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

751-7602-00 V
Exam in
Applied Statistical Methods
in Animal Sciences
SS 2021

Date: 31^{st} May 2021

Name:

Legi-Nr:

Problem	Maximum Number of Points	Number of Points Reached
1	16	
2	28	
3	9	
4	38	
Total	91	

Questions in German are in italics

Problem 1: Linear Regression

The same dataset is analysed with two different regression models. The R-Output of both analyses is given by Output A and Output B.

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) Give the formulas of both statistical models which belong to Output A and Output B. Where is the main difference between both models? *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

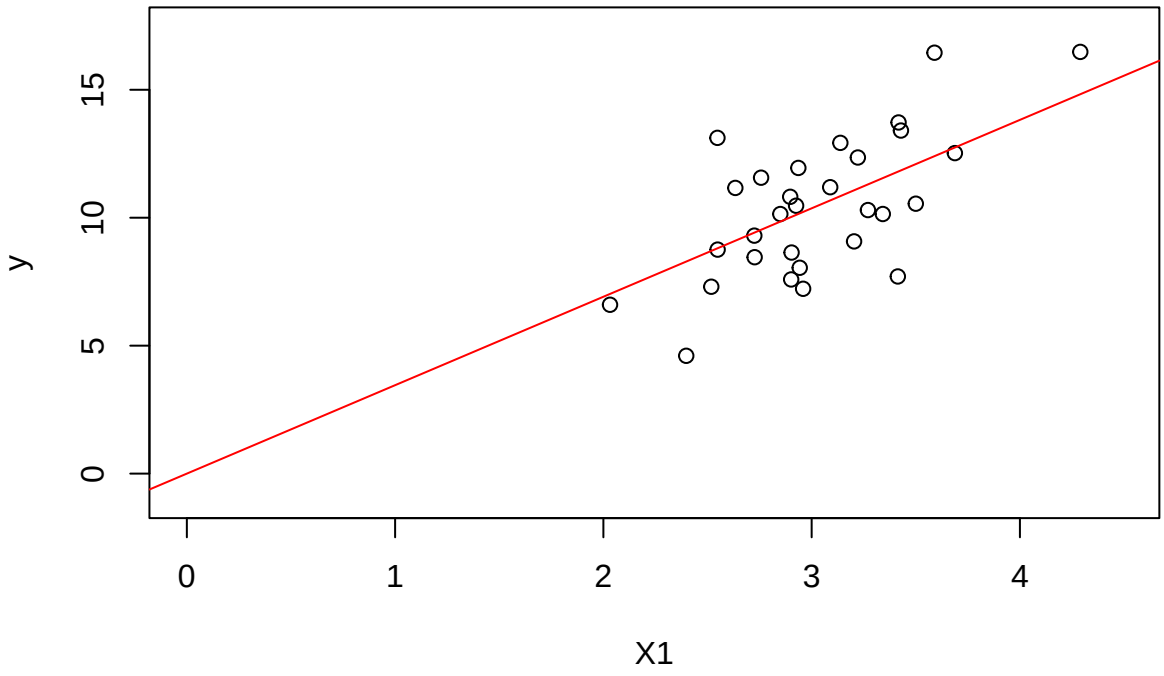
Solution

b) For both analyses a plot was produced. Assign Plots 1 and 2 to Outputs A and B.

