Peter von Rohr

Institut für Agrarwissenschaften

D-USYS

ETH Zürich

751-7602-00 V
Lösungen zur Prüfung
Angewandte Statistische Methoden
in den Nutzierwissenschaften
FS 2018

Datum: 28. Mai 2018

Name:

Legi-Nr:

Aufgabe	Maximale Punktzahl	Erreichte Punktzahl
1	26	
2	20	
3	9	
4	24	
Total	79	

Questions in English are in italics

Aufgabe 1: Lineare Regression

a) In der Tierzucht wird das Körpergewicht oft anhand einer linearen Regression auf den Brustumfang geschätzt. Zur Schätzung der Regressionsparameter werden die folgenden Daten verwendet.

In animal breeding body weight (Körpergewicht) is often estimated using a linear regression on chest size (Brustumfang). The regression coefficient ist estimated based on the following dataset.

14

Tier	Brustumfang	Gewicht
1	141	750
2	148	779
3	139	732
4	142	752
5	141	751
6	152	807
7	152	797
8	138	735
9	151	802
10	145	762
11	141	740
12	142	750

Das Resultat der Regression des Gewichts auf den Brustumfang lautet wie folgt

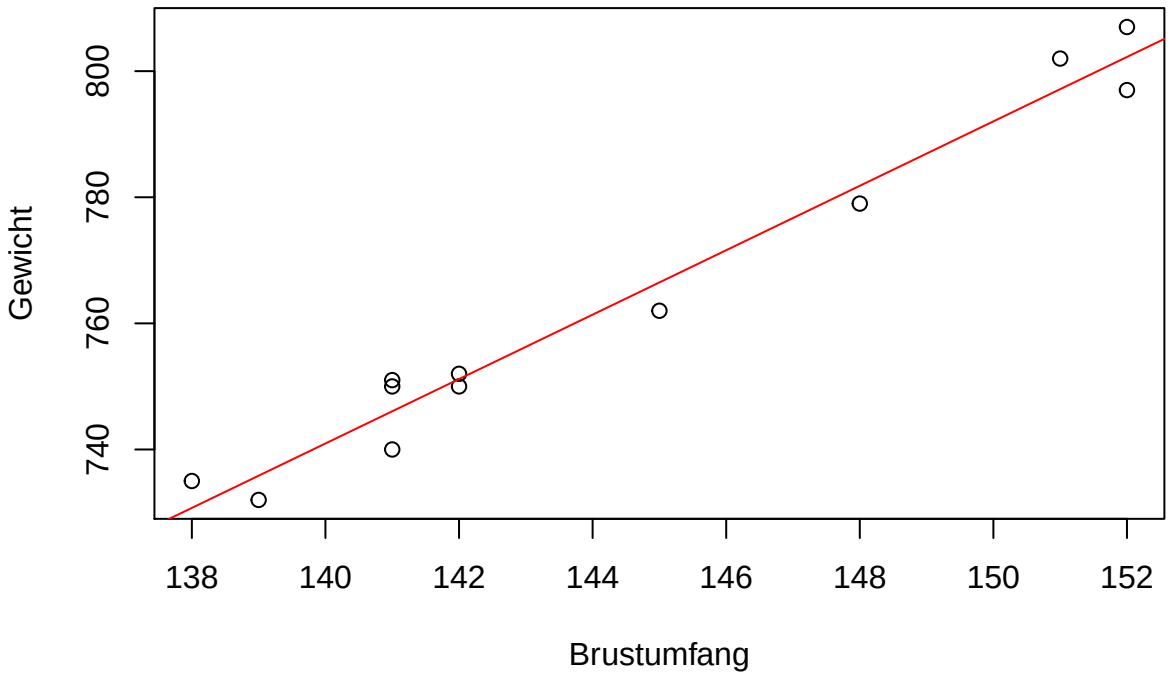
The results of the regression of weight on chest size is given below

```
##
## Call:
## lm(formula = Gewicht ~ Brustumfang, data = tbl_bu_ge_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.0548 -4.0006 -0.1634  4.3909  4.9452
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   25.750     39.277   0.656    0.527
## Brustumfang    5.109      0.272  18.784 3.96e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.621 on 10 degrees of freedom
## Multiple R-squared:  0.9724, Adjusted R-squared:  0.9697
## F-statistic: 352.8 on 1 and 10 DF,  p-value: 3.96e-09
```

