10.4 Alternative Verfahren

Im vorherigen Abschnitt haben wir Varianzkomponenten mit Hilfe der Varianzanalyse geschätzt. Wir haben gesehen, dass Schätzungen für die Varianzkomponenten berechnet werden können, indem wir die Beziehung zwischen den Erwartungswerten von Summenquadraten und den Varianzkomponenten verwendeten. Anstelle der Erwartungswerte wurden die empirischen Summenquadrate den Schätzwerten für die Varianzkomponenten gleichgesetzt.

Ein Nachteil der Varianzanalyse ist, dass sie in Abhängigkeit der Datenkonstellation negative Schätzwerte liefern kann. Da Varianzkomonenten als Quadrate definiert sind, und eine Erweiterung in die komplexe Zahlenmenge biologisch schwer interpretierbar ist, sind diese negativen Schätzwerte unbrauchbar. Als Ausweg hat man andere Schätzverfahren für Varianzkomponenten entwickelt, bei denen das Problem von negativen Schätzwerten nicht auftritt. Zwei von diesen Schätzverfahren wollen wir im folgenden noch etwas genauer anschauen.

10.5 Likelihood basierte Verfahren

Das **Maximum Likelihood** (ML) Verfahren wurde anfangs des 20. Jahrhunderts von R.A. Fisher entwickelt. ML ist ein allgemeines Schätzverfahren um unbekannte Parameter aus Daten zu schätzen. Es wird also nicht nur für die Schätzung von Varianzkomponenten verwendet. Nehmen wir an, dass es sich bei den beobachteten Daten um kontinuierliche Grössen handelt. Das heisst, die beobachteten Werte sind im wesentlichen reelle Zahlen. Bei ML geht man davon aus, dass die beobachteten Daten einer bestimmten Dichteverteilung - zum Beispiel einer multivariaten Normalverteilung - folgen. Diese Dichteverteilung ist abhängig von unbekannten Parametern, welche aus den Daten geschätzt werden sollen. Sobald wir es mit diskreten Daten zu tun haben, dann können diese nur gewisse Werte annehmen und anstelle der Dichteverteilung der kontinuierlichen Daten, folgen die diskreten Daten einer Wahrscheinlichkeitsverteilung. In den folgenden Abschnitten nehmen wir für die Erklärung von ML kontinuierliche Daten an. Das Verfahren funktioniert aber auch für diskrete Daten.

10.5.1 Dichteverteilung von Beobachtungen

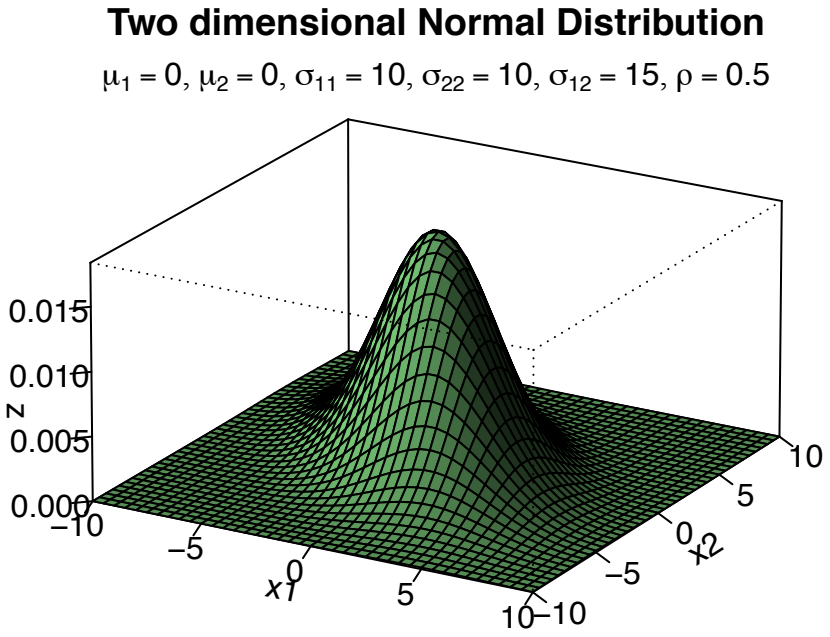
Wir haben einem Vektor y mit n Beobachtungswerten. Wir nehmen an, dass diese Daten einer bestimmten Dichteverteilung folgen. Als Beispiel für eine Verteilung können wir uns die multivariate oder multidimensionale Normalverteilung vorstellen. Der Begriff **multivariat** bedeutet, dass die Normalverteilung sich über mehrere Dimensionen ausdehnt. Bei n Beobachtungen im Datensatz dehnt sich die gewählte Normalverteilung für y über exakt n Dimensionen aus. Allgemein ist eine reelle n-dimensionale Zufallsvariable Y normalverteilt, wenn sie eine Dichteverteilung