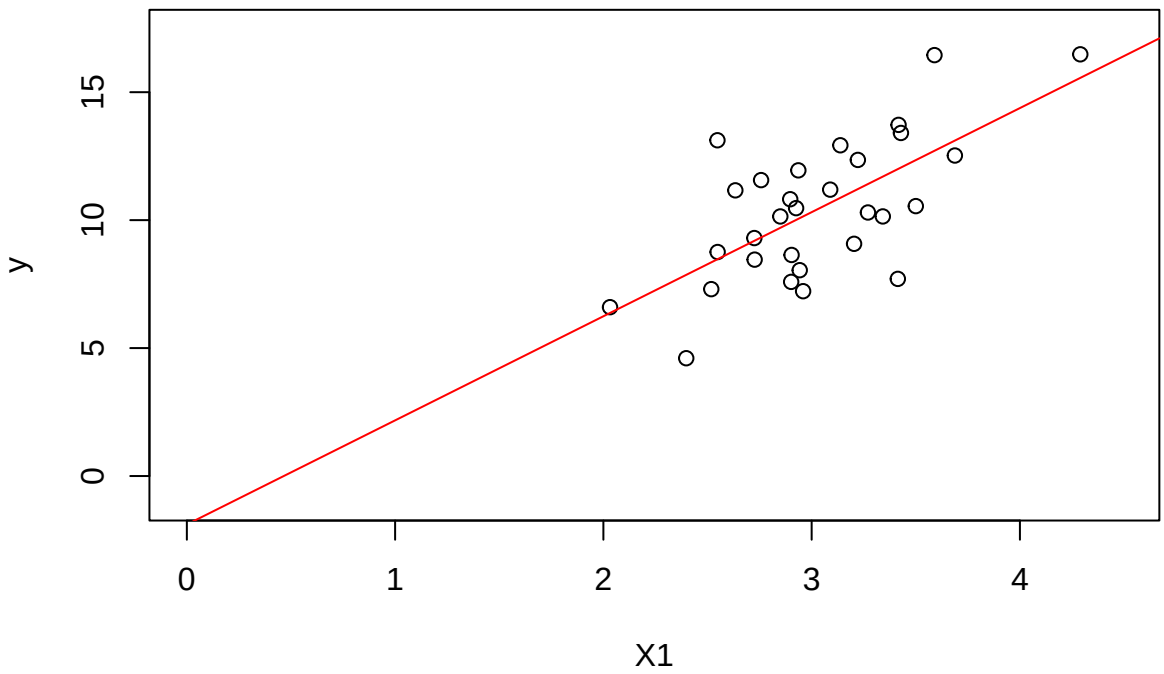
Für die zwei Analysen wurden auch zwei Plots gemacht. Ordnen Sie die Plots 1 und 2 den Outputs A und B zu.

