Peter von Rohr
Institute of Agricultural Sciences
D-USYS
ETH Zurich

751-7602-00 V
Solutions for Exam
Applied Statistical Methods
in Animal Sciences
SS 2020

Date: 25th May 2020

Name: Firstname Name

Legi-Nr: LegiNr

Problem	Maximum Number of Points	Number of Points Reached
1	14	
2	15	
3	32	
4	19	
Total	80	

Questions in German are in italics

Problem 1: Traditional Breeding versus Genomic Selection

In Livestock breeding three different sources of information are available

1. Performance data
2. Pedigree information
3. Genomic Marker Data

In der Tierzucht stehen uns drei verschiedene Informationsquellen zur Auswahl

1. *Leistungsdaten*
2. *Pedigreeinformationen*
3. *Genomische Markerdaten*

a) Complete the following table indicating which source of information is used in either traditional breeding or in genomic selection.

Vervollständigen Sie die folgende Tabelle und geben Sie an, welche Informationsquelle in der traditionellen Tierzucht oder der genomischen Selektion verwendet werden.

6

Source of Information	Traditional Breeding	Genomic Selection
Performance data		
Pedigree information		
Genomic Marker Data		

Solution

Source of Information	Traditional Breeding	Genomic Selection
Performance data	yes	yes
Pedigree information	yes	yes
Genomic Marker Data	no	yes

b) In the two step approach of genomic selection the reference population is used to estimated marker effects. Animals with high reliabilities ($B\%$) are included in the reference population. Let us assume that we have a population with 10569 breeding animals. What is the expected number of animals in the reference population, if we require a reliability of at least 59% for animals to be included into the reference population. We assume the reliabilities on the percentage scale to follow a normal distribution with mean 41% and standard deviation of 11%

Im Zwei-Schritt Verfahren der genomischen Selektion werden die Markereffekte anhand einer Referenzpopulation geschätzt. Tiere mit einem hohen Bestimmtheitsmass ($B\%$) werden in die Referenzpopulation aufgenommen. Wir nehmen an, dass wir eine Poplation von einer totalen Grösse von 10569 Tiere haben. Was ist die erwartete Anzahl Tiere in der Referenzpopulation, wenn wir ein Bestimmtheitsmass von mindestens 59% *für Tiere in der Referenzpopulation verlangen. Wir nehmen an, dass das Bestimmtheitsmass auf der Prozentskala einer Normalverteilung mit Mittelwert* 41% *und Standardabweichung* 11% *folgt.*

4

Solution

The size of the reference population can be determined by the proportion of the normal density that is above the threshold of 59%. The mean and the standard deviation are given in the problem description. Hence the proportion is given as

```
prop_ref <- 1- pnorm(n_req_rel, mean = n_mean_rel, sd = n_sd_rel)
cat(" * Proportion of animals in reference population: ", prop_ref, "\n")
```

```
##  * Proportion of animals in reference population:  0.05088175
```

The number of animals is the proportion times the total number of animals in the population.

```
n_nr_ref <- prop_ref * n_pop_size
cat(" * The number of animals in the reference population is: ", floor(n_nr_ref), "\n")
```

```
##  * The number of animals in the reference population is:  537
```

c) To be able to obtain good estimates of marker effects, the minimum size of the reference population is set to 1000 animals. What is the minimum reliability ($B\%$) to get to a reference population of 1000 animals using the same distributional assumptions for the reliabilities as under 1b).

Damit wir zuverlässige Markereffektschätzungen erhalten müssen wir eine Referenzpopulation von mindestens 1000 Tiere haben. Wie muss die untere Grnze für das Bestimmtheitsmass ($B\%$) festgelegt werden, damit wir eine Referenzpopulation von 1000 Tieren erhalten. Die Annahmen betreffend der Verteilung der Bestimmtheitsmasse können Sie aus der Aufgabe 1b) übernehmen.

4

Solution

Given that we want to have a reference population of at least 1000, we can fix the proportion of the reference animals compared to the total population.

```
n_prop_ref <- n_min_ref_size/n_pop_size
cat(" * Proportion of reference animals: ", round(n_prop_ref, digits = 3), "\n")
```

```
##  * Proportion of reference animals:  0.095
```

With the proportion, we determine the quantile

```
quant_rel <- qnorm(n_prop_ref, mean = n_mean_rel, sd = n_sd_rel, lower.tail = FALSE)
cat(" * Minimum reliability: ", floor(quant_rel), "\n")
```

```
##  * Minimum reliability:  55
```

Problem 2: Fixed Linear Effect Model

Hyperketonemia (HYK) is a metabolic disorder in cattle characterized by elevated levels of blood ketone bodies. The disorder affects early lactating cows. Ketone bodies such as beta-hydroxy-butyrate (BHB) and acetone (ACE) can be diagnosed in blood, milk and urine. For diagnostic reasons it is interesting to be able to predict ketone bodies such as BHB in the blood based on measurements of BHB and ACE in the milk. In a publication by Chandler et al. (2018) the following regression coefficients were found for cows which were not diagnosed of HYK. In that regression model BHB in blood serum (bBHB) was the response variable and a number of predictors either measured in a milk sample or observed from the cows performance record were used. These predictors are listed in the column entitled by the term `Variable` in the table below.

