Züchtungslehre - BLUP von Zuchtwerten

Peter von Rohr

2016-11-11

Einführung

Die in diesem Kapitel eingeführten und besprochenen Verfahren werden in der **traditionellen Zuchtwertschätzung** verwendet. Unter traditionellen Zuchtwertschätzungen verstehen wir die Auswertungen, bei welchen Zuchtwerte aufgrund von phänotypischen Leistungen und Pedigreeinformationen geschätzt werden. Im Gegensatz dazu steht die **genomische Selektion** bei welcher Zuchtwerte aufgrund von genomischen Informationen ermittelt werden.

Die Schätzung von Zuchtwerten mit der Regressionsmethode oder mit dem Selektionsindex verlangt, dass wir die phänotypischen Leistungen um bekannte Umwelteinflüsse korrigieren können. Diese Korrektur wurde jeweilen im Populationsmittel zusammengefasst. Im routinemässigen Betrieb, wo Zuchtwerte aufgrund von Daten aus dem Feld geschätzt werden sollen, sind die Umwelteinflüsse sehr verschieden und a priori nicht bekannt. Somit brauchen wir ein Verfahren, mit welchem wir gleichzeitig den Einfluss von verschiedenen Umweltfaktoren schätzen können und Zuchtwerte voraussagen können.

BLUP-Verfahren

Das BLUP Verfahren ist für die traditionelle Zuchtwertschätzung die Methode der Wahl. BLUP wurde ab 1949 unter der Leitung von Charles Henderson entwickelt. Im Gegensatz zu anderen Methoden, wie dem Selektionsindex, erlaubt BLUP die simultane Schätzung von Umwelteffekten zusammen mit der Vorhersage der Zuchtwerte.

Die Abkürzung BLUP steht für **Best Linear Unbiased Prediction** und beschreibt damit die Eigenschaften der geschätzten zufälligen Effekte in statistischen Modell. Was diese Eigenschaften genau bedeuten werden wir später noch genauer betrachten. Der Ausgangspunkt von BLUP ist ein sogenanntens **lineares gemischtes Modell** (linear mixed model). Ein Beispiel für ein solches lineares gemischtes Modell ist in Gleichung (1) gezeigt. Ein statistisches Modell, welches neben den fixen Effekten b und den zufälligen Resteffekten b zusätzlich noch weitere zufällige Effekte u aufweist, wird als ein gemischtes Modell bezeichnet.

Modell

$$y = Xb + Zu + e \tag{1}$$

mit
- y: Vektor der Beobachtungswerte
- b: Vektor der fixen Effekte
- u: Vektor der zufälligen Effekte
- e: Vektor der zufälligen Resteffekte
- X: Inzidenzmatrix zur Verknüpfung der Beobachtungen mit den fixen Effekten
- Z: Inzidenzmatrix zur Verknüpfung der Beobachtungen mit den zufälligen Effekten

Erwartungswerte und Varianzen

Die Erwartungswerte und Varianzen der Zufallsvariablen im Modell (1) sind definiert als

$$E \begin{bmatrix} y \\ u \\ e \end{bmatrix} = \begin{bmatrix} Xb \\ 0 \\ 0 \end{bmatrix} \tag{2}$$

Aufgrund der Definition der Erwartungswerte in (2) können wir erkennen, dass die zufälligen Effekte u und e jeweilen einen Erwartungswert von 0 haben und somit als Abweichung von einem allgemeinen Populationsmittel definiert sind. Dieses allgemeine Mittel entspricht dem Erwartungswert der phänotypischen Beobachtungen y und ist gleich Xb.

$$var \begin{bmatrix} y \\ u \\ e \end{bmatrix} = \begin{bmatrix} ZGZ^T + R & ZG & R \\ GZ^T & G & 0 \\ R & 0 & R \end{bmatrix} \tag{3}$$

Die Covarianzmatrix der zufälligen Resteffekte e wird mit R bezeichnet und die der zufälligen Effekte u wird mit G bezeichnet. Die Covarianz $cov(u, e^T)$ zwischen den zufälligen Effekte u und e wird als 0 angenommen.

Die fixen Effekte b werden als fixe Faktorstufen bezeichnet und haben somit keine Varianz, d.h. dass $var(b) = 0$. Auch alle Covarianzen zwischen den zufälligen Effekten u und e und den fixen Effekten b sind 0.

