Peter von Rohr

Institut für Agrarwissenschaften

D-USYS

ETH Zürich

751-7602-00 V

Lösungen zur Prüfung

Angewandte Statistische Methoden in den Nutzierwissenschaften

FS 2017

Name:

Legi-Nr:

Aufgabe	Maximale Punktzahl	Erreichte Punktzahl
1	10	
2	16	
3	14	
4	22	
Total	62	

Aufgabe 1: Genomische Selektion

a) In der genomischen Selektion werden häufig Zielgrössen verwendet, welche auf BLUP-Zuchtwerten basieren. Was wird in der klassischen Zuchtwertschätzung als Zielgrösse verwendet? Wo liegen die Vor- und die Nachteile der jeweilen verwendeten Zielgrössen? Füllen Sie die nachfolgende Tabelle aus und geben Sie je einen Vor- und einen Nachteil der Zielgrössen vor und nach der Einführung der genomischen Selektion an.

6

Lösung:

Punkt	klassische Zuchtwertschätzung	Genomische Selektion
Zielgrössen	phänotypische Beobachtungen oder Leistungen	Deregressierte Zuchtwerte
Vorteile	unabhängig von Schätzverfahren, direkt beobachtbar oder messbar	frühe und breite Verfügbarkeit
Nachteile	späte Verfügbarkeit, uneinheitliche Definition	abhängig von Schätzverfahren, berechnete Grösse, Reduktion der Varianz und Schrumpfung zum Elterndurchschnitt

b) Angenommen wir würden rohe BLUP-Zuchtwerte als Zielgrössen in der genomischen Zuchtwertschätzung verwenden, welche Nachteile hätte das? Nennen Sie zwei Nachteile.

2

Lösung:

1. Reduktion der Varianz und
2. Schrumpfung zum Elterndurchschnitt

c) Wie lautet die Korrekturmassnahme zur Behebung der unter Aufgabe b) genannten Nachteile und auf welcher Grösse basiert diese Massnahme?

2

Lösung:

1. Deregression
2. Bestimmtheitsmass der geschätzten Zuchtwerte

Aufgabe 2: Lineare Regression

Wir haben den gleichen Datensatz mit zwei unterschiedlichen linearen Regressionsmodellen analysiert. Der R-Output dieser beiden Analysen ist nachfolgend als Output A und Output B gegeben.

Output A

```
##
## Call:
## lm(formula = y ~ X1, data = dfSimData)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.2899 -1.4864  0.2526  1.2982  4.6501
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -1.8929     2.6536  -0.713    0.482
## X1            4.0680     0.8675   4.689 6.49e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.108 on 28 degrees of freedom
## Multiple R-squared:  0.4399, Adjusted R-squared:  0.4199
## F-statistic: 21.99 on 1 and 28 DF,  p-value: 6.487e-05
```

Output B

```
##
## Call:
## lm(formula = y ~ -1 + X1, data = dfSimData)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.0925 -1.4013 -0.0846  1.6308  4.3171
##
## Coefficients:
##    Estimate Std. Error t value Pr(>|t|)
## X1   3.4557     0.1247   27.71   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.09 on 29 degrees of freedom
## Multiple R-squared:  0.9636, Adjusted R-squared:  0.9623
## F-statistic: 767.6 on 1 and 29 DF,  p-value: < 2.2e-16
```

a) Geben Sie die Formeln der beiden statistischen Modelle an, welche zu Output A und Output B geführt haben. Wo liegt der hauptsächliche Unterschied zwischen den beiden Modellen