Hyperketonemia (HYK) ist eine Stoffwechselstörung beim Rind, welche sich durch erhöhte Werte von Ketonkörpern im Blut manifestiert. Die Störung betrifft vor allem Milchrinder zu Beginn der Laktation. Ketonkörper wie Beta-Hydroxy-Butyrat (BHB) und Aceton (ACE) können im Blut, in der Milch und im Urin diganostiziert werden. Aus Gründen der Diagnostik ist es interessant Ketonkörper, wie z.B. BHB im Blut über Messungen von BHB und ACE in der Milch schätzen zu können. In einer Publikation von Chandler et al. (2018) wurden die folgenden Regressionskoeffizienten für Kühe, welche keine HYK-Diagnose erhielten, gefunden. In diesem Regressionsmodell wurde BHB im Blut (bBHB) als Zielgrösse festgelegt. Verschiedene Milchinhaltsstoffe oder Leistungsparameter der untersuchten Kühe dienten als unabhängige Variablen verwendet.

Variable	Regression Coefficients (Non-HYK)
Intercept	-2.380
Milk acetone, mmol/L	1.220
Milk protein, %	-0.100
Fat-to-protein ratio	0.240
Production on test day, kg	-0.003
Gestation length, d	0.010

a) Compute the predicted value of 'bBHB' for the three cows which are described in the following table.

Schätzen Sie die 'bBHB'-Werte der drei Kühe, welche in der nachfolgenden Tabelle beschrieben sind.

9

Variable	Observed Values of Cow A	Observed Values of Cow B	Observed Values of Cow C
Milk acetone, mmol/L	0.116	0.003	0.094
Milk protein, %	3.400	3.320	3.320
Milk fat, %	4.290	4.240	4.300
Production on test day, kg	22.200	22.600	24.000
Gestation length, d	276.000	275.000	274.000

Solution

The predicted value ($\hat{y}$) for `bBHB` can be computed using the formula for the linear regression. So for cow i this means

$$\hat{y}_i = \hat{b}_0 + x_i^T \cdot \hat{b}$$

where $\hat{b}_0$ is the estimate for the intercept, $\hat{b}$ is the vector of estimated regression coefficients and x_i corresponds to the vector of observed predictor values for cow i.

```
vec_cow_A <- c(1,vec_mace[1],vec_mprot[1], vec_mfat[1]/vec_mprot[1], vec_mprod[1], vec_gest_len[1])
vec_cow_B <- c(1,vec_mace[2],vec_mprot[2], vec_mfat[2]/vec_mprot[2], vec_mprod[2], vec_gest_len[2])
vec_cow_C <- c(1,vec_mace[3],vec_mprot[3], vec_mfat[3]/vec_mprot[3], vec_mprod[3], vec_gest_len[3])

bBHB_cow_A <- crossprod(vec_cow_A, tbl_bhb_reg$`Regression Coefficients (Non-HYK)`)
bBHB_cow_B <- crossprod(vec_cow_B, tbl_bhb_reg$`Regression Coefficients (Non-HYK)`)
bBHB_cow_C <- crossprod(vec_cow_C, tbl_bhb_reg$`Regression Coefficients (Non-HYK)`)

tbl_bhb_blood <- tibble::tibble(Cow = c("A","B","C"),
                                `Blood BHB` = c(bBHB_cow_A,bBHB_cow_B,bBHB_cow_C))
knitr::kable(tbl_bhb_blood,
             booktabs = TRUE,
             longtable = TRUE)
```

Cow	Blood BHB
A	0.4177435
B	0.2803660
C	0.3815234

b) According to the publication by Chandler et al. (2018), the mean levels of BHB and Aceton found in the blood serum and in the milk are higher for cows with a diagnosed HYK compared to cows which are HYK-free. The different levels are listed in the table below. Is it possible to use the regression coefficients listed in Problem 2a) to predict the levels of blood BHB for cows that have HYK? Please reason about your answer.

Gemäss der Publikation von Chandler et al. (2018) liegen die mittleren Werte für BHB und Aceton im Blut und in der Milch höher bei den Kühen, bei welchen eine HYK diagnostiziert wurde im Vergleich zu den HYK-freien Kühe. Die Werte sind in der nachfolgenden Tabelle angegeben. Ist es möglich den Blut-BHB-Wert einer auf HYK diagnostizierten Kuh anhand der unter Aufgabe 2a aufgelisteten Regressionskoeffizienten zu schäten? Bitte begründen Sie Ihre Antwort.

4

Variable	Non-HYK Cows	HYK Cows
Blood Serum BHB, mmol/L	0.60	2.30
Milk BHB, mmol/L	0.07	0.15
Milk Acetone, mmol/L	0.08	0.69
Fat, %	4.37	5.10
Protein, %	3.36	2.97
Production on test-day, kg	23.90	22.20
Gestation length, d	275.00	282.00

Solution

Predictions cannot be done for HYK-cows based on regression coefficients that were estimated only on HYK-free cows. This would correspond to a case of "extra-polation" which is not suitable here due to the differences in the shown average values.