Ihre Aufgabe

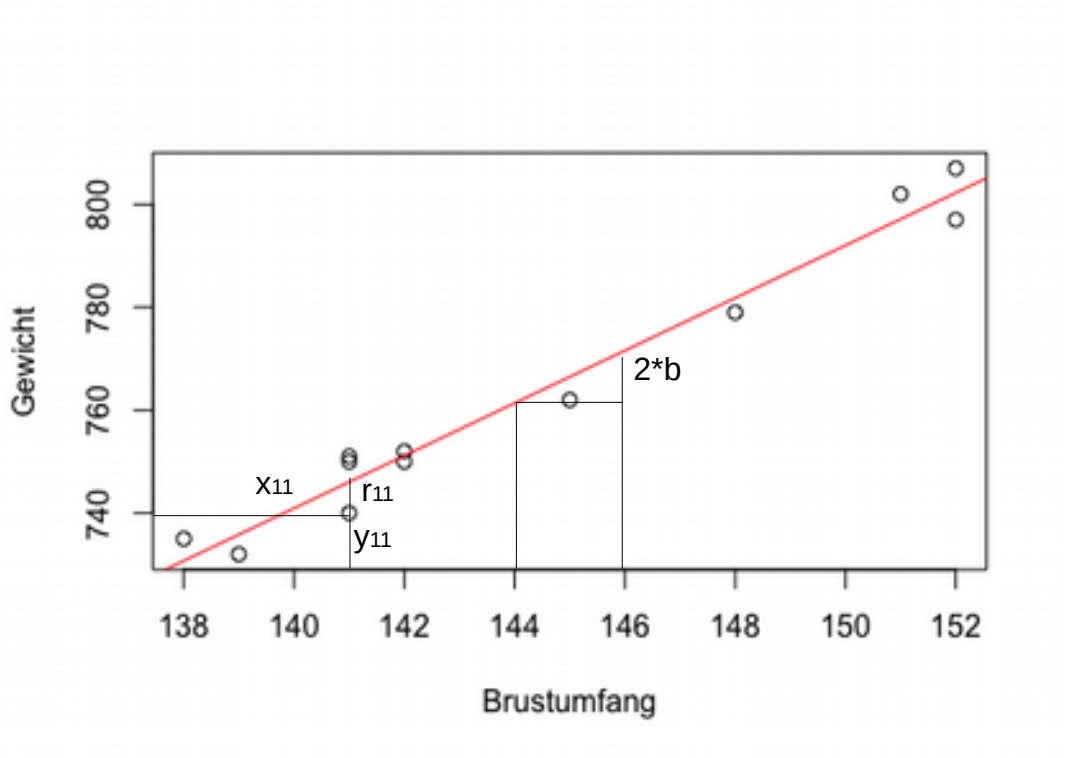
Zeichnen Sie für alle Beobachtungen die erklärende Variable (als x_1 bis x_{12}), die Zielgrösse (als y_1 bis y_{12}), die Residuen (als r_1 bis r_{12}) und die geschätzten Regressionskoeffizienten in den folgenden Plot ein.

Draw for all observations the explaining variable (x_1 to x_{12}), the target variable (y_1 to y_{12}), the residuals (r_1 to r_{12}) and the estimated regression coefficients into the following plot.



Lösung:

Die für die Beobachtung 11 gezeigten Grössen x_{11}, y_{11} und r_{11} sollen für alle Beobachtungen eingezeichnet werden



b) Für die ersten drei Tiere aus dem Datensatz aus Aufgabe 1a) werden SNP-Daten an fünf SNP-Loci erhoben. Mit diesen SNP-Daten wird versucht den genetischen Anteil des Gewichts der drei Tiere zu schätzen. In einer ersten Analyse wurden die SNP-Effekte mit einer Regressionsanalyse geschätzt. Die Resultate sind im nachfolgenden Output gezeigt. Wo liegt das zentrale Problem dieser Analyse? Schlagen Sie ein alternatives Analyseverfahren vor, welches diese Probleme nicht hat.

For the first three animals of the data set of 1a) SNP-data at five loci were collected. The genetic part of the trait body weight should be estimated based on the SNP-data for the three animals. In a first analysis, the SNP-effects were estimated with a regression analysis. The results are shown in the following output. Where is the central problem of this analysis? Propose an alternative analysis to regression analysis which does not have the observed problem.