2

Plot 1



Plot 2



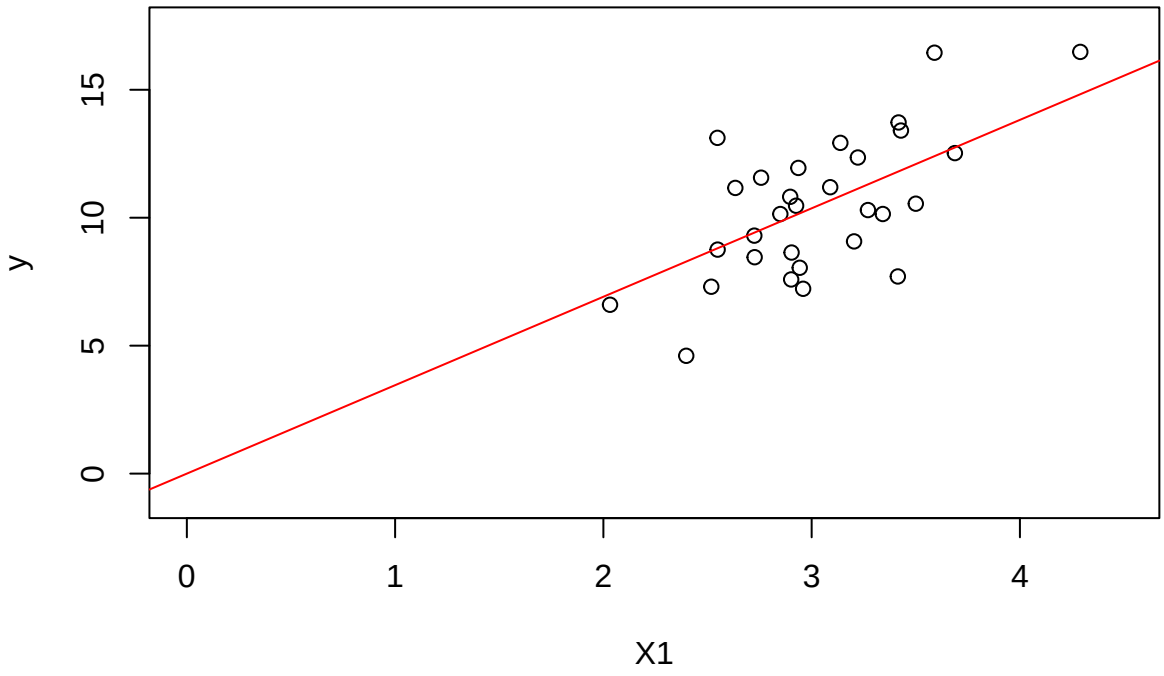
Solution

c) Enter the parameter estimates from Outputs A and B in Plots 1 and 2 by marking their lengths in the plots.

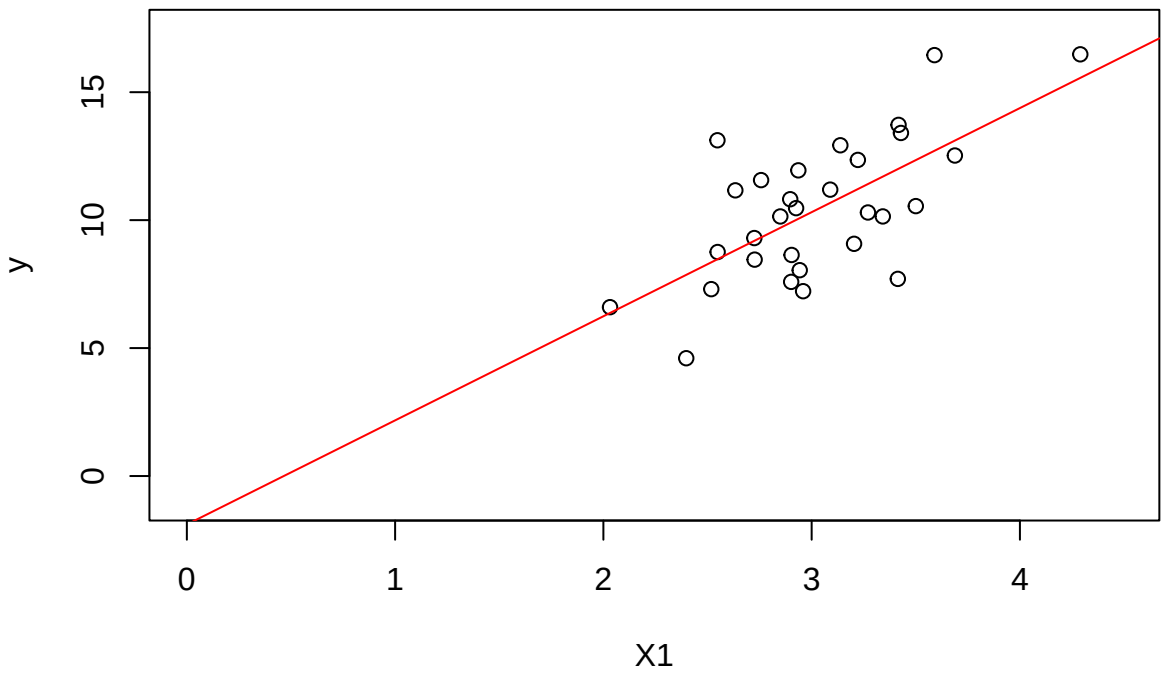
Zeichnen Sie die geschätzten Parameter (Estimate) aus den Outputs A und B in die Plots 1 und 2 ein

6

Plot 1



Plot 2



Solution

Problem 2: Bayes

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.

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.

Animal	BodyWeight	SlaughterWeight
1	522	200
2	516	199
3	523	205
4	540	224
5	530	209
6	549	213
7	543	209
8	547	219
9	549	220
10	524	204
11	535	206
12	540	209

a) Please specify the equation that models 'SlaughterWeight' (response variable) as a regression on 'BodyWeight' (predictor variable). Based on the specified regression equation and based on the dataset, create a table with knowns and unknowns.

Bitte geben Sie eine Modellgleichung, welche 'SlaughterWeight' (Zielgrösse) als Regression auf 'BodyWeight' (Predictorvariable) modelliert. Basierend auf der spezifizierten Regressionsgleichung und basierend auf dem Datensatz, geben Sie in einer Tabelle an, welche Grössen bekannt und welche unbekannt sind.