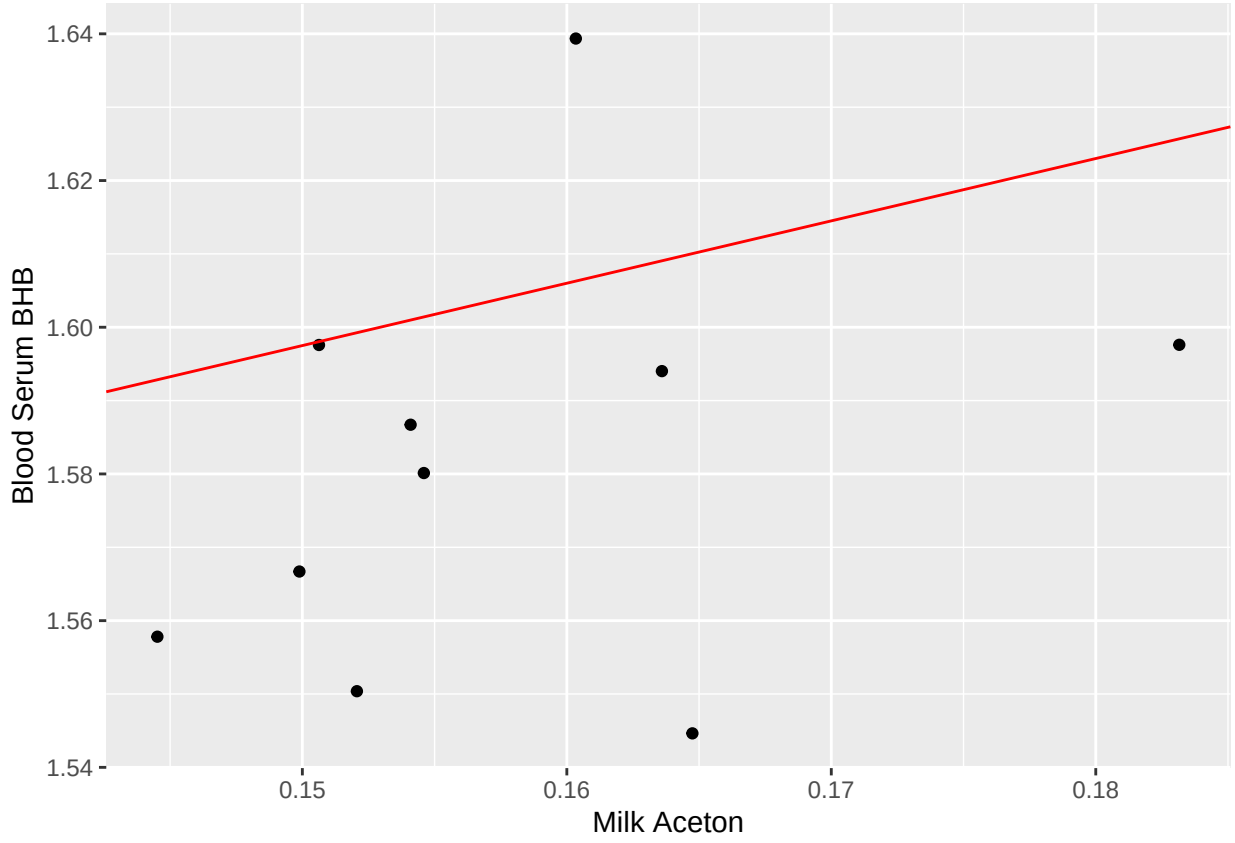
c) As shown in the table below, there are also breed differences in the levels of keton bodies in blood serum and in the milk. The following tables contains levels of blood-BHB and aceton in the milk for cows of the breeds Holstein and Jersey. When looking only at the intercept of the complete model and at the regression coefficient of milk-aceton for the two breeds the following two plots can be drawn. Please indicate which plot (A and B) belong to which breed.

Gemäss der folgenden Tabelle gibt es auch Rassenunterschiede bei den Werten der Ketonkörper im Blut und in der Milch. Die folgende Tabelle zeigt Blut-BHB und Milch Aceton für Kühe der Rassen Holstein und Jersey. Schaut man sich nur den Achsenabschnitt und den Regressionskoeffizienten von Milch-Aceton für die zwei Rassen an, dann entstehen die unten gezeigten Plots für die beiden Rassen. Bitte geben Sie an, welches Diagramm (A oder B) zu welcher Rasse gehört.