$$f_Y(y) = \frac{1}{\sqrt{(2\pi)^n \ det(\Sigma)}} \ exp\left(-\frac{1}{2}(y-\mu)^T\Sigma^{-1}(y-\mu)\right)$$

mit	μ	Erwartungsvektor der Länge n
	Σ	Covarianzmatrix mit Dimension $n \times n$
	$det()$	Determinante

besitzt. Abgekürzt schreibt man auch $Y \sim \mathcal{N}_n(\mu, \Sigma)$.

Eine graphische Darstellung für eine zweidimensionale Normalverteilung, das heisst hier wäre $n = 2$, ist im nachfolgenden Plot gezeigt.



Die für y gewählte Dichteverteilung ist in der Regel von unbekannten Parametern abhängig. Bei der eindimensionalen Normalverteilung sind das der Erwartungswert μ und die Varianz σ^2. Wir definieren den Parametervektor θ als einen Vektor, der alle unbekannten Parameter einer gewissen Verteilung enthält. Bei der eindimensionalen Normalverteilung ist

$$\theta = \begin{bmatrix} \mu \\ \sigma^2 \end{bmatrix}$$

10.5.2 Likelihood Funktion

Die gewählte Dichteverteilung für die Beobachtungen y bestimmt die Dichte $f(y|\theta)$ der Daten gegeben die unbekannten Parameter. Diese Funktion $f(y|\theta)$ liefert bei bekannten Werten von θ die Dichtewerte in Abhängigkeit der Beobachtungen y. Bevor die Daten beobachtet werden, kann $f(y|\theta)$ als Funktion der unbekannten Daten behandelt werden und liefert `a priori` Information zur Dichte von möglichen Daten bei gegebenen Verteilungsparametern θ. Sobald aber die Daten beobachtet sind, dann sind diese fix und können nicht mehr verändert werden. Dann macht es keinen Sinn mehr $f(y|\theta)$ als Funktion von y anzuschauen. Da aber die Parameter unbekannt sind, liegt es auf der Hand $f(y|\theta)$ als Funktion der unbekannten Parameter θ zu betrachten. Wir definieren also die Funktion $L(\theta)$ als

$$L(\theta) = f(y|\theta) \tag{10.14}$$

Die Funktion $L(\theta)$ heisst **Likelihood**. Aufgrund der Definition von $L(\theta)$ können wir sagen, dass je besser ein Dichteverteilung mit gegebenem Parametervektor θ die beobachteten Daten y beschreibt desto grösser ist der entsprechende Likelihood-Wert. Aufgrund dieses Arguments scheint es vernünftig, die unbekannten Parameter θ so zu wählen, dass $L(\theta)$ maximal wird. Genau das wird im ML-Schätzverfahren umgesetzt. Wir definieren für eine gewählte Dichteverteilung der Beobachtung die Likelihoodfunktion. Dann maximieren wir $L(\theta)$ im Bezug auf θ und wählen den Wert für θ als Schätzer, welcher $L(\theta)$ maximiert. Formal schreiben wir das als

$$\hat{\theta}_{ML} = argmax_\theta \; L(\theta)$$

10.5.3 Beispiel für ein Regressionsmodell

Als erstes Beispiel schauen wir uns an, wie wir die Restvarianz σ^2 in einem Regressionsmodell (10.15) mit dem ML-Verfahren schätzen können.