7

Solution

- Regression Model:
- Table of knowns and unknowns

b) The following programming code in R does a Bayesian estimation of the unknowns of the regression model. Please complete the code where indicated (lines after comment starting with "TODO") such that estimates of unknowns are obtained.

Der nachfolgende Programmcode in R ergibt eine Bayes'sche Schätzung der unbekannten im Regressionsmodell. Bitte vervollständigen Sie den nachfolgenden Programmcode so, dass die Schätzungen der Unbekannten Grössen im Regressionsmodell resultieren. Die zu ergänzenden Stellen sind mit einem Kommentar, welcher mit dem Wort "TODO" beginnt, markiert.

15

Hint

- The dataset is available at: https://charlotte-ngs.github.io/gelasmss2021/data/asm_exam_p02.csv.

Solution

```
01 # read the data
02 s_data_p02_path <- "https://charlotte-ngs.github.io/gelasmss2021/data/asm_exam_p02.csv"
03 tbl_reg_sw_bw <- readr::read_csv2(file = s_data_p02_path)
04
05 # take number of observations from tbl_reg_sw_bw
06 n_nr_obs <- nrow(tbl_reg_sw_bw)
07
08 # define Matrix X
09 X <- matrix(c(rep(1,n_nr_obs), tbl_reg_sw_bw$BodyWeight), ncol = 2)
10 # observations as vector y
11 y <- tbl_reg_sw_bw$SlaughterWeight
12 # fix constants
13 nuRes <- 4
14 varResidual <- 1
15 scaleRes <- varResidual * (nuRes - 2)/nuRes
16 mu <- mean(y)
17 ycorr <- y - mu
18 # intialise estimates for intercept, slope and residual variance
19 beta <- c(0,0)
20 meanBeta <- c(0, 0)
21 sigma <- 1
22 meanSigma <- 0
23 # loop over iterations of the Gibbs Sampler
24 niter <- 1000
25 for (iter in 1:niter){
26   # sampling intercept
27   w  <-  y - X[, 2] * beta[2]
28   x <- X[, 1]
29   xpxi <- 1/(t(x) %*% x)
30   # TODO: compute mean of conditional distribution
31   betaHat <-
32   # TODO: draw sample of intercept from normal distribution
33   beta[1] <-
34   # sampling slope
35   w <- y - X[, 1] * beta[1]
36   x <- X[, 2]
37   xpxi <- 1/(t(x) %*% x)
38   # TODO: compute mean of conditional distribution
```

```
39   betaHat <-
40   # TODO: draw sample for slope from normal distribution
41   beta[2] <-
42   # sample residual variance
43   sigma <- (crossprod(ycorr) + nuRes * scaleRes) / rchisq(1, n_nr_obs + nuRes)
44  # TODO: save current sample of beta to meanBeta and sigma to meanSigma
45   meanBeta <-
46   meanSigma <-
47   # output every 200th sample
48   if (iter %% 200 == 0){
49     cat(" * Iteration:         ", iter, "\n")
50     cat(" * Intercept:         ", beta[1], "\n")
51     cat(" * Slope:             ", beta[2], "\n")
52     cat(" * Residual Variance: ", sigma, "\n")
53   }
54 }
55
56 # output estimates
57 cat(" * Bayes Estimates\n")
58 cat(" * Intercept:         ", meanBeta[1]/niter, "\n")
59 cat(" * Slope:             ", meanBeta[2]/niter, "\n")
60 cat(" * Residual Variance: ", meanSigma/niter, "\n")
```

c) We assume that in addition to the 12 animals shown above there is an additional animal with a body weight of 513 kg. In the slaughterhouse the slaughter weight could not be measured, hence it is missing. How are such missing observations handled in a Bayesian analysis? Please fill out the table with the knowns and unknowns once again taking into account the fact that the observation of the slaughterweight for animal 13 is missing.