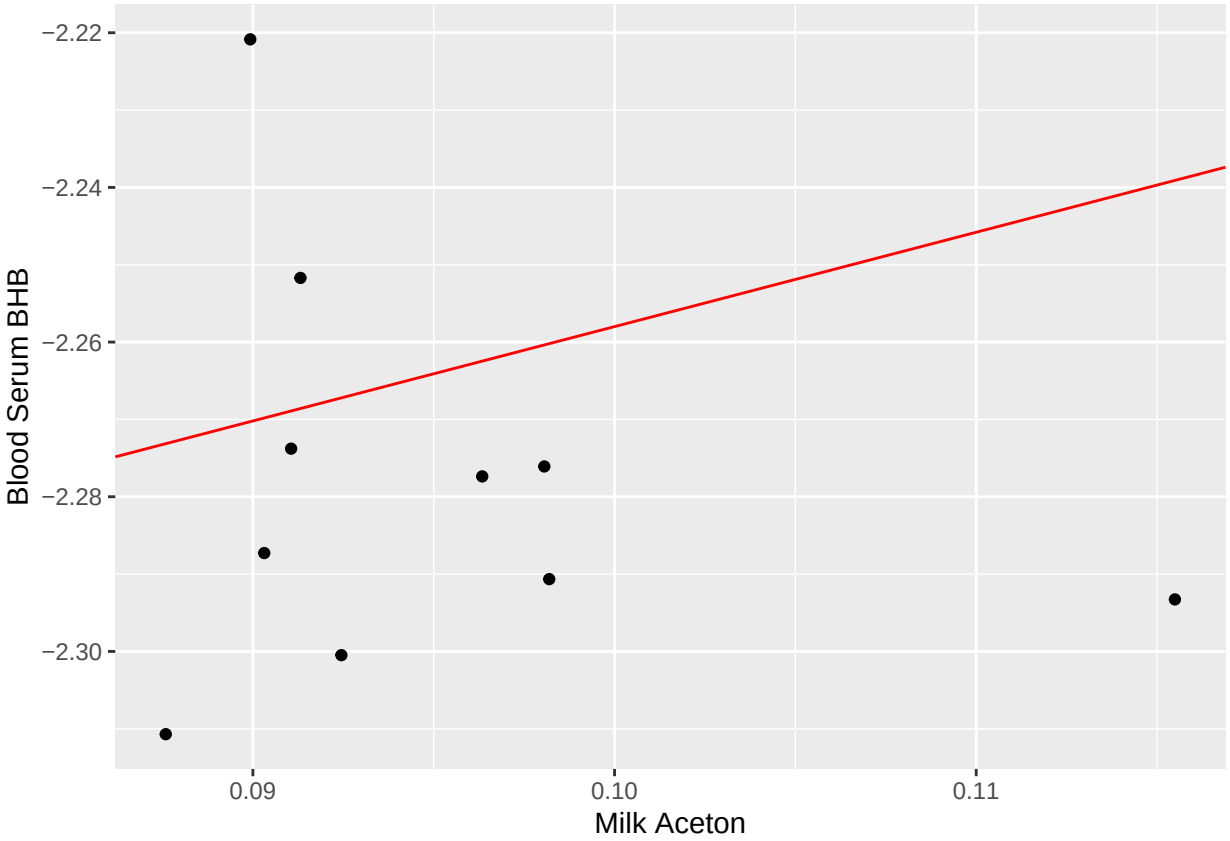
2

Variable	Holstein	Jersey
Mean Blood Serum BHB, mmol/L	0.800	0.920
Standard deviation Blood Serum BHB, mmol/L	0.030	0.030
Mean Acetone, mmol/L	0.101	0.159
Standard deviation Acetone, mmol/L	0.010	0.010
Intercept	-2.380	1.470
Regression Coefficient	1.220	0.850

Plot A



Plot B



Solution

Plot A is based on Jersey and Plot B is based on Holstein.

Problem 3: Genomic BLUP

The dataset shown below on blood-BHB is used to quantify the influence of 5 SNP-loci. The only fixed effect in this model is the herd.

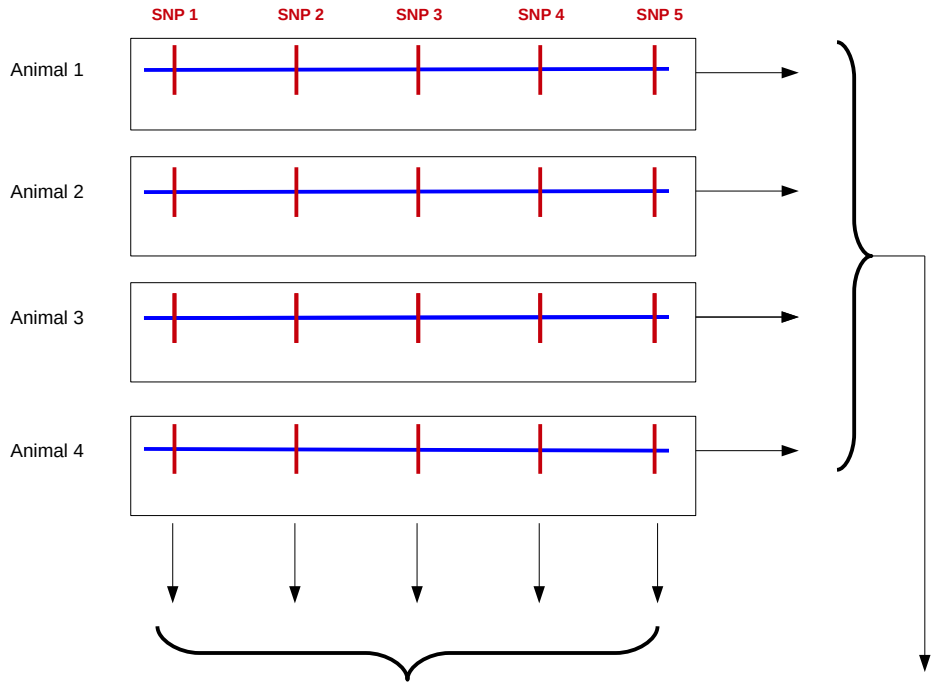
Der unten gezeigte Datensatz zu BHB im Blut soll verwendet werden um den Einfluss von 5 SNP-Loci zu quantifizieren. Der Betrieb wird als fixer Effekt berücksichtigt.