8

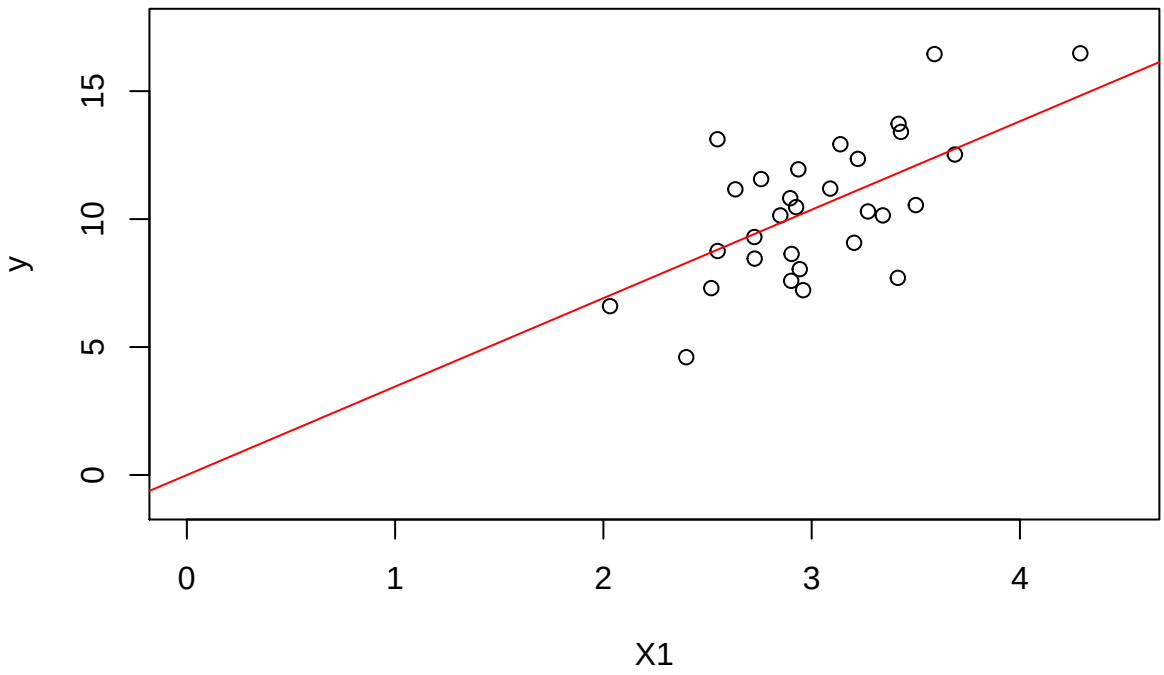
Lösung

- Modell für Output A: $y_i = \beta_0 + \beta_1 X_{1i} + \epsilon_i$
- Modell für Output B: $y_i = \beta_1 X_{1i} + \epsilon_i$
- Der Hauptunterschied liegt darin, dass im Modell von Output A ein Achsenabschnitt β_0 angepasst wird und im Modell von Output B nicht.

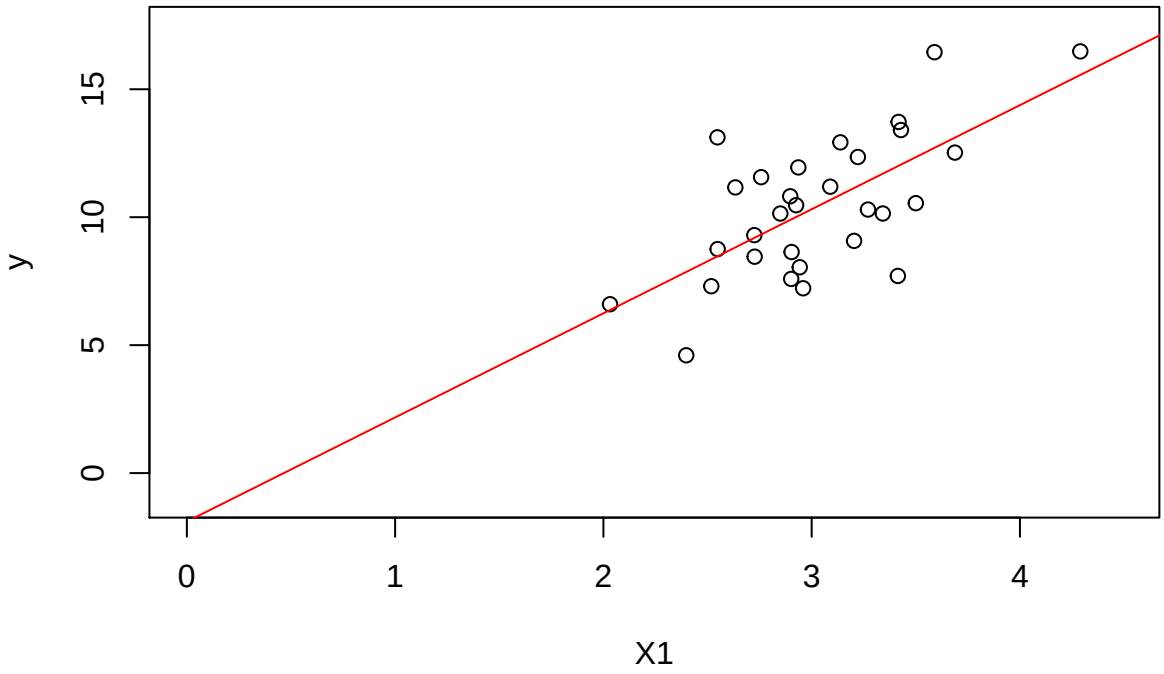
b) Für die zwei Analysen wurden auch zwei Plots gemacht. Ordnen Sie die Plots 1 und 2 den Outputs A und B zu.