$$y = Xb + e \tag{10.15}$$

Unter der Annahme, dass die Beobachtungen y einer multivariaten Normalverteilung folgen, können wir die bedingte Dichteverteilung aller Daten gegeben bekannte Parameter schreiben als

$$f_Y(y|b, \sigma^2) = (2\pi\sigma^2)^{-n/2} exp\left(-\frac{1}{2\sigma^2}(y - Xb)^T\ (y - Xb)\right) \tag{10.16}$$

Fassen wir die Dichteverteilung in (10.16) als Funktion der Parameter b und σ^2 auf, so resultiert daraus die folgende Likelihoodfunktion

$$L(b, \sigma^2) = (2\pi\sigma^2)^{-n/2} exp\left(-\frac{1}{2\sigma^2}(y - Xb)^T\ (y - Xb)\right) \tag{10.17}$$

Schätzwerte für die unbekannten Parameter b und σ^2 erhalten wir indem wir $L(b, \sigma^2)$ mit Bezug auf b und auf σ^2 maximieren. Allgemein finden wir das Maximum einer Funktion durch differenzieren und Nullsetzen der ersten Ableitung (Steigung). Die so erhaltenen Nullstellen der Steigung müssen mit höheren Ableitungen überprüft werden, ob sie tatsächlich ein Maximum darstellen. Für das Differenzieren verwenden wir nicht die Likelihoodfunktion $L(b, \sigma^2)$ direkt, sondern deren Logarithmus zur Basis e.

$$l(b, \sigma^2) = \log(L(b, \sigma^2)) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}(y - Xb)^T\ (y - Xb)$$

Der Grund für diese Transformation ist, dass $l(b, \sigma^2)$ in der Regel viel einfacher zu differenzieren ist als $L(b, \sigma^2)$. Die Transformation auf die logarithmische Skala ändert nichts an der Position der auftretenden Extrema. Für uns heisst das, dass wo immer $l(b, \sigma^2)$ ein Maximum hat, hat auch $L(b, \sigma^2)$ ein Maximum.

Obwohl wir eigentlich nur an der Schätzung für die Varianzkomponente σ^2 interessiert sind, bekommen wir mit dem ML-Verfahren auch eine Schätzung für den Vektor b. Wir berechnen die partielle Ableitung von $l(b, \sigma^2)$ nach b und nach σ^2, setzen diese gleich Null und haben dann Kandidaten für mögliche Schätzwerte.

$$\begin{aligned} \frac{\partial l(b, \sigma^2)}{\partial b} &= -\frac{1}{2\sigma^2}(-(y^T X)^T - X^T y + 2X^T Xb) \\ &= -\frac{1}{2\sigma^2}(-2X^T y + 2X^T Xb) \end{aligned} \tag{10.18}$$

Das Maximum von $l(b, \sigma^2)$ kann dort auftreten, wo die Ableitung in (10.18) gleich 0 ist. Die Untersuchung höherer Ableitung würde ergeben, dass diese Nullstelle der Ableitung wirklich ein Maximum darstellt. Somit folgen die sogenannten Normalgleichungen

$$X^T y = X^T X\hat{b}$$

Daraus folgt die ML-Schätzung für b als

$$\hat{b} = (X^T X)^{-1} X^T y \tag{10.19}$$

Der ML-Schätzer für b in (10.19) setzt voraus, dass die Matrix X vollen Kolonnenrang p hat. Das heisst, keine zwei oder mehr Kolonnen von X sind linear abhängig voneinander. Der ML-Schätzer $\hat{b}$ für b entspricht dem Schätzer, welcher wir schon mit Least Squares gefunden hatten.