Tier	bBHB	Herd	SNP1	SNP2	SNP3	SNP4	SNP5
1	1.567	1	-1	0	1	1	1
2	1.598	2	1	1	0	1	0
3	1.598	1	-1	0	-1	1	-1
4	1.639	1	1	1	-1	1	0

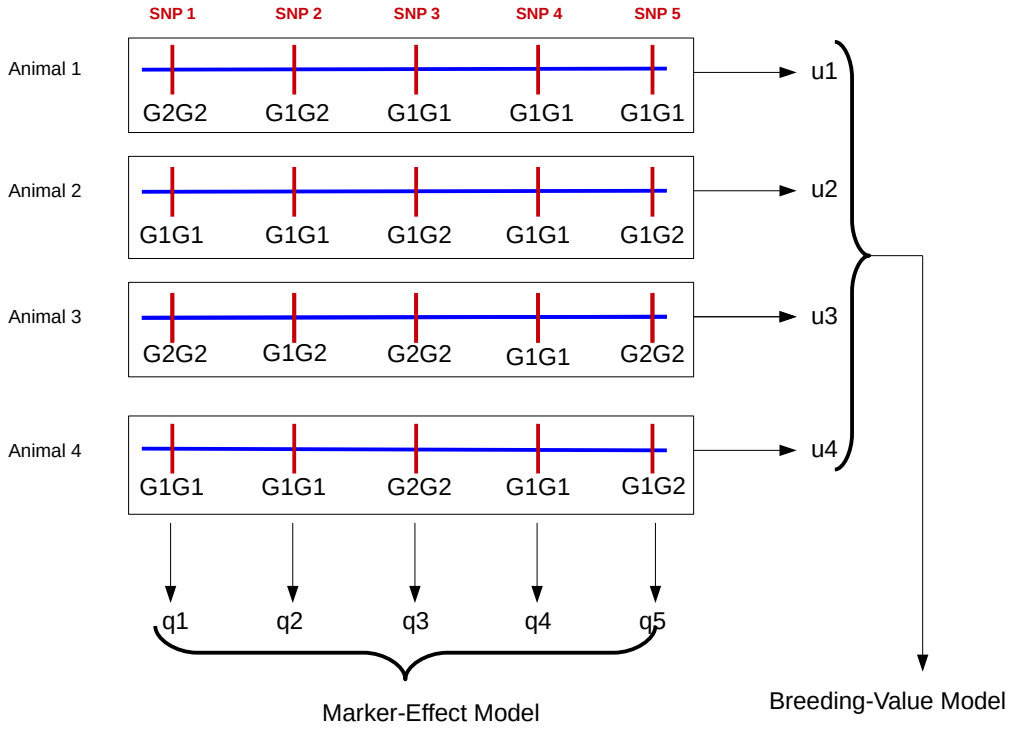
a) GBLUP models can either be specified as marker-effect models or as breeding-value models. Please complete the following diagram which illustrates the difference between the two types of models. Enter the SNP-genotypes for all animals at each SNP-position. Use G_1 as the allele with the positive effect.

GBLUP Modelle können entweder als Marker-Effekt-Modelle oder als Zuchtwertmodelle angegeben werden. Bitte vervollständigen Sie das folgende Diagramm, welches den Unterschied zwischen den beien Modelltypen erklärt. Bitte notieren Sie die SNP-Genotypen der Tiere an allen SNP-Positionen. Verwenden Sie G_1 als das Allele mit der positiven Wirkung.

8



Solution


SNP 1
SNP 2
SNP 3
SNP 4
SNP 5
Animal 1
G2G2 G1G2 G1G1 G1G1 G1G1
u1
Animal 2
G1G1 G1G1 G1G2 G1G1 G1G2
u2
Animal 3
G2G2 G1G2 G2G2 G1G1 G2G2
u3
Animal 4
G1G1 G1G1 G2G2 G1G1 G1G2
u4
q1 q2 q3 q4 q5
Marker-Effect Model
Breeding-Value Model

b) Use a marker-effect model for the above given dataset. Specify all model components and enter the numbers from the dataset into the model. Setup the mixed model equations to solve for the parameters to be estimated. You can assume a ratio of 1 between the genetic and the residual variance components.

Verwenden Sie ein Marker-Effekt Modell für den oben gegebenen Datensatz. Geben Sie alle Modellkomponenten an und verwenden Sie die Zahlen des Datensatzes als Information im Modell. Stellen Sie die Mischmodellgleichungen auf für die Lösung nach den zu schätenden Parameter. Sie können ein Verhältnis von 1 *annehmen zwischen der genetischen Varianz und der Varianz der Resteffekte.*

12

Solution

1. Model

$$y = Xb + Wq + e$$

where

- y is the vector of observations.

$$y = \begin{bmatrix} 1.651 \\ 1.585 \\ 1.586 \\ 1.696 \end{bmatrix}$$

- b is the vector of fixed effect levels for the two farms

$$b = \begin{bmatrix} b_1 \\ b_2 \end{bmatrix}$$

- X is the design matrix for the fixed effects

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

- q is the vector of marker effects for the five SNP-markers

$$q = \begin{bmatrix} q_1 \\ q_2 \\ q_3 \\ q_4 \\ q_5 \end{bmatrix}$$

- W is the indicator matrix for the SNP-effects

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

- e is the vector of random error terms

$$e = \begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ e_4 \end{bmatrix}$$

2. Expected values and Variance-Covariance Matrix

- Expected values of all random components

$$E \begin{bmatrix} e \\ q \\ y \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ Xb \end{bmatrix}$$

- Variance-Covariance Matrix

$$var \begin{bmatrix} e \\ q \\ y \end{bmatrix} = \begin{bmatrix} R & 0 & R \\ 0 & Q & QW^T \\ R & WQ & V \end{bmatrix}$$

where $R = I\sigma_e^2$, $Q = I\sigma_q^2$ and $V = WQW^T + R$. The quantities σ_e^2 and σ_q^2 correspond to the residual variance component and the variance component attributed to SNPs.