2

Plot 1



Plot 2



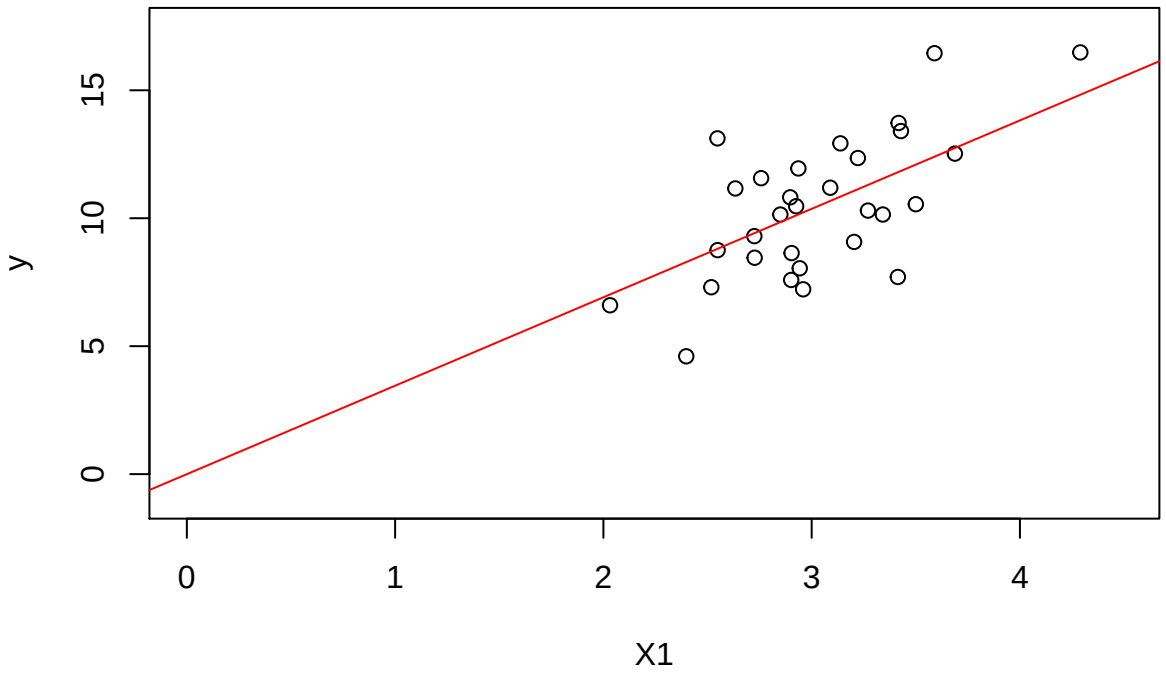
Lösung:

- Plot 1 gehört zu Output B
- Plot 2 gehört zu Output A

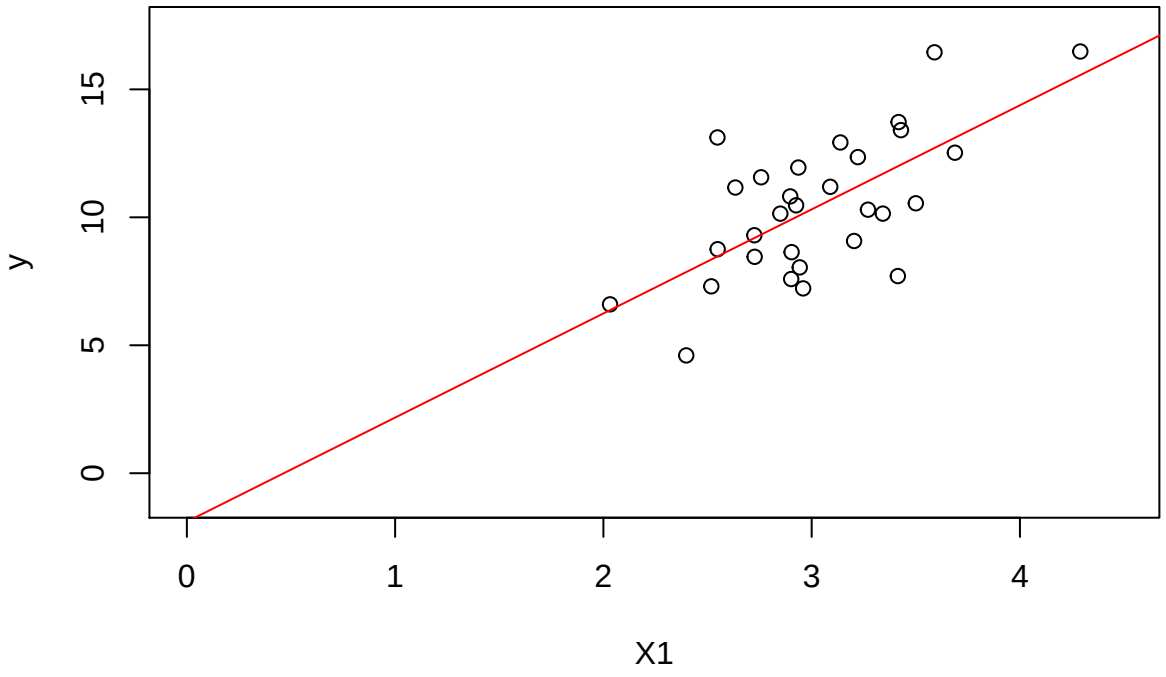
c) Zeichnen Sie die geschätzten Parameter (Estimate) aus den Outputs A und B in die Plots 1 und 2 ein

