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 2019

Date: 27th May 2019

Name:

Legi-Nr:

Problem	Maximum Number of Points	Number of Points Reached
1	34	
2	15	
3	25	
4	18	
Total	92	

Questions in German are in italics

Problem 1: Genomic BLUP

We are given the following data set for the trait Average Daily Gain (`ADG`).

Gegeben ist der folgende Datensatz für das Merkmal Tageszunahme (‘ADG‘).

Animal	ADG	Sire	Dam	SNP1	SNP2	SNP3	SNP4
1	1285	NA	NA	1	1	0	0
2	943	NA	NA	0	-1	-1	0
3	1051	1	2	1	0	0	1

a) Use a linear mixed effects model that is based on genomic breeding values (GBLUP) for the given data set. A general mean μ is the only fixed effect that is considered in the model. Specify all model components including the expected values and the variances of the random effects. Input the numeric information from the dataset into the model.

Verwenden Sie ein lineares gemischtes Modell basierend auf genomischen Zuchtwerten (GBLUP) für den gegebenen Datensatz. Als einzigen fixen Effekt nehmen wir ein allgemeines Mittel μ an. Spezifizieren Sie alle Modellkomponenten, inklusive der Erwartungswerte und der Varianzen der zufälligen Effekte.

19

Solution

$$y = Xb + Zg + e$$

where

y vector of length n with observations
b general mean as the only fixed effect
X incidence matrix linking elements in b to observations
g vector of length t with random genomic breeding values
Z incidence matrix linking elements in g to observations
e vector of length n of random error terms

Inserting the data into the model

$$y = \left[\begin{array}{r} 1285 \\ 943 \\ 1051 \end{array}\right], b = \left[\begin{array}{c} \mu \end{array}\right], X = \left[\begin{array}{c} 1 \\ 1 \\ 1 \end{array}\right], g = \left[\begin{array}{c} g_1 \\ g_2 \\ g_3 \end{array}\right], Z = \left[\begin{array}{ccc} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{array}\right], e = \left[\begin{array}{c} e_1 \\ e_2 \\ e_3 \end{array}\right]$$

The expected values of the random components are: $E\left[g\right] = 0$, $E\left[e\right] = 0$ and $E\left[y\right] = Xb = 1_n\mu$

The variances are $var(g) = G * \sigma_g^2$ where G is the genomic relationship matrix, $var(e) = R = I * \sigma_e^2$ and $var(y) = ZGZ^T + R$.

b) Compute the genomic relationship matrix G from the SNP marker data given in the above table.

Berechnen Sie die genomische Verwandtschaftsmatrix G basierend auf den SNP-Markerdaten, welche im Datensatz gegeben sind.

9

Solution

The genomic relationship matrix is computed based on the incidence matrix W in a marker effect model.

$$W = \left[\begin{array}{rrrr} 1 & 1 & 0 & 0 \\ 0 & -1 & -1 & 0 \\ 1 & 0 & 0 & 1 \end{array}\right],$$

The following function is used to compute the genomic relationship matrix G

```r
#' Compute genomic relationship matrix based on data matrix
computeMatGrm <- function(pmatData, pnfreq = NULL) {
  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
  if (is.null(pnfreq)){
    freq <- apply(matData, 2, mean) / 2
  } else {
    freq <- rep(pnfreq, ncol(matData))
  }
  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_G_agd <- computeMatGrm(pmatData = mat_W_agd)
```

From that the genomic relationship matrix is

$$G = \left[\begin{array}{rrr} 0.800 & -0.800 & 0.000 \\ -0.800 & 1.200 & -0.400 \\ 0.000 & -0.400 & 0.400 \end{array}\right],$$

c) Sire 1 and dam 2 have also animals 4 and 5 as offspring. Animals 4 and 5 do not have phenotypic observations. Is it possible to determine a ranking based on the predicted breeding values of the three full-sibs 3, 4 and 5 using a traditional BLUP animal model (without SNP-Information) or is it possible with genomic breeding values or with both or with none? Please explain your answer.