3. Mixed Model equations

The mixed model equations are give by

$$\begin{bmatrix} X^TX & X^TW \\ W^TX & W^TW + I * \lambda \end{bmatrix} \begin{bmatrix} \hat{b} \\ \hat{q} \end{bmatrix} = \begin{bmatrix} X^Ty \\ W^Ty \end{bmatrix}$$

with $\lambda = \sigma_e^2/\sigma_q^2$

The solution

```
mat_XTX <- crossprod(mat_X)
mat_XTW <- crossprod(mat_X, mat_W)
mat_WTX <- crossprod(mat_W, mat_X)
mat_WTWIlambda <- crossprod(mat_W) + diag(1,nrow = n_nr_snp) * lambda
mat_coef <- rbind(cbind(mat_XTX, mat_XTW), cbind(mat_WTX, mat_WTWIlambda))
mat_rhs <- rbind(crossprod(mat_X, vec_y), crossprod(mat_W, vec_y))
mat_sol <- solve(mat_coef, mat_rhs)
vec_sol_farm <- mat_sol[1:2,]
vec_sol_snp <- mat_sol[(3:nrow(mat_sol)),]
mat_sol_g <- crossprod(t(mat_W), vec_sol_snp)
```

The solution vector $\hat{s}$ for the herd effects and the marker effects is

$$\hat{s} = \begin{bmatrix} 1.6504 \\ 1.5468 \\ 0.0255 \\ 0.0127 \\ 0.0054 \\ 0 \\ 0.0181 \end{bmatrix}$$

The solution for the farm effects $\hat{b}$ are the first two components of $\hat{s}$

$$\hat{b} = \begin{bmatrix} 1.6504 \\ 1.5468 \end{bmatrix}$$

The solutions for the marker effects $\hat{q}$ are

$$\hat{q} = \begin{bmatrix} 0.0255 \\ 0.0127 \\ 0.0054 \\ 0 \\ 0.0181 \end{bmatrix}$$

The direct genomic breeding values $\hat{g}$ are obtained by multiplying the matrix W by the predicted marker effects $\hat{q}$. Hence,

$$\hat{g} = W \cdot \hat{q}$$

$$\hat{g} = \begin{bmatrix} -0.002 \\ 0.0382 \\ -0.0489 \\ 0.0329 \end{bmatrix}$$

c) Use a breeding-value model for the above given dataset. Specify all model components and enter the numbers from the dataset into the model. Setup the mixed model equations to solve for the parameters to be estimated. You can assume a ratio of 1 between the genetic and the residual variance components.

Verwenden Sie ein Zuchtwertmodell für den oben gegebenen Datensatz. Geben Sie alle Modellkomponenten an und verwenden Sie die Zahlen des Datensatzes als Information im Modell. Stellen Sie die Mischmodellgleichungen auf für die Lösung nach den zu schätenden Parameter. Sie können ein Verhältnis von 1 *annehmen zwischen der genetischen Varianz und der Varianz der Resteffekte.*

12

Solution

1. Model

$$y = Xb + Zg + e$$

where y, X, b and e are the same as under 3b)

- g: vector of random genomic breeding values

$$g = \begin{bmatrix} g_1 \\ g_2 \\ g_3 \\ g_4 \end{bmatrix}$$

- Z: Indidence matrix linking genomic breeding values to observations.

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

2. Expected values and Variance-Covariance Matrix

- Expected values for the random components

$$E \begin{bmatrix} e \\ g \\ y \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ Xb \end{bmatrix}$$

- Variance-covariance matrix

$$var \begin{bmatrix} e \\ g \\ y \end{bmatrix} = \begin{bmatrix} R & 0 & R \\ 0 & G & GZ^T \\ R & ZG & V \end{bmatrix}$$

with $R = I * \sigma_e^2$ (σ_e^2 residual variance component), $G = H * \sigma_g^2$ (σ_g^2 the genetic variance component) and $V = ZGZ^T + R$.

3. Mixed Model equations

The mixed model equations are

$$\begin{bmatrix} X^TX & X^TZ \\ Z^TX & Z^TZ + H^{-1} * \lambda \end{bmatrix} \begin{bmatrix} \hat{b} \\ \hat{g} \end{bmatrix} = \begin{bmatrix} X^Ty \\ Z^Ty \end{bmatrix}$$

with $\lambda = \sigma_e^2/\sigma_g^2$ and H corresponding to the genomic relationship matrix.

The genomic relationship matrix H can be computed with the following function.

```
computeMatGrm <- function(pmatData) {
  matData <- pmatData
  # check the coding, if matData is -1, 0, 1 coded, then add 1 to get to 0, 1, 2 coding
  if (min(matData) < 0) matData <- matData + 1
  # Allele frequencies, column vector of P and sum of frequency products
  freq <- apply(matData, 2, mean) / 2
  P <- 2 * (freq - 0.5)
  sumpq <- sum(freq*(1-freq))
  # Changing the coding from (0,1,2) to (-1,0,1) and subtract matrix P
  Z <- matData - 1 - matrix(P, nrow = nrow(matData),
                            ncol = ncol(matData),
                            byrow = TRUE)
  # Z%*%Zt is replaced by tcrossprod(Z)
  return(tcrossprod(Z)/(2*sumpq))
}
mat_H <- computeMatGrm(pmatData = mat_W)
n_cfact <- 0.05
mat_Hstar <- mat_H + n_cfact * diag(1, nrow = n_anz_snp_tiere)
```

The computed matrix H is computationally singular. This can be corrected by adding a small fraction of the numerator relationship matrix to H. Because, we do not have any information about the pedigree of the animals in the dataset, we replace the numerator relationship matrix by the identity matrix I which is the same as assuming that the animals are unrelated. Therefore we get a new genomic relationship matrix H^* by the following formula

$$H^* = H + 0.05 * I$$

From now on the matrix H^* is used as the genomic relationship matrix.