4

```
##
## Call:
## lm(formula = Gewicht ~ 0 + SNP1 + SNP2 + SNP3 + SNP4 + SNP5,
##     data = tbl_snp_ge_data)
##
## Residuals:
## ALL 3 residuals are 0: no residual degrees of freedom!
##
## Coefficients: (2 not defined because of singularities)
##      Estimate Std. Error t value Pr(>|t|)
## SNP1        9         NA      NA       NA
## SNP2       29         NA      NA       NA
## SNP3       NA         NA      NA       NA
## SNP4      741         NA      NA       NA
## SNP5       NA         NA      NA       NA
##
## Residual standard error: NaN on 0 degrees of freedom
## Multiple R-squared:      1,  Adjusted R-squared:    NaN
## F-statistic:   NaN on 3 and 0 DF,  p-value: NA
```

Lösung:

- Anzahl Beobachtungen n ist kleiner als die Anzahl Parameter p. Somit können die unbekannten Parameter nicht mit Least Squares geschätzt werden. Deshalb die seltsamen Resultate.
- Eine Alternative ist LASSO, GBLUP oder ein Bayes'sches Verfahren.

c) Gegeben sind die folgenden beiden Resultate (als 'summary()' und 'plot()') zweier Regressionsanalysen (mit 'lm()'). Welcher 'summary()'-Output gehört zu welchem Plot? Zeichnen Sie die geschätzten Regressionskoeffizienten in die Plots ein.

Given are the following results (using 'summary()' and 'plot()') from two regression analyses (using 'lm()'). Which 'summary()'-Output belongs to which plot? Show the estimated regression coefficient by drawing its value into the plot.

8

```
## Warning: `data_frame()` is deprecated, use `tibble()`.
## This warning is displayed once per session.
```

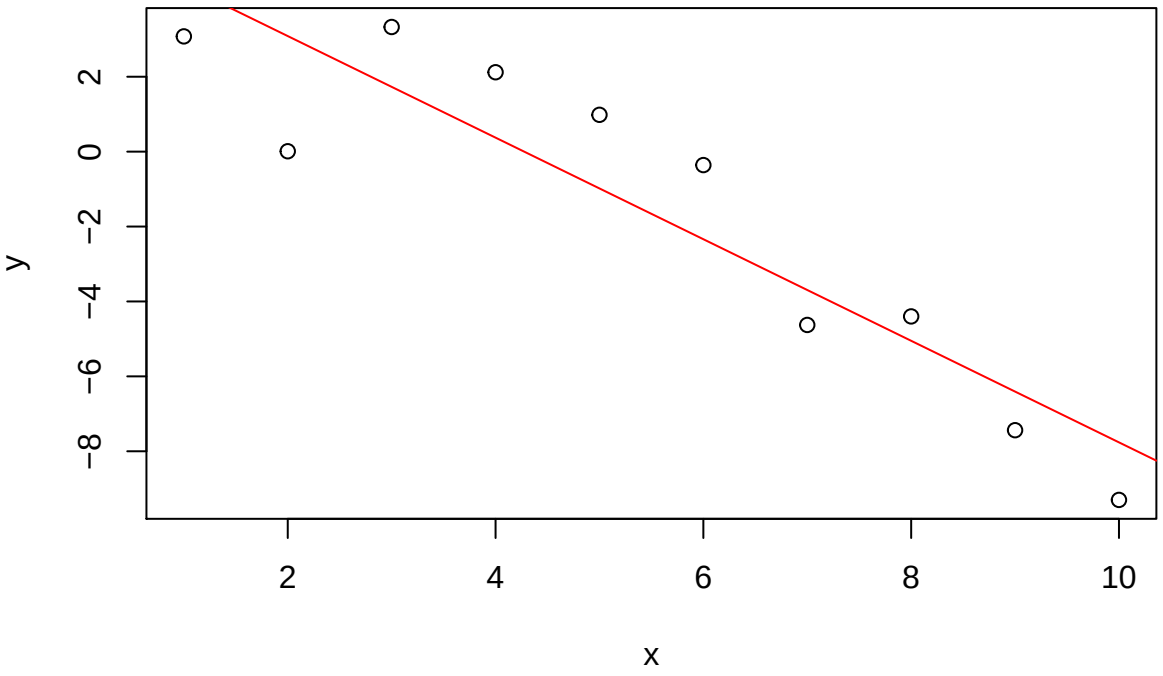
summary()-Output 1:

```
##
## Call:
## lm(formula = y ~ x, data = tbl_task1c_o1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.70275 -0.50765 -0.22913  0.09658  2.58440
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   3.2245     0.8361   3.857 0.004831 **
## x             0.7862     0.1347   5.835 0.000389 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.224 on 8 degrees of freedom
## Multiple R-squared:  0.8097, Adjusted R-squared:  0.7859
## F-statistic: 34.04 on 1 and 8 DF,  p-value: 0.0003895
```