Wir nehmen an, dass zusätzlich zu den 12 Tieren, welche in der Tabelle oben gezeigt wurden, noch ein zusätzliches Tier mit einem Lebendgewicht von 513 kg hinzukommt. Im Schlachthof konnte das Schlachtgewicht vom Tier 13 nicht erfasst werden und fehlt somit. Wie wird diese fehlende Beobachtung in einer Bayes'schen Analyse behandelt? Bitte ergänzen Sie die Tabelle mit den bekannten und den unbekannten Grössen unter Berücksichtigung der fehlenden Beobachtung des Schlachtgewichts von Tier 13.

6

Solution

Expanded Table with knowns and unknowns

Problem 3: LASSO

a) LASSO is an alternative procedure to Least Squares to estimate parameters of a linear model. Which of the following equations belongs to least square and which belongs to LASSO.

LASSO ist ein alternatives Parameterschätzverfahren zu Least Squares. Ordnen Sie die nachfolgenden Gleichungen zu den beiden Verfahren Least Squares und LASSO zu.

4

$$\hat{\beta}_1 \quad = \quad argmin_\beta \left\{ ||y - X\beta||^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\}$$

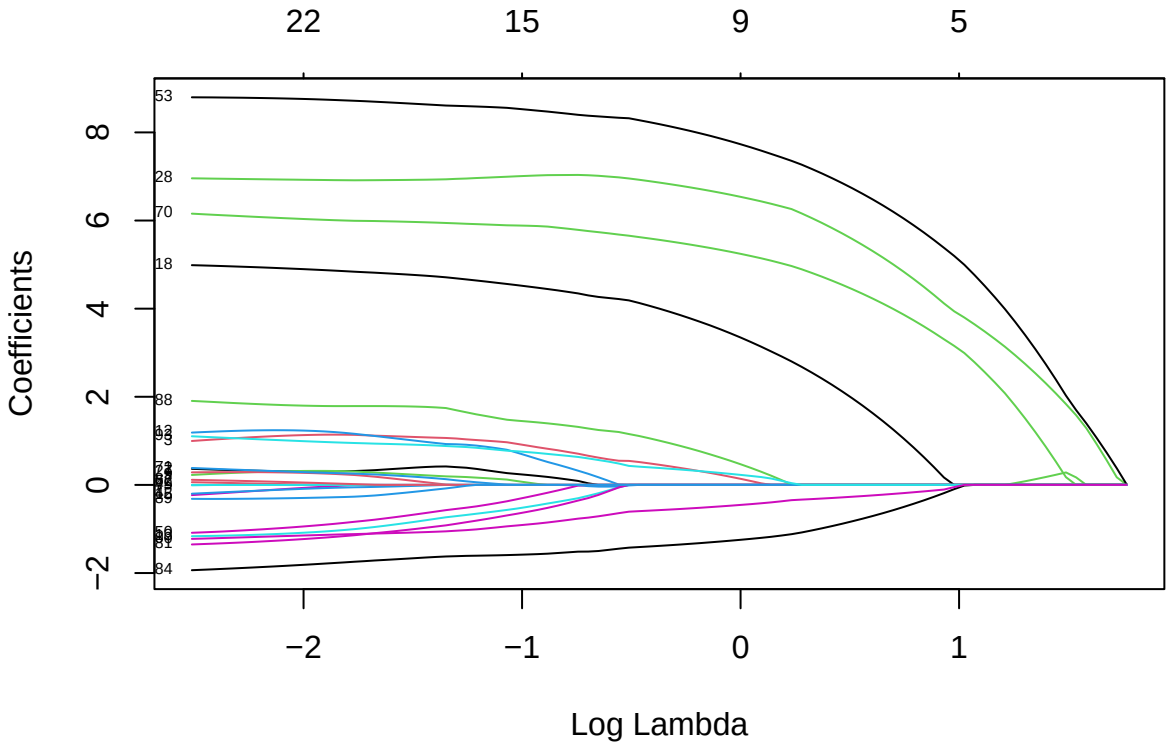
$$\hat{\beta}_2 \quad = \quad argmin_\beta ||y - X\beta||^2$$

Solution

b) We analyse a genomic dataset with 25 animals which are genotyped at 100 SNP-locations. Only 5 SNPs have an effect on the observed trait. The SNP-effects are estimated with LASSO. The results of this analysis are shown in the two plots below. The second plot can be used to determine the penalty-term (Log Lambda), such that a minimum number of SNP-effects are considered and such that the mean-squared error is still as small as possible (right dotted). Which are the 5 SNPs (indicate the numbers) with the largest absolute effects, when the penalty-term is determined based on the second plot.