The solution of the mixed model equations can be obtained, as follows.

```
mat_XTX <- crossprod(mat_X)
mat_XTZ <- crossprod(mat_X, mat_Z)
mat_ZTX <- crossprod(mat_Z, mat_X)
mat_ZTZHinvlambda <- crossprod(mat_Z) + solve(mat_Hstar) * lambda
mat_coef <- rbind(cbind(mat_XTX, mat_XTZ), cbind(mat_ZTX, mat_ZTZHinvlambda))
mat_rhs <- rbind(crossprod(mat_X, vec_y), crossprod(mat_Z, vec_y))
mat_sol <- solve(mat_coef, mat_rhs)
vec_sol_farm <- mat_sol[1:2,]
vec_sol_g <- mat_sol[(3:nrow(mat_sol)),]
```

The solutions for the herds are

$$\hat{b} = \begin{bmatrix} 1.6533 \\ 1.5582 \end{bmatrix}$$

The solutions for the genomic breeding values break

$$\hat{g} = \begin{bmatrix} -0.0061 \\ 0.0268 \\ -0.0445 \\ 0.0237 \end{bmatrix}$$

Problem 4: Bayes

a) In a Bayesian data analysis, we differentiate between known and unknown quantities. Please, separate the relevant quantities for the dataset and the model shown below into known and unknown quantities.

In einer Bayes'schen Datenanalyse wird zwischen bekannten und unbekannten Grössen unterschieden. Machen Sie die Einteilung für den unten gezeigten Datensatz und das angegebene Modell.

13

Tier	bBHB	Herd	SNP1	SNP2	SNP3	SNP4	SNP5
1	1.567	1	-1	0	1	1	1
2	1.598	2	1	1	0	1	0
3	1.598	1	-1	0	-1	1	-1
4	1.639	1	1	1	-1	1	0

For the above shown dataset, we assume the following model

Für den Datensatz nehmen wir das folgende Modell an

$$y = Xb + Zq + e$$

where: Herd is modelled as a fixed effect (b) and SNP-effects (q) are taken as random effects.

Grösse	bekannt	unbekannt
y_1		
y_2		
y_3		
y_4		
b_{H1}		
b_{H2}		
X		
q_1		
q_2		
q_3		
q_4		
q_5		
Z		

Solution

{

Grösse	bekannt	unbekannt
y_1	y	
y_2	y	
y_3	y	
y_4	y	
b_{H1}	n	
b_{H2}	n	
X	y	
q_1	n	
q_2	n	
q_3	n	
q_4	n	
q_5	n	
Z	y	

}

b) In a first step of the Bayesian analysis, the influence of the 'herd' on the response variable 'bBHB' is to be quantified. This is done by the following junk of R-code. Compute the Bayesian estimate for the intercept and the effects of the two herds, based on the output shown below.

In einem ersten Schritt der Bayes'schen Analyse soll der Einfluss des Betriebs ('herd') auf die Zielgrösse 'bBHB' abgeschätzt werden. Dies wird mit dem folgenden R-programm gemacht. Berechnen Sie die Bayes'sche Schätzung für den Achsenabschnitt und die Effekte der beiden Betriebe aufgrund des unten gezeigten Outputs.

3

```
### # Matrix X as incidence matrix for beta0 and beta1
X <- cbind(1,model_matrix_bhb)
### # y as vector of observations
y <- tbl_herd_bhb_gen_data$bBHB
### # starting values
beta = c(0, 0, 0)
# loop for Gibbs sampler
niter = 10 # number of samples
for (iter in 1:niter) {
# sampling intercept
w = y - X[, 2] * beta[2] - X[, 3] * beta[3]
x = X[, 1]
xpxi = 1/(t(x) %*% x)
betaHat = t(x) %*% w * xpxi
beta[1] = rnorm(1, betaHat, sqrt(xpxi)) # using residual var = 1
# sampling herd 1
w = y - X[, 1] * beta[1] - X[, 3] * beta[3]
x = X[, 2]
xpxi = 1/(t(x) %*% x)
betaHat = t(x) %*% w * xpxi
beta[2] = rnorm(1, betaHat, sqrt(xpxi)) # using residual var = 1
# sampling herd 2
w = y - - X[, 1] * beta[1] - X[, 2] * beta[2]
x = X[, 3]
xpxi = 1/(t(x) %*% x)
betaHat = t(x) %*% w * xpxi
beta[3] = rnorm(1, betaHat, sqrt(xpxi)) # using residual var = 1

# output current sample
cat("iteration: ", iter,
    " -- intercept: ", beta[1],
    " -- herd1: ", beta[2],
    " -- herd2: ", beta[3], "\n")
}
```

```
## iteration:  1  -- intercept:  1.213823  -- herd1:  1.314682  -- herd2:  1.654015
## iteration:  2  -- intercept:  0.5292787  -- herd1:  2.543715  -- herd2:  2.092518
## iteration:  3  -- intercept:  -1.165233  -- herd1:  2.762176  -- herd2:  2.209852
## iteration:  4  -- intercept:  -1.592898  -- herd1:  3.983947  -- herd2:  1.334666
## iteration:  5  -- intercept:  -1.552891  -- herd1:  3.158203  -- herd2:  -0.4103593
## iteration:  6  -- intercept:  -0.8488247  -- herd1:  2.824447  -- herd2:  2.819446
## iteration:  7  -- intercept:  -1.299396  -- herd1:  2.097807  -- herd2:  -0.4249774
## iteration:  8  -- intercept:  0.2625197  -- herd1:  1.155759  -- herd2:  1.68273
## iteration:  9  -- intercept:  0.2280009  -- herd1:  0.5810335  -- herd2:  1.652214
## iteration:  10  -- intercept:  1.176788  -- herd1:  0.8273104  -- herd2:  3.324785
```

