

# ASMNW - Lösung 1

*Peter von Rohr*

*2017-02-24*

## Kontrollfrage 1

Welches Modell wurde vor der Genomischen Selektion zur Zuchtwertschätzung verwendet und welche Tiere bekamen in diesem Modell Zuchtwerte?

### Lösung

- BLUP Tiermodell
- alle Tiere bekommen einen Zuchtwert

## Kontrollfrage 2

Beim gängigen Verfahren zur genomischen Zuchtwertschätzung braucht es mehrere Schritte, wie sehen diese aus?

### Lösung

- Schritt 1: Schätzung der  $a$ -Werte in der Referenzpopulation
- Schritt 2: Schätzung genomischer Zuchtwerte für Tiere ausserhalb der Referenzpopulation durch Aufsummieren der für die Tiere relevanten  $a$ -Effekte

## Kontrollfrage 3

Was bedeuten die  $a$ -Werte in den Modellen der genomischen Zuchtwertschätzung und welchem genetischen Modell werden diese entnommen?

### Lösung

- Die  $a$ -Werte sind Allel-Substitutionseffekte
- Sie stammen aus dem Substitutionsmodell

## Kontrollfrage 4

Im Paper zur Deregression (auf dem Stick oder unter: <http://gsejournal.biomedcentral.com/articles/10.1186/1297-9686-41-55>) stehen nach Gleichung (8) zwei Probleme, weshalb mit BLUP geschätzte Zuchtwerte nicht ideal sind als Beobachtungen in genomischer Zuchtwertschätzung. Fassen Sie diese zwei Probleme mit Ihren Worten kurz zusammen.

## Lösung

1. Durch das Hinzufügen des Schätzfehlers wird die Varianz des Schätzers im Vergleich zur Varianz der phänotypischen Beobachtungen reduziert. Das würde zu einer Unterschätzung der Allel-Substitutionseffekte führen
2. Durch die Eigenschaften von BLUP werden die geschätzten Zuchtwerte zum Durchschnitt der Eltern gedrückt (shrinkage estimation). Das Ausmass, wie stark die einzelnen Schätzwerte zu den Elterndurchschnitten gedrückt werden, hängt vom Informationsgehalt, d.h. vom Bestimmungsmass ab. Dies verfälscht aber die Schätzwerte von Allelsubstitutionseffekten.

## Aufgabe 1: Modellierung

In einem kleinen Beispieldatensatz sind die SNP-Genotypen für 5 Tiere gegeben. Für jedes Tier liegen Typisierungsergebnisse an 10 SNPs vor. Die Bezeichnung  $(G_k G_l)_{ij}$  steht für den Genotypen für Tier  $i$  an der SNP-Position  $j$  mit den Allelen  $k$  und  $l$ . Da wir nur SNPs mit zwei Allelen betrachten können als  $k$  und  $l$  nur entweder 0 oder 1 sein. Wir nehmen an, dass das Allel 0 immer das Allel mit der gewünschten Ausprägung ist. Im Substitutionseffekt ignorieren wir alle Dominanzeffekte, d.h. alle  $d$ -Werte werden auf 0 gesetzt. In der folgenden Tabelle sind die SNP-Genotypen für alle Tiere aufgelistet, wobei die Indices  $i$  und  $j$  weggelassen wurden.

	Tier 1	Tier 2	Tier 3	Tier 4	Tier 5
SNP1	$G_0 G_0$	$G_0 G_1$	$G_0 G_0$	$G_0 G_1$	$G_0 G_0$
SNP2	$G_0 G_0$	$G_0 G_1$	$G_0 G_0$	$G_0 G_0$	$G_0 G_1$
SNP3	$G_0 G_0$	$G_0 G_1$	$G_0 G_1$	$G_0 G_1$	$G_0 G_1$
SNP4	$G_0 G_0$	$G_0 G_0$	$G_0 G_0$	$G_0 G_1$	$G_0 G_1$
SNP5	$G_0 G_1$	$G_0 G_0$	$G_0 G_0$	$G_0 G_1$	$G_0 G_0$
SNP6	$G_0 G_1$	$G_0 G_1$	$G_0 G_1$	$G_0 G_0$	$G_0 G_0$
SNP7	$G_1 G_1$	$G_0 G_1$	$G_0 G_1$	$G_0 G_1$	$G_0 G_1$
SNP8	$G_0 G_1$	$G_0 G_0$	$G_0 G_1$	$G_0 G_0$	$G_0 G_1$
SNP9	$G_0 G_0$	$G_0 G_0$	$G_0 G_0$	$G_0 G_0$	$G_0 G_1$
SNP10	$G_0 G_1$	$G_0 G_0$	$G_1 G_1$	$G_0 G_1$	$G_0 G_0$

Wir möchten aufgrund des gegebenen Datensatzes die  $a$ -Werte schätzen. Dafür verwenden wir das folgende Modell

$$\hat{g}_d = 1\mu + Ma + \epsilon_d \quad (1)$$

wobei  $\hat{g}_d$  Vektor der deregressierten BLUP-Zuchtwerte  
 $\mu$  allgemeines Mittel  
 $a$  Allelsubstitutionseffekte  
 $M$  Inzidenzmatrix, welche  $\hat{g}_d$  und  $a$  verknüpft  
 $\epsilon_d$  zufällige Resteffekte

Wir nehmen an, dass für jedes Tier nur ein deregressierter Zuchtwert vorliegt.