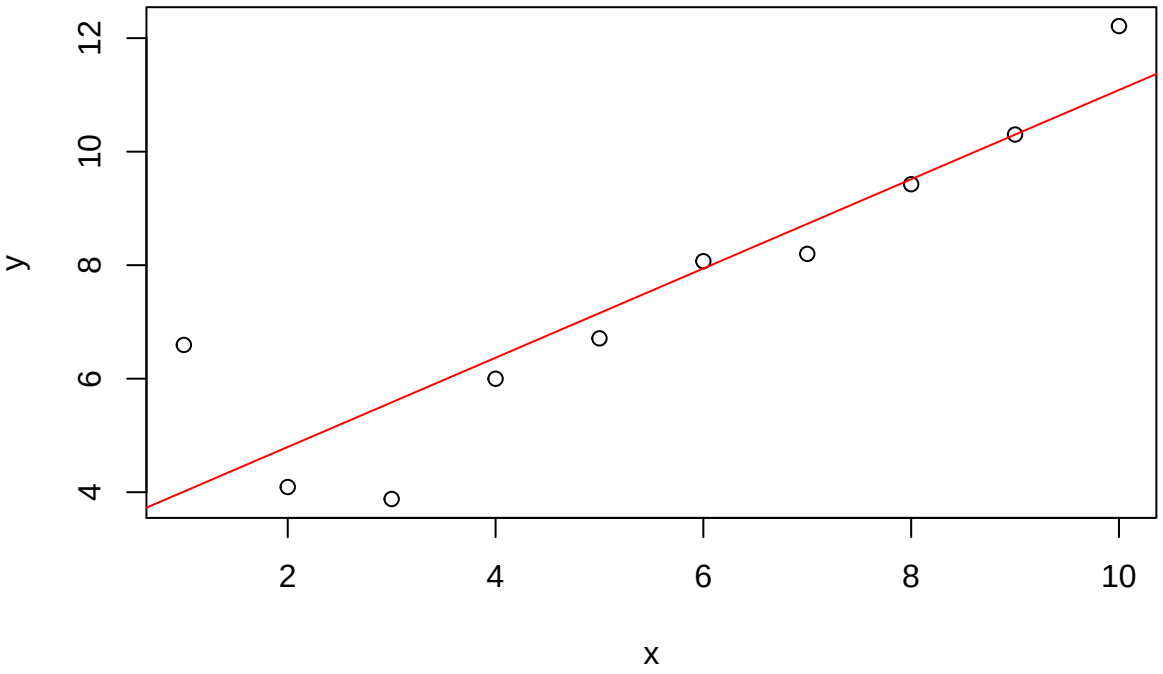
summary()-Output 2:

```
##
## Call:
## lm(formula = y ~ x, data = tbl_task1c_o2)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -3.075 -1.281 -0.141  1.709  1.979
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   5.7977     1.3101   4.425 0.002210 **
## x            -1.3560     0.2111  -6.422 0.000204 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.918 on 8 degrees of freedom
## Multiple R-squared:  0.8375, Adjusted R-squared:  0.8172
## F-statistic: 41.24 on 1 and 8 DF,  p-value: 0.0002043
```

Plot 1:



Plot 2:



Lösung:

Plot 2 gehört zu Output 1 und Plot 1 zu Output 2. Die Regressionskoeffizienten entsprechen den Steigungen der roten Geraden.

Aufgabe 2: Genomisches BLUP

a) Zum Datensatz aus Aufgabe 1b) wird für die drei Tiere zusätzlich zum Gewicht und zu den SNP-Informationen noch das Geschlecht ermittelt. Der Datensatz ist in der nachfolgenden Tabelle dargestellt.

The dataset given in 1b) is augmented with the sex of the three animals. The data set is given in the following table.