   Vater 1 und Mutter 2 haben auch Tiere 4 und 5 als Nachkommen. Die Tiere 4 und 5 haben keine phänotypischen Beobachtungen. Ist es möglich eine Rangierung der Vollgeschwister 3, 4 und 5 aufgrund der geschätzten Zuchtwerte zu erstellen, falls die Zuchtwerte mit einem traditionellen BLUP-Tiermodell (ohne SNP-Informationen) geschätzt wurden oder, wenn genomische Zuchtwerte verwendet werden oder mit beiden Methoden oder mit keinen der beiden Methoden? Bitte begründen Sie Ihre Antwort.

6

Solution

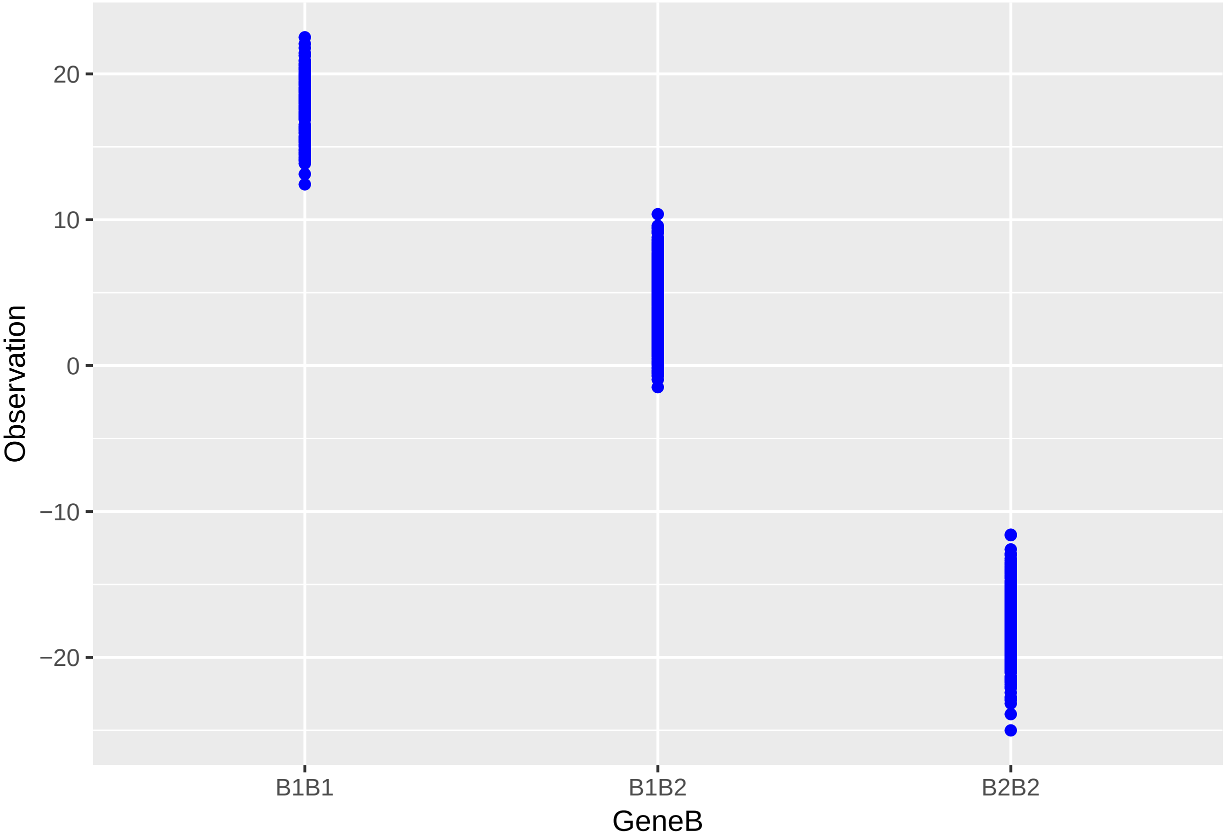
With traditional BLUP animal model breeding values, animals 4 and 5 get the same predicted breeding values. The correspond to the mean of the predicted breeding values of parents 1 and 2. Hence it is not possible to determine which of 4 and 5 is better based on traditional BLUP animal model breeding values. Because animal 3 has an own-performance record, its predicted breeding value will be different.