6

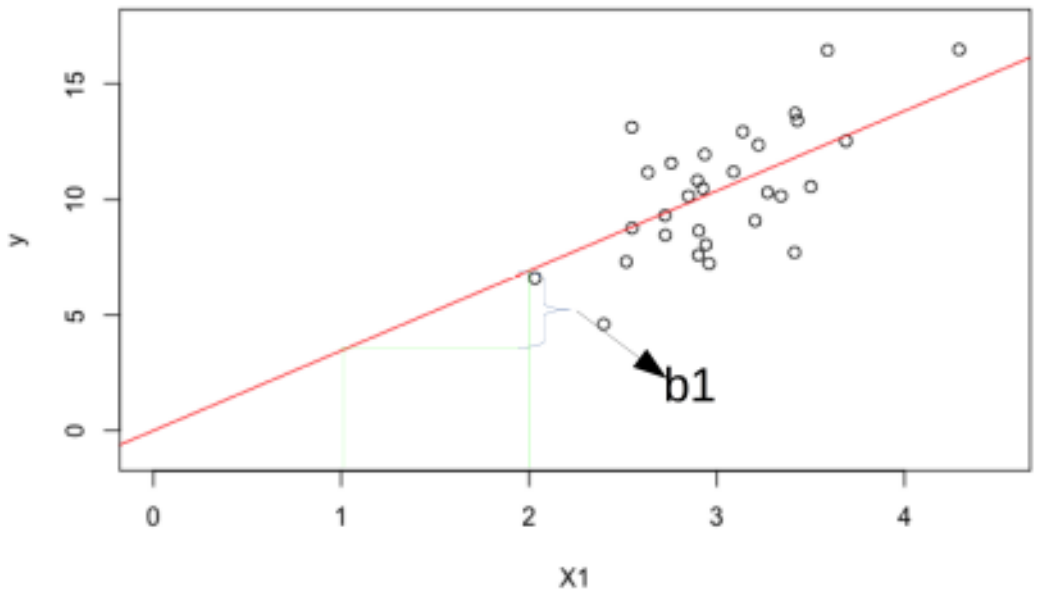
Plot 1

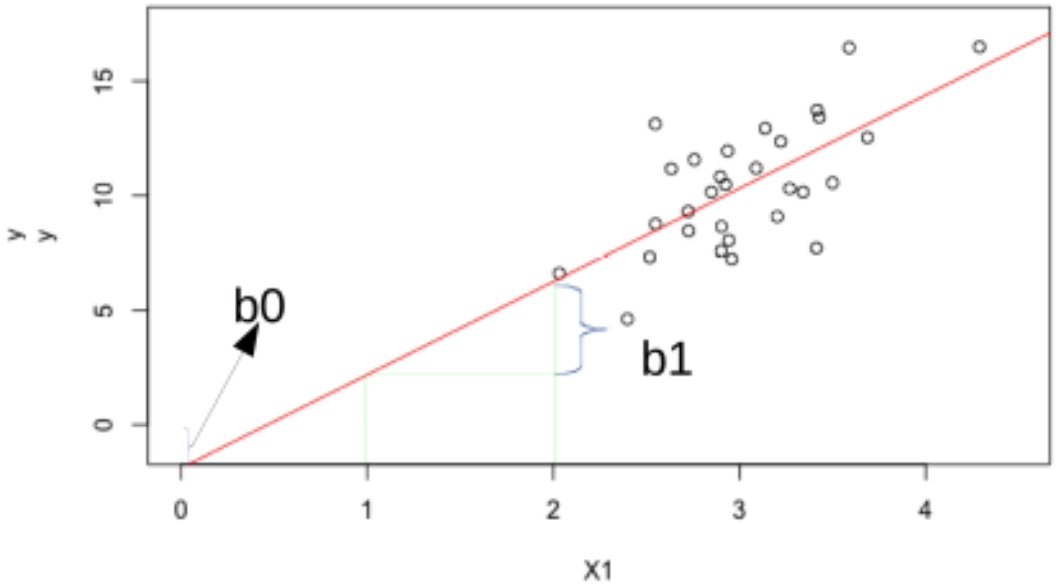


Plot 2



Lösung:





Aufgabe 3: LASSO und Bayes

a) In der genomischen Zuchtwertschätzung sind die SNP-Genotypen die hauptsächliche Informationsquelle. Für die statistische Modellierung dieser Daten können wir ein einfaches lineares Regressionsmodell verwenden. Weshalb kann bei der genomischen Zuchtwertschätzung Least Squares nicht als Schätzmethode verwendet werden?

4

Lösung

- $n << p$, wobei n die Anzahl Beobachtungen und p die Anzahl Parameter
- Für die Berechnung der Least Squares-Parameterschätzungen muss die Designmatrix X vollen Kolonnenrang haben, da nur dann $(X^T X)$ invertierbar ist.

b) LASSO (Least Absolute Shrinkage and Selection Operator) ist eine Alternative zu Least Squares. Worin unterscheiden sich LASSO und Least Squares?

4

Lösung

- Bei LASSO wird ein sogenannter Strafterm berücksichtigt.
- Strafterm führt dazu, dass gewisse Parameterschätzwerte β_j auf 0 gesetzt werden, weshalb auch eine Selektion der erklärenden Variablen passiert (Selection). Dies gibt es in Least Squares nicht.
- Durch die Berücksichtigung des Strafterms werden die Parameterschätzwerte auch gegen 0 gedrückt (Shrinkage). Dies gibt es in Least Squares nicht.
- Der Strafterm ist eine Funktion des Absolutbetrages des Parametervektors (Absolute). Dies gibt es in Least Squares nicht.

c) Unterschiede zwischen Bayesianer und Frequentisten

- Frequentisten unterscheiden in einer statistischen Analyse zwischen Daten und Parameter. Wie lautet die äquivalente Unterscheidung in einer Bayes'schen Analyse?
- Fehlende Daten werden in einer frequentistischen Datenanalyse ignoriert. Was passiert damit in einer Bayes'schen Analyse
- Aus welchem Grund muss die Bedingung $n > p$ in einer Bayes'schen Analyse nicht gelten?