6

Tier	Gewicht	Geschlecht	SNP1	SNP2	SNP3	SNP4	SNP5
1	750	F	1	0	1	1	0
2	779	M	1	1	0	1	0
3	732	F	-1	0	-1	1	-1

Ihre Aufgabe

Die Daten sollen mit einem Ridge-Regression (RR) BLUP Modell analysiert werden. Stellen Sie das RR-BLUP Modell in Matrix-Vektor-Schreibweise für die gezeigten Daten auf, wobei das Gewicht die Zielgrösse darstellt und das Geschlecht als fixer Effekt behandelt werden soll. Füllen Sie für jeden Vektor und für jede Matrix auch die konkreten Zahlen aus den Daten ab, soweit dies möglich ist.

The above shown dataset is to be analysed using a Ridge-Regression (RR) model. Setup the RR-BLUP model for the shown data. Use a matrix-vector notation. Body weight is the target variable and sex should be used as fixed effect. Fill the matrices and the vectors with the numbers coming from the data, as it is possible.

Lösung

In RR-BLUP werden die einzelnen SNP-Effekte als zufällige Effekte im Modell berücksichtigt. Somit erhalten wir das folgende Modell

$$y = Xb + Wq + e$$

wobei:

- y der Vektor der Beobachtungen mit

$$y = \begin{bmatrix} 750 \\ 779 \\ 732 \end{bmatrix}$$

- b der Vektor der unbekannten fixen Effekte des Geschlechts mit

$$b = \begin{bmatrix} b_F \\ b_M \end{bmatrix}$$

- X die Designmatrix, welche fixe Effekte zu Beobachtungen verknüpft

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

- q der Vektor der unbekannten zufälligen SNP-Effekte

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

- W die Designmatrix, welche die SNP-Effekte und die Beobachtungen miteinander verknüpft.

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

- e der Vektor der unbekannten zufälligen Fehlerterme

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

b) Verwenden Sie die gleichen Daten, wie in Aufgabe 2a) und stellen Sie dafür das genomische BLUP-Modell (GBLUP) Modell auf. Füllen Sie für jeden Vektor und für jede Matrix die Zahlen aus dem Datensatz ab, soweit dies möglich ist.

Use the same data as in 2a and setup the genomic BLUP (GBLUP) model. Fill the matrices and the vectors with the numbers coming from the data, as it is possible.

6

Lösung

Im GBLUP Model werden die genomischen SNP-Effekte pro Tier in einem zufälligen Effekt zusammengefasst. Das Modell lautet somit

$$y = Xb + Zg + e$$

wobei:

- y der Vektor der Beobachtungen mit

$$y = \begin{bmatrix} 750 \\ 779 \\ 732 \end{bmatrix}$$

- b der Vektor der unbekannten fixen Effekte des Geschlechts mit

$$b = \begin{bmatrix} b_F \\ b_M \end{bmatrix}$$

- X die Designmatrix, welche fixe Effekte zu Beobachtungen verknüpft

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

- g der Vektor der unbekannten zufälligen genomischen Zuchtwert

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

- Z die Designmatrix, welche die SNP-Effekte und die Beobachtungen miteinander verknüpft.

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

- e der Vektor der unbekannten zufälligen Fehlerterme

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

c) Angenommen es gibt ein zusätzliches viertes Tier zu den Tieren im Datensatz von 2a), welches nur SNP-Informationen aufweist. Stellen Sie die Mischmodellgleichungen für das GBLUP-Modell auf und schätzen sie den genomischen Zuchtwert für das vierte Tier.

Suppose that we are given an additional animal with only SNP-Information. Setup the mixed-model equations for the GBLUP-Model and estimate the genomic breeding value for the fourth animal.

8

Tier	Gewicht	Geschlecht	SNP1	SNP2	SNP3	SNP4	SNP5
1	750	F	1	0	1	1	0
2	779	M	1	1	0	1	0
3	732	F	-1	0	-1	1	-1
4	NA	NA	1	1	0	-1	0

Lösung:

Basierend auf den Mischmodellgleichungen, wir bekommen:

$$\begin{bmatrix} X^TX & X^TZ & 0 \\ Z^TX & Z^TZ + G^{11} & G^{12} \\ 0 & G^{21} & G^{22} \end{bmatrix} \begin{bmatrix} \hat{b} \\ \hat{g}_1 \\ \hat{g}_2 \end{bmatrix} = \begin{bmatrix} X^Ty \\ Z^Ty \\ 0 \end{bmatrix}$$

wobei die Matrix G^{11} der inversen genomischen Verwandtschaftsmatrix G, welche den Tieren mit Beobachtungen und Genotypen entspricht. Die Matrix G^{22} entspricht dem Teil der inversen genomischen Verwandtschaftsmatrix, welcher zu den Tieren gehört, welche nur SNP-Genotypen aufweisen. Für unser Beispiel sind die Tiere 1 bis 3 in G^{11} enthalten und das Tier 4 ist in G^{22}.