With genomic information the breeding values of 3, 4 and 5 will be different, because of the genomic information.

Problem 2: Single Gene Trait

We assume that a given trait is determined by just one genetic locus called B with two alleles $B1$ and $B2$. The frequency of the $B1$ allele is 0.35. The following figure shows a plot of the observed values of the trait against the three genotypes $B1B1$, $B1B2$ and $B2B2$.

Ein Merkmal wird von einem Genort namens B mit zwei Allelen $B1$ und $B2$ bestimmt. Die Frequenz des $B1$-Allels entspricht 0.35. Das folgende Diagramm zeigt die beobachteten Werte des Merkmals gegen die drei Genotypen $B1B1$, $B1B2$ und $B2B2$ geplottet.



When fitting a linear model with the observed values as response and the genotypes as the predictor, the following is the result of the `summary()` function.

Ein lineares Model der beobachteten Werte als Zielgrösse und die Genotypen als unabhängige Variablen führt zum folgenden Ergebnis.

```
##
## Call:
## lm(formula = Observation ~ 0 + GeneB, data = tbl_all_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.2718 -1.4581  0.0385  1.4880  6.1449
##
## Coefficients:
##            Estimate Std. Error t value Pr(>|t|)
## GeneBB1B1  17.87047    0.19035   93.88   <2e-16 ***
## GeneBB1B2   4.33810    0.09869   43.96   <2e-16 ***
## GeneBB2B2 -17.73346    0.10510 -168.73   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.128 on 997 degrees of freedom
## Multiple R-squared:  0.9752, Adjusted R-squared:  0.9751
## F-statistic: 1.307e+04 on 3 and 997 DF,  p-value: < 2.2e-16
```

a) What are the values for a and d in the mono-genic model.

Bestimmen Sie die Werte a und d aus dem Ein-Gen Modell.

2

Solution

$$a = 17.8$$

$$d = 4.27$$

b) Compute the genotypic values for the three genotypes $B1B1$, $B1B2$ and $B2B2$.

Berechnen Sie die genotypischen Werte der drei Genotypen $B1B1$, $B1B2$ and $B2B2$.

3

Solution

Genotype	Genotypic Value
B1B1	17.8
B1B2	4.27
B2B2	-17.8

c) Compute the population mean μ and the breeding values BV_{11}, BV_{12} and BV_{22} for the three genotypes $B1B1$, $B1B2$ and $B2B2$. (Use $a = 10$ and $d = 0$, if you could not solve Problem 2a).

Berechnen Sie das Populationsmittel μ und die Zuchtwerte BV_{11}, BV_{12} und BV_{22} für die drei Genotypen $B1B1$, $B1B2$ und $B2B2$. (Verwenden Sie $a = 10$ und $d = 0$, falls Sie Aufgabe 2a nicht lösen konnten.)

10

Solution

The population mean

$$\mu = (p - q) * a + 2pqd = (0.35 - 0.65) * 17.8 + 2 * 0.35 * 0.65 * 4.27 = -3.4$$

The substitution effect α corresponds to

$$\alpha = a + (q - p) * d = 17.8 + (0.65 - 0.35) * 4.27 = 19.08$$

$$BV_{11} = 2q\alpha = 2 * 0.65 * 19.08 = 24.81$$
$$BV_{12} = (q - p)\alpha = (0.65 - 0.35) * 19.08 = 5.72$$
$$BV_{22} = -2p\alpha = -2 * 0.35 * 19.08 = -13.36$$

Problem 3: Fixed Linear Effects Model

The following table contains body weight and slaughter weight for 12 animals. Before the farmer sells the animal to the slaughter house, it is weighed on the farm. The slaughter weight is determined by the slaughter house. The following regression of slaughter weight on body weight gives the association between the two traits.

