 1 & -1 & -1 \\ 1 & -1 & -1 \\ 1 & -1 & -1 \\ 1 & -1 & -1 \end{pmatrix}$$

Und  $\hat{\beta} = (X^T X)^{-1} X^T \bar{y}$  ergäbe:

$$\hat{\beta} \approx \begin{pmatrix} 9,267 \\ 2,533 \\ 0,733 \end{pmatrix}$$

Eine weitere Möglichkeit für die Wahl einer Designmatrix wäre die, dass man die letzte Spalte der ursprünglichen Designmatrix streicht:

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

Mit dieser Designmatrix ergibt sich der folgende Schätzer:

$$\hat{\beta} \approx \begin{pmatrix} 6 \\ 5,8 \\ 4 \end{pmatrix}$$