Aufgrund der letzten Zeile der Mischmodellgleichungen folgt der genomische Zuchtwert $\hat{g}_4$ vom Tier 4 dem folgenden Ausdruck.

$$\hat{g}_4 = -\left(G^{22}\right)^{-1} G^{21}\hat{g}_1$$

Aufgabe 3: LASSO

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

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.

4

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

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

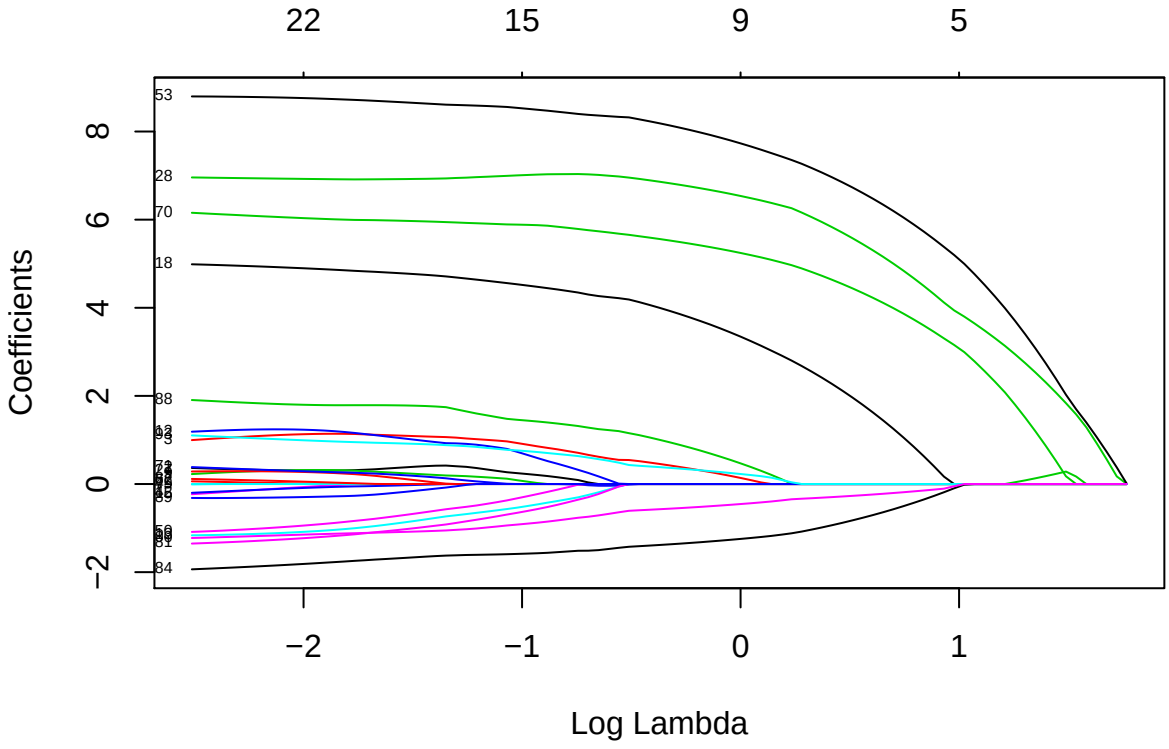
Lösung

Die erste Gleichung gehört zu LASSO und die zweite zu Least Squares.

b) 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.