### Ihr Aufgabe:

Stellen Sie den Vektor  $a$  und die Matrix  $M$  für den gezeigten Genotypendatensatz und das Modell (1) auf.

NB

Diese Aufgabe dient nur der Anschauung. Für den praktischen Einsatz wäre der Datensatz viel zu klein.

## Lösung

In der gezeigten Tabelle zählen wir als erstes die Anzahl der  $G_0$ -Allele. Diese speichern wir in einer Matrix.

```
matPosAllels <- matrix(NA, nrow = nrow(dfGenotypes), ncol = ncol(dfGenotypes))
matPosAllels[dfGenotypes == "$G_0G_0$"] <- 2
matPosAllels[dfGenotypes == "$G_0G_1$"] <- 1
matPosAllels[dfGenotypes == "$G_1G_1$"] <- 0
print(matPosAllels)
```

```
##      [,1] [,2] [,3] [,4] [,5]
## [1,]    2    1    2    1    2
## [2,]    2    1    2    2    1
## [3,]    2    1    1    1    1
## [4,]    2    2    2    1    1
## [5,]    1    2    2    1    2
## [6,]    1    1    1    2    2
## [7,]    0    1    1    1    1
## [8,]    1    2    1    2    1
## [9,]    2    2    2    2    1
## [10,]   1    2    0    1    2
```

Die gesuchte Matrix  $M$  hat Werte 1, 0 und -1 für die Genotypen  $G_0G_0$ ,  $G_0G_1$  und  $G_1G_1$ . Wir müssen als bei den Einträgen der oben gezeigten Matrix  $\$1$  abziehen. Da die Orientierung der Gleichungen im Modell so ist, dass jede Zeile für ein Tier steht müssen wir die Matrix noch transponieren.

```
matM <- t(matPosAllels-1)
#print(matM)
knitr::kable(matM)
```

1	1	1	1	0	0	-1	0	1	0
0	0	0	1	1	0	0	1	1	1
1	1	0	1	1	0	0	0	1	-1
0	1	0	0	0	1	0	1	1	0
1	0	0	0	1	1	0	0	0	1

Der Vektor  $a$  hat die Länge gleich der Anzahl SNPs in unserem Beispiel also 10.

## Aufgabe 2: Reduktion der Varianz

Bei der BLUP-Zuchtwertschätzung haben die geschätzten Zuchtwerte im Vergleich zu den phänotypischen Werten eine reduzierte Varianz. Dies können wir an folgendem Beispiel mit R zeigen. Wir verwenden dazu einen Datensatz aus einer Übung der Züchtungslehre. Der Datensatz wird mit folgendem Befehl eingelesen:

```
dfLmm <- read.csv2(file =
  "http://charlotte-ngs.github.io/LivestockBreedingAndGenomics/w7/z1_w7_u5_DataLmm.csv")
```

Die Struktur der Daten können wir mit dem Befehl `str` anzeigen. Die Kolonne `y` enthält die beobachteten Daten.

```
str(dfLmm)
```

```
## 'data.frame': 240 obs. of 3 variables:  
## $ ID : int 1 1 1 1 1 1 1 1 1 1 ...  
## $ FixerFactor: int -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 ...  
## $ y : num -35.6 -34 -35.4 -33.7 -37.4 ...
```

BLUP-Zuchtwerte werden mit dem Package `pedigreemm` geschätzt.

```
library(pedigreeemm)
```

```
## Loading required package: lme4
```

```
## Loading required package: Matrix
```

```
nAnzAnim <- 6
```

```
pedP1 <- pedigree(sire = as.integer(c(NA,NA,1, 1,4,5)),  
                 dam = as.integer(c(NA,NA,2,NA,3,2)),  
                 label = as.character(1:nAnzAnim))
```

```
fitReml <- pedigreeemm(formula = y ~ FixerFactor + (1 | ID),  
                      data = dfLmm,  
                      pedigree = list(ID = pedP1))
```

Die geschätzten Zuchtwerte erhalten wir aus dem Slot `u` aus dem Resultat-Objekt `fitReml`. Der Befehl

```
fitReml@u
```

```
## [1] 0.2223957 0.8957371 -0.5791965 -1.9366629 2.5201804 -0.5211967
```

zeigt den Vektor der geschätzten Zuchtwerte. Die Funktion `var()` kann nun verwendet werden um die Varianz der Beobachtungen mit der Varianz der geschätzten Zuchtwerte zu vergleichen.

```
var(dfLmm$y)
```

```
## [1] 1227.159
```

```
var(fitReml@u)
```

```
## [1] 2.300128
```