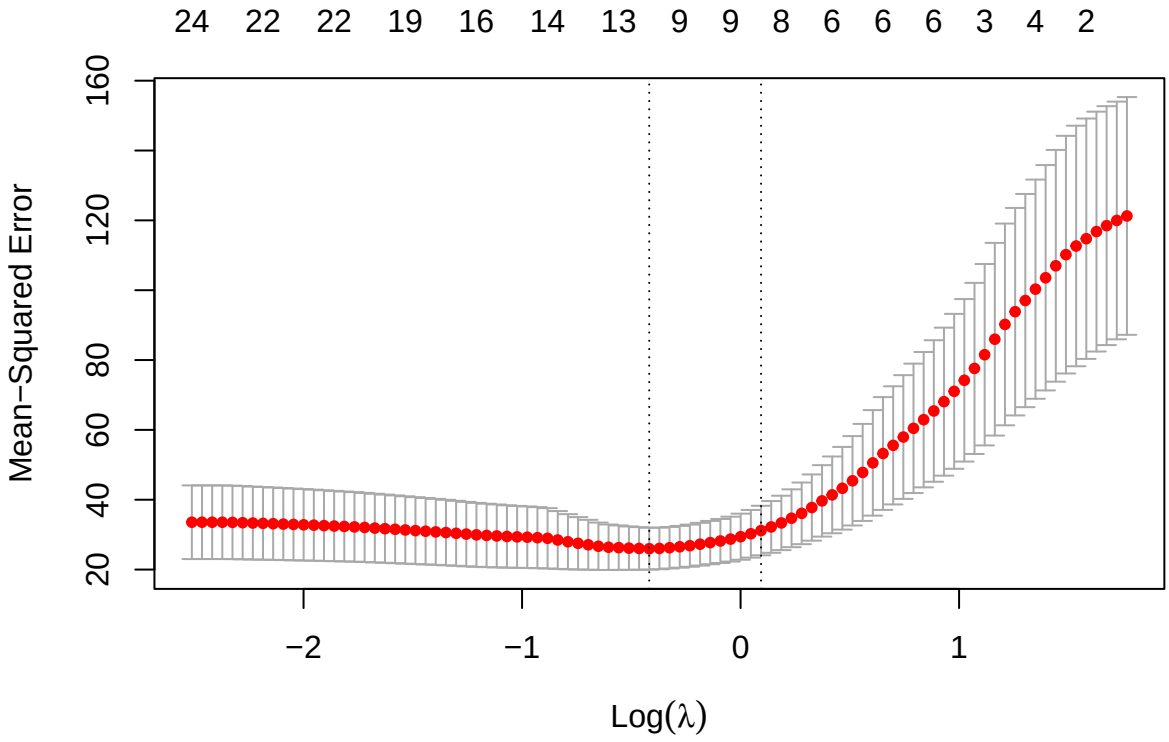
Wir analysieren einen genomischen Datensatz mit 25 Tieren, welche Daten an 100 SNP-Positionen aufweisen. Davon haben nur 5 SNP einen Effekt auf das gemessene Merkmal. Die SNP-Effekte werden mit LASSO geschätzt. Die Resultate sind in den nachfolgenden Plots gezeigt. Im zweiten Plot können wir den Strafterm (Log Lambda) so bestimmen, dass möglichst wenige SNPs berücksichtigt werden und dass gleichzeitig der mittlere quadrierte Fehler minimal bleibt (rechte gestrichelte Linie). Welches sind die 5 SNPs (bitte Nummern angeben) mit den grössten absoluten Effekten, wenn wir den Strafterm aufgrund des zweiten Plots bestimmen.

5



Der Strafterm kann aufgrund der linken gestrichelten Linie bestimmt werden.

The penalty-term can be determined based on the right dotted line.