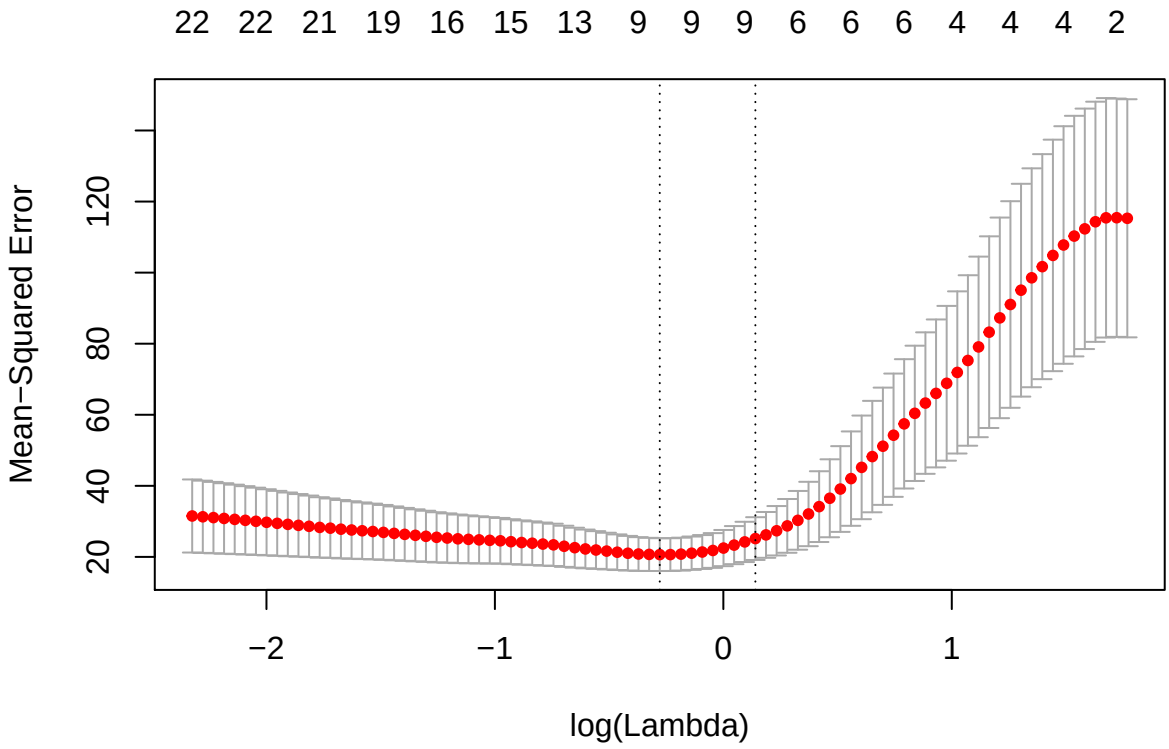
We analyse a genomic dataset with 25 animals which are genotyped at 100 SNP-locations. Only 5 SNPs have an effect on the observed trait. The SNP-effects are estimated with LASSO. The results of this analysis are shown in the two plots below. The second plot can be used to determine the penalty-term (Log 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.

5



Der Strafterm kann aufgrund der linken gestrichelten Linie bestimmt werden.

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



Lösung

Die fünf SNP mit den grössten absoluten Effekten sind: 53, 28, 70, 18, 84

Aufgabe 4: Bayes

a) In einer Bayes'schen Datenanalyse wird zwischen bekannten und unbekannten Grössen unterschieden. Machen Sie die Einteilung für den Datensatz aus Aufgabe 2c).

In a Bayesian data analysis, we differentiate between known and unknown quantities. Do this differentiation for the dataset of 2c).

12

Tier	Gewicht	Geschlecht	SNP1	SNP2	SNP3	SNP4	SNP5
1	750	F	1	0	1	1	0
2	779	M	1	1	0	1	0
3	732	F	-1	0	-1	1	-1
4	NA	NA	1	1	0	-1	0

Für den Datensatz nehmen wir das folgende Modell an

For the above shown dataset, we assume the following model

$$y = Xb + Zq + e$$

wobei: das Geschlecht als fixen Effekt (b) und die SNP-Effekte (q) als zufällig modelliert werden.

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

Lösung

Grösse	bekannt
y_1	ja
y_2	ja
y_3	ja
y_4	ja
b_M	nein
b_F	nein
q_1	nein
q_2	nein
q_3	nein
q_4	nein
q_5	nein
Z	ja

b) Für das nachfolgend gegebene lineare Regressionsmodell soll die Parameter mit einer Bayes'schen Analyse geschätzt werden. Die Analyse wird anhand des gezeigten R-Codeblocks gemacht. Berechnen Sie die Bayes'schen Schätzwerte für den Achsenabschnitt β_0 und den Regressionsparameter β_1 aufgrund des Outputs des gezeigten Programmcodes.

Parameters of the following given model are to be estimated using a Bayesian analysis. The analysis is done using the shown R-code below. Compute the Bayesian estimator for the intercept β_0 and the regression parameter β_1 based on the output of the R-program.