3

Lösung

- Bayesianter unterscheiden zwischen bekannten und unbekannten Grössen.
- Fehlende Daten werden als unbekannte Grössen behandelt und aus den vorhandenen Daten mitgeschätzt
- Da die Berücksichtigung der a priori Information eine Schätzung auch mit sehr wenigen Beobachtungen erlaubt

d) In einer Bayes'schen Analyse basieren die Schätzung der unbekannten Grössen auf der sogenannten a posteriori-Verteilung. Aus welchen Komponenten besteht diese a posteriori Verteilung

3

Lösung

- Bayessche Likelihood
- a priori Verteilung
- Normalisierungskonstante

Aufgabe 4: Genomisches BLUP

a) Worin besteht der Unterschied zwischen RR-BLUP und GBLUP und wie werden die SNP-Informationen in RR-BLUP und in GBLUP berücksichtigen?

4

Lösung

In RR-BLUP werden die einzelnen SNP-Genotypen als fixe oder zufällige Effekte modelliert. Somit entspricht die Anzahl der unbekannten Parameter der Anzahl an SNPs. Die Berücksichtigung der SNP-Information erfolgt also über die direkte Modellierung der SNP-Effekte.

In GBLUP werden die genetischen Effekte aller SNP-Effekte pro Tier in einem zufälligen Effekt zusammengefasst. Die Berücksichtigung der SNP-Information erfolgt über die genomische Verwandtschaftsmatrix.

b) Wenn wir uns die Grösse der entstehenden Gleichungssysteme anschauen, welche Methode RR-BLUP oder GBLUP ergibt die kleineren Gleichungssysteme? Begründen Sie Ihre Antwort.

2

Lösung

Da in den meisten genomischen Analysen $n < p$ ist, dann haben wir in GBLUP die kleineren Gleichungssysteme. Die Begründung ist, dass in GBLUP die Anzahl der unbekannten Parameter der Anzahl Tiere n entspricht und in RR-BLUP ist die Anzahl der unbekannten Parameter gleich der Anzahl SNPs p.

c) Gegeben ist der folgende Datensatz. Bei den SNPs gilt G_1 als Allel mit der positiven Wirkung. Stellen Sie die Modelle als Formeln und die Gleichungssysteme für RR-BLUP auf. Verwenden Sie bei den Gleichungssystemen so weit als möglich die im Datensatz gegebenen Zahlenwerte. Als fixen Effekt können Sie ein allgemeines Mittel μ annehmen. Das Verhältnis zwischen Restvarianz und genetischer Varianz λ betrage $\lambda = 1$.

8

	Tier 1	Tier 2
SNP1	G_0G_0	G_1G_1
SNP2	G_0G_1	G_0G_1
SNP3	G_0G_0	G_1G_1
y	5.2	31.99

Lösung

Das Modell für RR-BLUP lautet:

$$y = 1_n\mu + Wq + e \tag{1}$$

wobei

- y Vektor der Länge n mit phänotypischen Beobachtungen
- μ allgemeines Mittel
- q Vektor der zufälligen additiven Effekte aller SNPs
- W Inzidenzmatrix, welche Genotypen für die SNPscodiert
- e Vektor der Resteffekte

Die Komponenten μ und q sind unbekannt und müssen aus den Daten geschätzt werden.

$$q = \begin{bmatrix} q_1 \\ q_2 \\ q_3 \end{bmatrix}$$

Die Matrix W verbindet die genetischen Effekte mit den Tieren. Die Elemente sind codiert als 0, 1 oder 2 je nachdem wie viele Allele mit positiver Wirkung in einem Genotyp enthalten sind. Die Matrix W lautet somit

$$W = \begin{bmatrix} 0 & 1 & 0 \\ 2 & 1 & 2 \end{bmatrix}$$

Der Vektor der Beobachtungen y ist definiert als

$$y = \begin{bmatrix} 5.20 \\ 31.99 \end{bmatrix}$$

Der Vektor der unbekannten Resteffekte ϵ ist definiert als

$$\epsilon = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \end{bmatrix}$$

Das Gleichungssystem als Ganzes lautet somit

$$\begin{bmatrix} 5.20 \\ 31.99 \end{bmatrix} = \begin{bmatrix} 1.00 \\ 1.00 \end{bmatrix} \mu + \begin{bmatrix} 0 & 1 & 0 \\ 2 & 1 & 2 \end{bmatrix} \begin{bmatrix} q_1 \\ q_2 \\ q_3 \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \end{bmatrix}$$

d) Verwenden Sie den gleichen Datensatz und die gleichen Annahmen, wie unter Aufgabe 4c) und stellen sie das Modell und das Gleichungssystem gleich wie in Aufgabe 4c, aber für GBLUP auf.