Aus den bisher getroffenen Annahmen betreffend der Varianzen und Covarianzen können die anderen Elemente der Matrix in (3) berechnet werden. Die Varianz $var(y)$ der phänotypischen Beobachtungen y wird mit V bezeichnet und berechnet sich als

$$\begin{aligned} var(y) &= var(Xb + Zu + e) \\ &= var(Zu) + var(e) \\ &= Z * var(u) * Z^T + var(e) \\ &= Z * G * Z^T + R = V \end{aligned}$$

Die Covarianz $cov(y, u^T)$ zwischen den phänotypischen Beobachtungen y und den zufälligen Effekten u ergibt sich als

$$\begin{aligned} cov(y, u^T) &= cov(Xb + Zu + e, u^T) \\ &= cov(Zu, u^T) + cov(e, u^T) \\ &= Z * cov(u, u^T) \\ &= Z * G \end{aligned}$$

Eigenschaften der Schätzwerte

Wie schon erwähnt beschreibt die Abkürzung BLUP die Eigenschaften der Schätzwerte $\hat{u}$ der zufälligen Effekte u. Die Bedeutung dieser Eigenschaften wollen wir jetzt genauer analysieren.

- **Linear**: fixe Effekte Kb plus zufällige Effekte Mu werden mit einer Linearkombination By der Beobachtungswerte geschätzt.
- **Unbiased**: die Schätzungen für die fixen Effekte plus die zufälligen Effekte sollen unverzerrt (unbiased) oder erwartungstreu sein. Somit ist $E[Kb + Mu] = E[By]$. Daraus folgt aufgrund der Definitionen zum Modell (1), dass $Kb = BXb$

- **Best**: unter allen linearen unverzerrten Schätzern weist der BLUP-Schäter die kleinste Fehlervarianz auf. Das heisst $var(By - (Kb + Mu))$ ist minimal

Aus den BLUP-Eigenschaften lassen sich die folgenden Schätzgleichungen ableiten.

$$\hat{b} = (X^T V^{-1} X)^- X^T V^{-1} y \tag{4}$$

$$\hat{u} = G Z^T V^{-1} (y - X\hat{b}) \tag{5}$$

mit $\hat{b}$ Lösungsvektor der fixen Effekte
$\hat{u}$ Lösungsvektor der zufälligen Effekte
$()^-$ verallgemeinerte Inverse

Zur Lösung der Gleichungen (4) und (5) wird die Inverse V^{-1} der Covarianzmatrix aller phänotypischen Beobachtungen gebraucht. Da diese Matrix sehr gross ist, ist deren Berechnung schon bei kleinen Datenmengen praktisch nicht mehr durchführbar. Charles Henderson hat aber gezeigt, dass die folgenden sogenannten **Mixed Model Equations** zu den gleichen Lösungen führt, wie die Gleichungen (4) und (5).

$$\begin{bmatrix} X^T R^{-1} X & X^T R^{-1} Z \\ Z^T R^{-1} X & Z^T R^{-1} Z + G^{-1} \end{bmatrix} \begin{bmatrix} \hat{b} \\ \hat{u} \end{bmatrix} = \begin{bmatrix} X^T R^{-1} y \\ Z^T R^{-1} y \end{bmatrix} \tag{6}$$

Da die Matrix R im Gegensatz zu V diagonal oder zumindest blockdiagonal ist, hat sie eine einfachere Struktur als V und ist somit einfacher zu invertieren. Deshalb ist das Lösen des Gleichungssystems in (6) einfacher als das Lösen von (4) und (5).

Sowohl bei den mixed model equations als auch bei den Gleichungen (4) und (5) werden bekannte Covarianzmatrizen und somit bekannte Varianzkomponenten für die zufälligen Effekte u und e vorausgesetzt. In der Praxis sind diese nicht bekannt und müssen aus Daten geschätzt werden. Dazu folgen mehr Informationen im Kapitel **Varianzkomponentenschätzung**.

Das Tiermodell

Das gemischte Modell, welches als zufällige Effekte die Zuchtwerte a der Tiere enthält, wird als **Tiermodell** ("animal model") bezeichnet. Das Tiermodell ist eine Weiterentwicklung des Vatermodells ("sire model"), bei welchem die Vatereffekte s als zufällige Effekte modelliert werden.

Das einfache Tiermodell wird mit der folgenden Beschreibung charaterisiert.

$$y = Xb + Za + e \tag{7}$$

mit y Vektor der phänotypischen Beobachtungen
b Vektor der fixen Effekte
a Vektor der zufälligen Effekte, welche den Zuchtwerten entsprechen
e Vektor der zufälligen Resteffekte
X Inzidenzmatrix der fixen Effekte
Z Inzidenzmatrix der Zuchtwerte

Die Erwartungswerte sind analog wie beim Modell (1) definiert, wenn der Vektor der zufälligen Effekte u durch den Vektor der Zuchtwerte a ersetzt wird. Bei der Betrachtung eines Merkmals sind die Covarianzmatrizen der zufälligen Effekte a und e gegeben als

$$Var(a) = G = A\ \sigma_a^2 \text{ und } G^{-1} = A^{-1}\sigma_a^{-2} \tag{8}$$

$$Var(e) = R = I\ \sigma_e^2 \text{ und } R^{-1} = I\ \sigma_e^{-2} \tag{9}$$

wobei A additiv genetische Verwandtschaftsmatrix
I Einheitsmatrix
σ_a^2 additiv genetische Varianz
σ_e^2 Restvarianz