10

Das Programm, welches den Gibbs Sampler umsetzt und die Stichproben von β_0 und β_1 erzeugt, sieht wie folgt aus. Berechnen Sie aus den gezeigten Stichproben eine Bayes'sch Schätzung für β_0 und β_1.

The program below implements the Gibbs Sampler to generate samples of β_0 and β_1. Compute the Bayesian estimates of β_0 and β_1 based on the samples shown in the output.

```
### # Matrix X as incidence matrix for beta0 and beta1
X <- cbind(1,dfDataRead$x)
### # y as vector of observations
y <- dfDataRead$y
### # starting values
beta = c(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 = X[, 1]
  xpxi = 1/(t(x) %*% x)
  betaHat = t(x) %*% w * xpxi
  beta[1] = rnorm(1, betaHat, sqrt(xpxi)) # using residual var = 1
  # sampling slope
  w = y - X[, 1] * beta[1]
  x = X[, 2]
  xpxi = 1/(t(x) %*% x)
  betaHat = t(x) %*% w * xpxi
  beta[2] = rnorm(1, betaHat, sqrt(xpxi)) # using residual var = 1
  # output current sample
  cat("iteration: ", iter, " -- beta0: ", beta[1], " -- beta1: ", beta[2], "\n")
}
```

```
## iteration:  1  -- beta0:  -2.381084  -- beta1:  0.8525947
## iteration:  2  -- beta0:  -2.759053  -- beta1:  1.264267
## iteration:  3  -- beta0:  -3.438714  -- beta1:  1.458278
## iteration:  4  -- beta0:  -3.831475  -- beta1:  1.793831
## iteration:  5  -- beta0:  -3.785601  -- beta1:  1.399844
## iteration:  6  -- beta0:  -3.591078  -- beta1:  1.742267
## iteration:  7  -- beta0:  -4.307856  -- beta1:  1.717601
## iteration:  8  -- beta0:  -3.660281  -- beta1:  1.520865
## iteration:  9  -- beta0:  -3.854872  -- beta1:  1.868808
## iteration:  10  -- beta0:  -4.208202  -- beta1:  2.052571
```

Lösung

Die Schätzwerte entsprechen den Mittelwerten aus den Samples diese sind dann

```
## iteration:  1  -- beta0:  -2.019969  -- beta1:  0.4436744
## iteration:  2  -- beta0:  -3.169093  -- beta1:  0.987194
## iteration:  3  -- beta0:  -3.269731  -- beta1:  1.376081
## iteration:  4  -- beta0:  -3.48327  -- beta1:  1.528404
## iteration:  5  -- beta0:  -3.576744  -- beta1:  1.498536
## iteration:  6  -- beta0:  -3.371717  -- beta1:  1.435582
## iteration:  7  -- beta0:  -3.763373  -- beta1:  1.238522
## iteration:  8  -- beta0:  -3.810358  -- beta1:  1.540668
## iteration:  9  -- beta0:  -3.746737  -- beta1:  1.482393
## iteration:  10  -- beta0:  -3.639203  -- beta1:  1.276948
```

```
## Estimates:
```

```
## [1] -3.38502  1.28080
```

c) Berechnen Sie den Standardfehler für die Bayes'schen Schätzungen von β_0 und β_1 aus 4b)
Compute the standard error of the Bayesian estimates for β_0 and β_1 of 4b)

2

Lösung

Die Standardfehler entsprechen den Standardabweichungen der Samples. Diese sind

```
## iteration:  1  -- beta0:  -2.620517  -- beta1:  0.5736311
## iteration:  2  -- beta0:  -2.758377  -- beta1:  1.112871
## iteration:  3  -- beta0:  -3.716958  -- beta1:  1.610732
## iteration:  4  -- beta0:  -4.21154  -- beta1:  1.91471
## iteration:  5  -- beta0:  -4.218863  -- beta1:  1.843484
## iteration:  6  -- beta0:  -4.174579  -- beta1:  1.968959
## iteration:  7  -- beta0:  -4.215382  -- beta1:  1.784188
## iteration:  8  -- beta0:  -4.183714  -- beta1:  2.048001
## iteration:  9  -- beta0:  -4.530491  -- beta1:  1.941263
## iteration:  10  -- beta0:  -4.651221  -- beta1:  2.408128

## Standard Errors:

## [1] 0.6977655 0.5211127
```