Die folgende Tabelle enthält Lebendgewicht (‘BodyWeight‘) und Schlachtgewicht (‘SlaughterWeight‘) für 12 Tiere. Vor der Schlachtung wird das Tier auf dem Betrieb noch gewogen. Das Schlachtgewicht wird im Schlachthof bestimmt. Die folgende Regression des Schlachtgewichts auf das Körpergewicht zeigt den Zusammenhang zwischen den beiden Merkmalen.

Animal	BodyWeight	SlaughterWeight
1	517	269
2	510	276
3	487	251
4	503	258
5	489	258
6	503	263
7	511	264
8	500	255
9	513	269
10	491	258
11	504	262
12	516	270

The result of the linear regression model are as follows.

Das Resultat der linearen Regression lauten wie folgt.

```
##
## Call:
## lm(formula = SlaughterWeight ~ BodyWeight, data = tbl_sw_bw)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.6315 -2.3371 -0.4092  1.2852  9.5908
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -28.2542    62.3902  -0.453 0.660314
## BodyWeight    0.5778     0.1238   4.665 0.000887 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.252 on 10 degrees of freedom
## Multiple R-squared:  0.6852, Adjusted R-squared:  0.6537
## F-statistic: 21.76 on 1 and 10 DF,  p-value: 0.0008873
```

a) How much does the slaughter weight change, if the body weight of the animal is changed by +1 kg?

Wie stark verändert sich das Schlachtgewicht, wenn das Körpergewicht eines Tieres um +1 kg zunimmt?

1

Solution

The change in slaughter weight is given by the regression coefficient which corresponds to: 0.578

b) Compute the residual standard deviation from the fitted values of the model. With which quantity of the R-output above can the result of your computation be verified?

Berechnen Sie die Standardabweichung der Residuen aufgrund der gefitteten Werte aus dem Modell. Mit welcher Grösse aus dem R-output in der Aufgabenstellung können Sie Ihre Berechnung verifizieren?

14

Solution

The standard deviation of the residuals is computed from the residuals r_i computed as $r_i = y_i - \hat{y}_i$. For this, we extend the table of the input data with the fitted values and the residuals.

Animal	BodyWeight	SlaughterWeight	FittedValues	Residuals
1	517	269	270.5	-1.454
2	510	276	266.4	9.591
3	487	251	253.1	-2.120
4	503	258	262.4	-4.365
5	489	258	254.3	3.724
6	503	263	262.4	0.635
7	511	264	267.0	-2.987
8	500	255	260.6	-5.632
9	513	269	268.1	0.857
10	491	258	255.4	2.568
11	504	262	262.9	-0.943
12	516	270	269.9	0.124

The residual standard deviation corresponds to

$$s_r^2 = \frac{1}{n-p}\sum_{i=1}^{n} r_i^2 = \frac{1}{10}\sum_{i=1}^{12} -1.4536199^2 + \ldots + 0.1241516^2 = 18.08$$

The square root of s_r^2 corresponds to the value given in the R-output named residual standard error. With that we get

$$s_r = \sqrt{18.08} = 4.25$$

c) The farmer wants to sell his animals whenever an animal is more than 500 kg. What is the expected slaughter weight that he can expect from an animal with a body weight of 500 kg?