8

Lösung

Das Modell für GBLUP lautet:

$$y = 1_n\mu + Zg + \epsilon \tag{2}$$

wobei

y	Vektor der Länge n mit phänotypischen Beobachtungen
μ	allgemeines Mittel
g	Vektor der Länge n mit zufälligen additiven SNP-Effekten pro Individuum
Z	Inzidenzmatrix, welche SNP-Effekte mit Beobachtungen verknüpft
ϵ	Vektor von zufälligen Resteffekten

Berechnung der genomischen Verwandtschaftsmatrix

```
### # use the data from task1
data <- matDesignW
nmarkers <- ncol(data)
sumpq <- 0
freq <- dim(data)[1]
P <- freq
for(i in 1:ncol(data)){
  (freq[i] <- ((mean(data[,i])/2)))
  (P[i] <- (2*(freq[i]-0.5)))
  (sumpq <- sumpq+(freq[i]*(1-freq[i])))
}
Z <- data
for(i in 1:nrow(data)){
  for(j in 1:ncol(data)){
    (Z[i,j] <- ((data[i,j]-1)-(P[j])))
  }
}
Zt <- t(Z)
ZtZ <- Z%*%Zt
G <- ZtZ/(2*sumpq)
```

Aufstellen der MMG und Berechnung der Lösungen

```
lamda <- 1
matG <- G
for(i in 1:nrow(G)){
  (matG[i,i] <- ((matG[i,i]+0.01)))
}
# matrix X
matX <- matrix(1,nrow=nAnzTiere,1)
matXtX <- crossprod(matX)
matZ <- diag(1,nAnzTiere)
matXtZ <- crossprod(matX,matZ)
matZtZ <- crossprod(matZ)
matCoeff <- cbind(rbind(matXtX,t(matXtZ)),rbind(matXtZ,matZtZ + lamda * solve(matG)))
```

```
vecRhs <- rbind(crossprod(matX,y),crossprod(matZ,y))
vecSol <- solve(matCoeff,vecRhs)
```

1. Die Elemente y, 1_n, μ und ϵ sind gleich wie im RR-BLUP Modell. Die Matrix Z und der Vektor g sind wie folgt definiert.

$$Z = \begin{bmatrix} 1.00 & 0.00 \\ 0.00 & 1.00 \end{bmatrix}$$

Der Vektor g enthält die genetischen Effekte pro Individuum über alle SNP.

$$g = \begin{bmatrix} g_1 \\ g_2 \end{bmatrix}$$

2. Die Genomische Verwandtschaftsmatrix, wie sie nach dem R-Programm aus [@CV2013] berechnet wird lautet

```
cat(" * Genomische Verwandtschaftsmatrix G:\n")
```

```
##  * Genomische Verwandtschaftsmatrix G:
```

```
print(G)
```

```
##           [,1]      [,2]
## [1,]  1.333333 -1.333333
## [2,] -1.333333  1.333333
```

3. Zur Berechnung der Lösung mit GBLUP müssen wir die entsprechenden Mischmodellgleichungen aufstellen. Diese lauten

$$\begin{bmatrix} 2.00 & 1.00 & 1.00 \\ 1.00 & 51.19 & 49.81 \\ 1.00 & 49.81 & 51.19 \end{bmatrix} \begin{bmatrix} \hat{\mu}_1 \\ \hat{g}_1 \\ \hat{g}_2 \end{bmatrix} = \begin{bmatrix} 37.19 \\ 5.20 \\ 31.99 \end{bmatrix}$$

Der Lösungsvektor lautet

$$\begin{bmatrix} \hat{\mu}_1 \\ \hat{g}_1 \\ \hat{g}_2 \end{bmatrix} = \begin{bmatrix} 18.60 \\ -9.75 \\ 9.75 \end{bmatrix}$$