Stellen wir nun die Mischmodellgleichungen gemäss (6) für das Tiermodell auf, so folgt

$$\begin{bmatrix} X^TR^{-1}X & X^TR^{-1}Z \\ Z^TR^{-1}X & Z^TR^{-1}Z + G^{-1} \end{bmatrix} \begin{bmatrix} \hat{b} \\ \hat{a} \end{bmatrix} = \begin{bmatrix} X^TR^{-1}y \\ Z^TR^{-1}y \end{bmatrix} \tag{10}$$

Setzen wir die Beziehungen aus (8) und (9) in die Mischmodellgleichungen des Tiermodells (10) ein und multiplizieren beide Seiten der Gleichung mit σ_e^2, so folgt

$$\begin{bmatrix} X^TX & X^TZ \\ Z^TX & Z^TZ + A^{-1}\ \alpha \end{bmatrix} \begin{bmatrix} \hat{b} \\ \hat{a} \end{bmatrix} = \begin{bmatrix} X^Ty \\ Z^Ty \end{bmatrix} \tag{11}$$

wobei α Verhältnis der Varianzen $\sigma_e^2/\sigma_a^2 = (1-h^2)/h^2$
h^2 Heritabilität

Das Gleichungssystem (11) kann kürzer geschrieben werden als

$$M * \hat{s} = r$$

wobei M Koeffizientenmatrix heisst
$\hat{s}$ Lösungsvektor
r Vektor der rechten Handseite

In (11) werden phänotypischen Leistungen verwandter Tiere über die Verwandtschaftsmatrix A miteinander verknüpft. Dadurch bekommen auch Tiere ohne phänotypischen Leistung im Beobachtungsvektor y einen geschätzten Zuchtwert. Hier wird klar, dass BLUP uns erlaubt, dass die Anzahl der zu schätzenden oder vorauszusagender Parameter grösser sein kann, als die Anzahl Beobachtungen. Dies wäre unter einem einfachen Regressionsmodell nicht möglich. In der Statistik wird dieser Vorgang auch als **Regularisierung** bezeichnet. Die Regularisierung ist auch bei der genomischen Selektion ein Thema. Dort werden wir noch andere Regularisierungsverfahren kennen lernen.

Ein Beispiel für das Tiermodell

Gegeben sei der folgende Datensatz für das Merkmal Gewichtszuwachs (WWG in kg) vor dem Absetzen bei Kälber an. Das Ziel ist, dass wir für alle Tiere Zuchtwerte für das Merkmal (WWG) schätzen. Die Varianzkomponenten σ_e^2 und σ_a^2 sind bekannt und haben die Werte $\sigma_e^2 = 40$ und $\sigma_a^2 = 20$. Somit ist das Varianzverhältnis $\alpha = 40/20 = 2$

Kalb	Geschlecht	Vater	Mutter	WWG
4	M	1	NA	4.5
5	F	3	2	2.9
6	F	1	2	3.9
7	M	4	5	3.5
8	M	3	6	5.0

Das Modell zur Beschreibung einer Beobachtung y_{ij} lautet

$$y_{ij} = b_i + a_j + e_{ij}$$

wobei	y_{ij}	beobachteter Wert für Merkmal WWG für Kalb j mit Geschlecht i
	b_i	fixer Effekt für Geschlecht i
	a_j	Zuchtwert für Kalb j
	e_{ij}	Resteffekt für Kalb j mit Geschlecht i

Setzen wir die beobachteten Werte aus der Datentabelle ein, dann folgt

$$\begin{aligned} 4.5 &= b_M + a_4 + e_{M4} \\ 2.9 &= b_F + a_5 + e_{F5} \\ 3.9 &= b_F + a_6 + e_{F6} \\ 3.5 &= b_M + a_7 + e_{M7} \\ 5.0 &= b_M + a_8 + e_{M8} \end{aligned}$$

In Matrix-Vektor-Schreibweise haben wir dann das bekannte Gleichungssystem des Tiermodells

$$y = Xb + Za + e \tag{12}$$

Die Vektoren und Matrizen in (12) sehen wie folgt aus

$$y = \begin{bmatrix} 4.50 \\ 2.90 \\ 3.90 \\ 3.50 \\ 5.00 \end{bmatrix}$$

$$b = \begin{bmatrix} \hat{b}_M \\ \hat{b}_F \end{bmatrix}$$

$$a = \begin{bmatrix} \hat{a}_1 \\ \hat{a}_2 \\ \hat{a}_3 \\ \hat{a}_4 \\ \hat{a}_5 \\ \hat{a}_6 \\ \hat{a}_7 \\ \hat{a}_8 \end{bmatrix}$$

$$e = \begin{bmatrix} e_{M4} \\ e_{F5} \\ e_{F6} \\ e_{M7} \\ e_{M8} \end{bmatrix}$$

Die Inzidenzmatrizen X und Z verknüpfen die Beobachtungen mit den jeweiligen Effekten.