Der Bauer möchte seine Tiere bei 500 *kg schlachten lassen. Welches Schlachtgewicht kann er bei einem Tier mit* 500 *kg Lebendgewicht erwarten?*

4

Solution

The expected slaughter weight $\widehat{sw}$ of an animal with body weight $bw = 500$ kg is computed as

$$\widehat{sw} = \hat{b}_0 + \hat{b}_{bw} * bw = -28.25 + 0.58 * 500 = 260.6$$

d) The farmer would like to use the same regression to predict the slaughter weight for his fattening calf with a body weight of 180 kg. What do you think about the predicted slaughter weight for the fattening calf? Give one reason for your answer. Compute the predicted slaughter weight for the calf based on the regression given above.

Der Bauer möchte die Regression auch für ein Mastkalb mit einem Körpergewicht von 180 *kg verwenden um das Schlachtgewicht zu schätzen. Ist das Ihrer Meinung nach eine gute Idee. Begründen Sie Ihre Antwort. Wie gross ist das geschätzte Schlachtgewicht für das Mastkalb?*

6

Solution

It is not a good idea, because this prediction is based on a body weight outside of the range of the body weights in the dataset. Hence this is extrapolation which is considered to be negative. The predicted slaughter weight would be

$$\widehat{sw}_{calf} = \hat{b}_0 + \hat{b}_{bw} * bw_{calf} = -28.25 + 0.58 * 180 = 75.7$$

Problem 4: Bayes

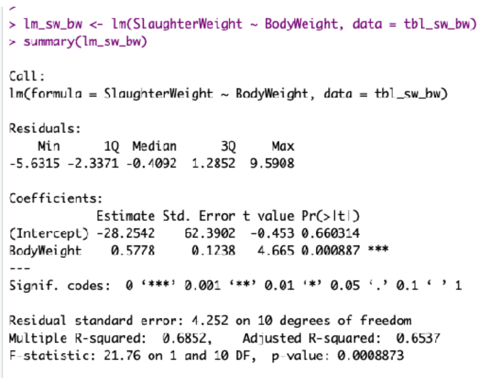
We use the same dataset body weight and slaughter weight as in Problem 3.

Wir verwenden den gleichen Datensatz zu Lebendgewicht und Schlachtgewicht, wie in Aufgabe 3.

Animal	BodyWeight	SlaughterWeight
1	517	269
2	510	276
3	487	251
4	503	258
5	489	258
6	503	263
7	511	264
8	500	255
9	513	269
10	491	258
11	504	262
12	516	270

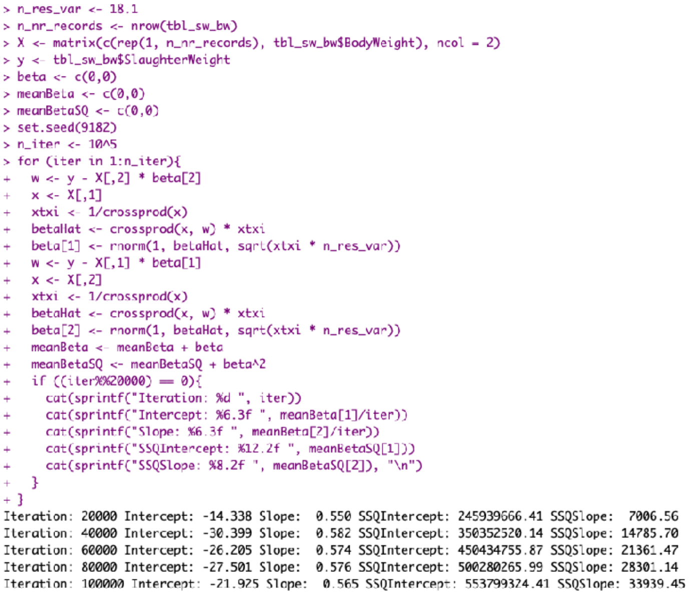
A fixed linear model is fit to the data using the two codeblocks 1 and 2 shown below.