Solution

The solution is to compute the average of the samples which then corresponds to the estimates.

```
### # Matrix X as incidence matrix for beta0 and beta1
X <- cbind(1,model_matrix_bhb)
### # y as vector of observations
y <- tbl_herd_bhb_gen_data$bBHB
### # starting values
beta = c(0, 0, 0)
### # vector for to store mean
mean_beta = c(0, 0, 0)
# loop for Gibbs sampler
niter = 10 # number of samples
for (iter in 1:niter) {
# sampling intercept
w = y - X[, 2] * beta[2] - X[, 3] * beta[3]
x = X[, 1]
xpxi = 1/(t(x) %*% x)
betaHat = t(x) %*% w * xpxi
beta[1] = rnorm(1, betaHat, sqrt(xpxi)) # using residual var = 1
# sampling herd 1
w = y - X[, 1] * beta[1] - X[, 3] * beta[3]
x = X[, 2]
xpxi = 1/(t(x) %*% x)
betaHat = t(x) %*% w * xpxi
beta[2] = rnorm(1, betaHat, sqrt(xpxi)) # using residual var = 1
# sampling herd 2
w = y - - X[, 1] * beta[1] - X[, 2] * beta[2]
x = X[, 3]
xpxi = 1/(t(x) %*% x)
betaHat = t(x) %*% w * xpxi
beta[3] = rnorm(1, betaHat, sqrt(xpxi)) # using residual var = 1

# update the average
mean_beta <- mean_beta + beta / niter
# output current sample
cat("iteration: ", iter,
    " -- intercept: ", beta[1],
    " -- herd1: ", beta[2],
    " -- herd2: ", beta[3], "\n")
}
```

```
## iteration:  1  -- intercept:  1.399134  -- herd1:  0.09158261  -- herd2:  1.802606
## iteration:  2  -- intercept:  1.054582  -- herd1:  0.6940887  -- herd2:  4.358546
## iteration:  3  -- intercept:  0.4910536  -- herd1:  0.8241545  -- herd2:  2.444604
## iteration:  4  -- intercept:  -0.1960709  -- herd1:  2.304435  -- herd2:  2.374846
## iteration:  5  -- intercept:  0.3390206  -- herd1:  1.501638  -- herd2:  1.462302
## iteration:  6  -- intercept:  0.1416927  -- herd1:  1.169535  -- herd2:  0.9136941
## iteration:  7  -- intercept:  0.5784199  -- herd1:  0.5054548  -- herd2:  2.344605
## iteration:  8  -- intercept:  0.8127417  -- herd1:  0.7585087  -- herd2:  2.214807
## iteration:  9  -- intercept:  0.1533818  -- herd1:  0.8072271  -- herd2:  2.600656
## iteration:  10  -- intercept:  0.3560969  -- herd1:  1.725096  -- herd2:  0.7098091
```

```
cat("\n")
```

```
cat(" * Estimates: intercept: ", mean_beta[1],
    " -- herd1: ", mean_beta[2],
    " -- herd2: ", mean_beta[3], "\n")
```

```
##  * Estimates: intercept:  0.5130052  -- herd1:  1.038172  -- herd2:  2.122648
```

c) How can the effect of the herds be quantified using the least squares method instead of a Bayesian approach? Please specify the command in R that you would use to get to the least squares estimates of the herd-effects. The data to be analysed is available in a dataframe called `tbl_herd_bhb_gen_data` with a column named 'bBHB' for the response variable and a column 'Herd' with the indicators of the herd.

Wie können die Betriebseffekte mit der Methode der kleinsten Quadrate anstelle eines Bayes'schen Ansatztes geschätzt werden? Bitte geben sie den R-Befehl, welcher die Betriebseffekte schätzt. Die zu analysierenden Daten sind in einem Datenframe namens `tbl_herd_bhb_gen_data` abgelegt. Der Datenframe enthält eine Kolonne 'bBHB' für die Zielgrösse und eine Kolonnen namens 'Herd' für die Betriebsinformationen.

3

Solution

Least squares estimates of herd effects are obtained by `lm()`.

```
lm_bhb_herd <- lm(bBHB ~ Herd, data = tbl_herd_bhb_gen_data)
summary(lm_bhb_herd)
```

```
##
## Call:
## lm(formula = bBHB ~ Herd, data = tbl_herd_bhb_gen_data)
##
## Residuals:
##          1          2          3          4
## -3.433e-02 -3.469e-18 -3.333e-03  3.767e-02
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.601333   0.020851   76.80  0.00017 ***
## Herd2       -0.003333   0.041703   -0.08  0.94357
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.03612 on 2 degrees of freedom
## Multiple R-squared:  0.003184,   Adjusted R-squared:  -0.4952
## F-statistic: 0.006389 on 1 and 2 DF,  p-value: 0.9436
```