Solution

Problem 4: Genomic BLUP

```
## [#.......] Reading VCF file..
## [##......] Chromosome:   4  Mbp:  77.35628  Region Size:  347.154 kb  Num of individuals: 95
## [##......] Before filtering  Num of variants: 567 Num of individuals: 95
## [###.....] After filtering  Num of variants: 567 Num of individuals: 95
## [#####...] Creating SIM object
## [#####...] Haplodata object created
## Downloading genetic map for chromosome  4
##  -> Downloading genetic map from: https://github.com/adimitromanolakis/geneticMap-GRCh37/raw/master/g
##  -> Saving genetic map to:  /var/folders/2v/jfsqj8zj2f122jcgy15nzfn00000gn/T//RtmpiRv29y/genetic_map_
##       -> Genetic map has 211115 entries
##    user  system elapsed
##   0.912   0.014   0.930
```

The data shown in the following table is available to predict genomic breeding values using different methods. The data is available from the URL shown below:

Die Daten in der nachfolgenden Tabelle sollen für die Schätzung von genomischen Zuchtwerten mit verschiedenen Methoden verwendet werden. Die Daten können vom folgenden URL heruntergeladen werden:

https://charlotte-ngs.github.io/gelasmss2021/data/asm_exam_p04.csv.

Animal	Sire	Dam	Sex	Observation	SNP1	SNP2	SNP3
1	NA	NA	M	34.4	0	1	0
2	NA	NA	M	47.7	1	1	0
3	NA	NA	F	28.0	0	0	1
4	NA	NA	F	25.9	0	0	0
5	1	3	M	16.4	-1	0	0
6	1	4	F	30.6	0	1	0
7	2	3	F	46.0	1	1	1
8	2	3	F	25.9	0	0	0
9	5	7	F	30.6	0	1	0
10	5	8	F	12.6	-1	0	0

a) Use the two-step approach to predict genomic breeding values. Because, the number of SNP is smaller than the number of animals, marker effects can be estimated using least squares. Please indicate the type of model and specify the all the model components used to estimated marker effects. Also, describe how the genomic breeding values are computed from the marker effects.

   Verwenden Sie die Zwei-Schritt Methode zur Schätzung der genomischen Zuchtwerte. Da die Anzahl SNP kleiner ist als die Anzahl Tiere im Datensatz können die Markereffekte mit Least Squares geschätzt werden. Bitte geben Sie den Modell-Typ an und spezifizieren Sie alle Komponenten des Modells, welches zur Schätzung der Markereffekte verwendet wird. Beschreiben Sie auch, wie Sie aus den Markereffekten die genomischen Zuchtwerte berechnen.

   8

Solution

b) Use a single-step marker-effect model to predict breeding values using the data shown above. Please specify the model type and all the components of the marker-effect model and describe how genomic breeding values are predicted. The ratio (λ) between the residual variance (σ_e^2) and the QTL-variance (σ_q^2) can assumed to be 1.

Verwenden Sie ein "Single-Step" Markereffektmodell für die Schätzung der genomischen Zuchtwerte. Bitte geben Sie den Modelltyp und alle Komponenten des Markereffektmodells and und beschreiben Sie, wie die genomischen Zuchtwerte berechnet werden. Das Verhältnis (λ) zwischen der Restvarianz (σ_e^2) und der QTL-Varianz (σ_q^2) kann als 1 *angenommen werden.*

15

Solution

c) Use a single-step Genomic BLUP (GBLUP) model to predict genomic breeding values. Please specify the type of model used and also list all the model components of the GBLUP model used. The ratio (λ) between the residual variance (σ_e^2) and the genetic variance (σ_g^2) can assumed to be 1.

Verwenden Sie ein "Ein-Schritt" genomisches BLUP (GBLUP) Modell zur Schätzung der genomischen Zuchtwerte. Bitte geben Sie den Modelltyp und alle Modellkomponenten an, welche im GBLUP-Modell vorkommen. Das Verhältnis (λ) zwischen der Restvarianz (σ_e^2) und der genetischen Varianz (σ_g^2) kann als 1 *angenommen werden.*

15

Solution