Codeblock 1



```
> lm_sw_bw <- lm(SlaughterWeight ~ BodyWeight, data = tbl_sw_bw)
> summary(lm_sw_bw)

Call:
lm(formula = SlaughterWeight ~ BodyWeight, data = tbl_sw_bw)

Residuals:
    Min      1Q  Median      3Q     Max 
-5.6315 -2.3371 -0.4092  1.2852  9.5908 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -28.2542    62.3902  -0.453 0.660314    
BodyWeight    0.5778     0.1238   4.665 0.000887 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.252 on 10 degrees of freedom
Multiple R-squared:  0.6852,	Adjusted R-squared:  0.6537 
F-statistic: 21.76 on 1 and 10 DF,  p-value: 0.0008873
```

Codeblock 2



```
> n_res_var <- 18.1
> n_nr_records <- nrow(tbl_sw_bw)
> X <- matrix(c(rep(1, n_nr_records), tbl_sw_bw$BodyWeight), ncol = 2)
> y <- tbl_sw_bw$SlaughterWeight
> beta <- c(0,0)
> meanBeta <- c(0,0)
> meanBetaSQ <- c(0,0)
> set.seed(9182)
> n_iter <- 10^5
> for (iter in 1:n_iter){
+   w <- y - X[,2] * beta[2]
+   x <- X[,1]
+   xtxi <- 1/crossprod(x)
+   betaHat <- crossprod(x, w) * xtxi
+   beta[1] <- rnorm(1, betaHat, sqrt(xtxi * n_res_var))
+   w <- y - X[,1] * beta[1]
+   x <- X[,2]
+   xtxi <- 1/crossprod(x)
+   betaHat <- crossprod(x, w) * xtxi
+   beta[2] <- rnorm(1, betaHat, sqrt(xtxi * n_res_var))
+   meanBeta <- meanBeta + beta
+   meanBetaSQ <- meanBetaSQ + beta^2
+   if ((iter%%20000) == 0){
+     cat(sprintf("Iteration: %d ", iter))
+     cat(sprintf("Intercept: %6.3f ", meanBeta[1]/iter))
+     cat(sprintf("Slope: %6.3f ", meanBeta[2]/iter))
+     cat(sprintf("SSQIntercept: %12.2f ", meanBetaSQ[1]))
+     cat(sprintf("SSQSlope: %8.2f ", meanBetaSQ[2]), "\n")
+   }
+ }
Iteration: 20000 Intercept: -14.338 Slope:  0.550 SSQIntercept: 245939666.41 SSQSlope:  7006.56 
Iteration: 40000 Intercept: -30.399 Slope:  0.582 SSQIntercept: 350352520.14 SSQSlope: 14785.70 
Iteration: 60000 Intercept: -26.205 Slope:  0.574 SSQIntercept: 450434755.87 SSQSlope: 21361.47 
Iteration: 80000 Intercept: -27.501 Slope:  0.576 SSQIntercept: 500280265.99 SSQSlope: 28301.14 
Iteration: 100000 Intercept: -21.925 Slope:  0.565 SSQIntercept: 553799324.41 SSQSlope: 33939.45 
```

a) Which of the two codeblocks shows a Bayesian approach and which one does the model fit using Least Squares?

Welcher der Codeblocks zeigt einen Bayes'schen Ansatz und welcher verwendet Least Squares zur Anpassung des fixen linearen Modells?

2

Solution

Codeblock 1 uses Least Squares and Codeblock 2 uses a Bayesian approach.

b) What are the estimate for the intercept and the regression coefficient for both approaches?

Wie gross sind die Schätzwerte für Achsenabschnitt und Regressionskoeffizient bei beiden Methoden?

8

Solution

Parameter	Least Squares	Bayesian
Intercept	-28.2542	-21.925
Regression Coefficient	0.5778	0.565

c) What are the standard errors of intercept and slope under the two approaches?

Wie gross sind die Schätzfehler des Achsenabschnitts und des Regressionskoeffizienten unter beiden Methoden?

8

Solution

Parameter	Least Squares	Bayesian
Intercept	62.3902	71.1149
Regression Coefficient	0.1238	0.1412