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

$$Z = \begin{bmatrix} 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

Aufstellen der Mischmodellgleichungen

Nun haben wir alle Elemente, die es braucht um die Mischmodellgleichungen (11) aufzustellen. Wir berechnen als erstes die Elemente der Koeffizientenmatrix M.

$$X^T X = \begin{bmatrix} 3 & 0 \\ 0 & 2 \end{bmatrix}$$

$$X^T Z = \begin{bmatrix} 0 & 0 & 0 & 1 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 \end{bmatrix}$$

$$Z^T Z = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

Aufgrund der Verwandtschaftsbeziehungen, welche in den Daten gegeben ist, können wir das Pedigree aufbauen.

```
##   sire  dam
## 1 <NA> <NA>
## 2 <NA> <NA>
## 3 <NA> <NA>
## 4    1 <NA>
## 5    3    2
## 6    1    2
## 7    4    5
## 8    3    6
```

Die Inverse Verwandtschaftsmatrix können wir aus dem Pedigree errechnen

$$A^{-1} = \begin{bmatrix} 1.833 & 0.500 & 0.000 & -0.667 & 0.000 & -1.000 & 0.000 & 0.000 \\ 0.500 & 2.000 & 0.500 & 0.000 & -1.000 & -1.000 & 0.000 & 0.000 \\ 0.000 & 0.500 & 2.000 & 0.000 & -1.000 & 0.500 & 0.000 & -1.000 \\ -0.667 & 0.000 & 0.000 & 1.833 & 0.500 & 0.000 & -1.000 & 0.000 \\ 0.000 & -1.000 & -1.000 & 0.500 & 2.500 & 0.000 & -1.000 & 0.000 \\ -1.000 & -1.000 & 0.500 & 0.000 & 0.000 & 2.500 & 0.000 & -1.000 \\ 0.000 & 0.000 & 0.000 & -1.000 & -1.000 & 0.000 & 2.000 & 0.000 \\ 0.000 & 0.000 & -1.000 & 0.000 & 0.000 & -1.000 & 0.000 & 2.000 \end{bmatrix}$$

Das letzte Element der Koeffizientenmatrix $Z^T Z + A^{-1} * \alpha$ ist somit

$$\begin{bmatrix} 3.667 & 1.000 & 0.000 & -1.333 & 0.000 & -2.000 & 0.000 & 0.000 \\ 1.000 & 4.000 & 1.000 & 0.000 & -2.000 & -2.000 & 0.000 & 0.000 \\ 0.000 & 1.000 & 4.000 & 0.000 & -2.000 & 1.000 & 0.000 & -2.000 \\ -1.333 & 0.000 & 0.000 & 4.667 & 1.000 & 0.000 & -2.000 & 0.000 \\ 0.000 & -2.000 & -2.000 & 1.000 & 6.000 & 0.000 & -2.000 & 0.000 \\ -2.000 & -2.000 & 1.000 & 0.000 & 0.000 & 6.000 & 0.000 & -2.000 \\ 0.000 & 0.000 & 0.000 & -2.000 & -2.000 & 0.000 & 5.000 & 0.000 \\ 0.000 & 0.000 & -2.000 & 0.000 & 0.000 & -2.000 & 0.000 & 5.000 \end{bmatrix}$$

Zum Aufstellen der rechten Handseite brauchen wir die folgenden beiden Vektoren

$$X^T y = \begin{bmatrix} 13.0 \\ 6.8 \end{bmatrix}$$

$$Z^T y = \begin{bmatrix} 0.0 \\ 0.0 \\ 0.0 \\ 4.5 \\ 2.9 \\ 3.9 \\ 3.5 \\ 5.0 \end{bmatrix}$$

Lösen wir nun das Gleichungssystem nach dem Lösungsvektor auf, dann erhalten wir Schätzwerte für die fixen Effekte b und die Zuchtwerte a. Für unser einfaches Beispiel können wir die Lösung über eine explizite Inversion der Koeffizientenmatrix M bekommen. Wir haben also

$$\hat{s} = M^{-1} * r$$

Die Zahlenwerte für den Lösungsvektor bekommen wir

$$\hat{s} = \begin{bmatrix} 4.359 \\ 3.404 \\ 0.098 \\ -0.019 \\ -0.041 \\ -0.009 \\ -0.186 \\ 0.177 \\ -0.249 \\ 0.183 \end{bmatrix}$$