Den ML-Schätzer für σ^2 finden wir analog zum Schätzer für b. Als erstes berechnen wir die partielle Ableitung von $l(b, \sigma^2)$ nach σ^2. Dann setzen wir diese gleich 0 und erhalten so den ML-Schätzer für σ^2.

$$\frac{\partial l(b,\sigma^2)}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}(y - Xb)^T\ (y - Xb) \tag{10.20}$$

Vorausgesetzt, dass $\sigma^2 \neq 0$, gilt

$$\frac{1}{\hat{\sigma}^2}(y - Xb)^T\ (y - Xb) - n = 0$$

Dieser Ausdruck kann nur dann gleich 0 sein, falls

$$\hat{\sigma}^2 = \frac{1}{n}(y - Xb)^T\ (y - Xb) \tag{10.21}$$

Schreiben wir den Ausdruck in (10.21) in der Summennotation, so erhalten wir

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - x_i^T b)^2 \tag{10.22}$$

Da in (10.22) der Vektor b unbekannt ist, setzen wir den Schätzer $\hat{b}$ aus (10.19) ein und erhalten so den ML-Schätzer für σ^2

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - x_i^T \hat{b})^2 \tag{10.23}$$

Vergleichen wir den ML-Schätzer aus (10.23) mit dem Schätzer, den wir bei Least Squares aufgrund der Residuen gefunden hatten, dann sind die beiden Schätzer nicht gleich. Der Schätzer für σ^2 aufgrund der Residuen ist definiert als

$$\hat{\sigma}^2_{Res} = \frac{1}{n-p}\sum_{i=1}^{n} r_i^2 \tag{10.24}$$

wobei $r_i^2 = y_i - x_i^T \hat{b}$ und p dem Kolonnenrang der Matrix X entspricht. Wir hatten auch gesehen, dass $\hat{\sigma}^2_{Res}$ erwartungstreu ist. Somit ist der ML-Schätzer für σ^2 nicht erwartungstreu, d.h. $E\left[\hat{\sigma}^2_{ML}\right] \neq \sigma^2$.

10.5.4 Beispiel für das gemischte lineare Modell

Im allgemeinen lineare gemischten Modell

$$y = Xb + Zu + e \tag{10.25}$$

gibt es mindestens zwei Varianzkomponenten, welche zu schätzen sind. Für die beiden zufälligen Effekte u und e haben wir angenommen, dass

$$var(e) = R = I * \sigma_e^2$$

und

$$var(u) = G.$$

Je nach Anwendung hat auch G eine einfache Struktur, d.h. wir können G zerlegen in eine bekannte Matrix A mal eine Varianzkomponente σ_u^2. Als Beispiel entspricht G im Tiermodell der Verwandtschaftsmatrix mal die additiv genetische Varianz, d.h. $G = A * \sigma_a^2$. Die Erwartungswerte der zufälligen Effekte u und e sind für beide gleich. Es gilt also

$$E[e] = 0 \text{ und } E[u] = 0$$

Aus diese Eigenschaften für die Erwartungswerte und die Varianzen folgt, dass für die Beobachtungen y gilt

$$E[y] = Xb \text{ und } var(y) = V$$

10.5.5 Likelihood für das gemischte Modell

Unter der Annahme, dass die Beobachtungen y einer multivariaten Normalverteilung folgen, d.h.

$$y \sim \mathcal{N}(Xb, V)$$

dann ist die entsprechende Likelihoodfunktion $L(b, V)$ definiert als

$$L(b,V) = (2\pi)^{n/2}\ det(V)^{1/2}\ exp\left\{-\frac{1}{2}(y-Xb)^T V^{-1}(y-Xb)\right\}$$

Auch hier transformieren wir die Funktion L wieder auf die logarithmische Skala und erhalten

$$l(b,V) = \log(L(b,V)) = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\log(det(V)) - \frac{1}{2}(y-Xb)^T V^{-1}(y-Xb)$$

Den ML-Schätzer für b erhalten wir durch Nullsetzen der partiellen Ableitung von $l(b, V)$ nach b. Als Resultat erhalten wir den bekannten verallgemeinerten Least Squares Schätzer

$$\hat{b} = \left(X^T V^{-1} X\right) X^T V^{-1} y \tag{10.26}$$

Die partielle Ableitung von $l(b, V)$ nach σ^2 entspricht

$$\frac{\partial l(b,V)}{\partial \sigma^2} = -\frac{1}{2} tr(V^{-1}\tilde{Z}\tilde{Z}^T) + \frac{1}{2}(y-Xb)^T V^{-1}\tilde{Z}\tilde{Z}^T V^{-1}(y-Xb) \tag{10.27}$$

wobei $tr()$ die Spur (Summe der Diagonalelemente) einer Matrix bezeichnet. Die Varianzkomponente σ^2 entspricht der Kombination von σ_e^2 und σ_u^2 und $\tilde{Z}$ entspricht der Inzidenzmatrix aus der kombinierten Varianzkomponente.

Setzt man die Ableitung in (10.27) gleich Null und setzt für b den Schätzer $\hat{b}$ aus (10.26), dann resultiert ein Gleichungssystem, dessen Lösung zum ML-Schätzer von σ^2 führt.

10.6 Restricted (Residual) Maximum Likelihood (REML)

ML-Schätzer von Varianzkomponenten haben die Eigenschaft, dass sie die Anzahl Freiheitsgrade (p), welche zur Schätzung der fixen Effekte verwendet werden, nicht berücksichtigen. Sie sind somit nicht erwartungstreu, was wir beim ML-Schätzer der Restvarianz für das einfache Regressionsmodell gesehen hatten.

Im Gegensatz zu ML berücksichtigen REML-Schätzungen von Varianzkomponenten die Anzahl Freiheitsgrade, welche für die Schätzung von fixen Effekten verwendet werden. Dies wird dadurch erreicht, dass die Likelihoodfunktion nicht als die Dichteverteilung von y gegeben die Parameter aufgestellt werden, sondern von einer Transformation $\tilde{y} = Ky$. Dabei wird die Matrix K so bestimmt, dass der Erwatungswert $E[\tilde{y}] = 0$ ist. Somit treten in der Likelihoodfunktion über $\tilde{y}$ keine fixen Effekte b mehr auf, welche auch noch geschätzt werden müssen.

Der eigentliche Prozess, wie man zu den Schätzungen kommt ist analog zum Maximum-Likelihood Verfahren, nur wird anstelle von y mit $\tilde{y}$ operiert.

10.7 Bayes'sche Ansätze (Ein Ausblick)

In der Statistik gibt es zwei fundamentale Philosophien, wie Datenanalysen gemacht werden sollen. Auf der einen Seite gibt es den **frequentistischen** Ansatz und auf der anderen Seite den **Bayes'schen** Ansatz. Alles was wir bis jetzt behandelt haben stammt aus der frequentistischen Welt.

Bayes'sche Ansätze sind so benannt, weil sie auf dem Satz von Bayes begründet sind. Dieser Satz enthält eigentlich nur die Definition der bedingten Wahrscheinlichkeit. Eine Bayes'sche Schätzung eines unbekannten Parameters θ aufgrund von Daten y, basiert immer auf der sogenannten **a posteriori** Verteilung ($P(\theta|\ y)$) des Parameters gegeben die Daten. Als eigentlicher Schätzwert wird dann meistens der Erwartungswert $E[\theta|y]$ verwendet.

Die a posteriori Verteilung des Parameters gegeben die Daten lässt sich gemäss Satz von Bayes berechnen als

$$P(\theta|y) = \frac{P(y|\theta) * P(\theta)}{P(y)}$$

wobei $P(y|\theta)$ der Likelihood entspricht und $P(\theta)$ als **a priori** Wahrscheinlichkeit des unbekannten Parameters bezeichnet wird. $P(y)$ steht für eine Normalisierungskonstante, welche keine weitere Bedeutung hat.