Peter von Rohr

Institut für Agrarwissenschaften

D-USYS

ETH Zürich

751-7602-00 V

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 in der klassischen Zuchtwertschätzung und der genomischen Selektion an.

6

Lösung:

Punkt	klassische Zuchtwertschätzung	Genomische Selektion
Zielgrössen		
Vorteile		
Nachteile		

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

2

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

2

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
```

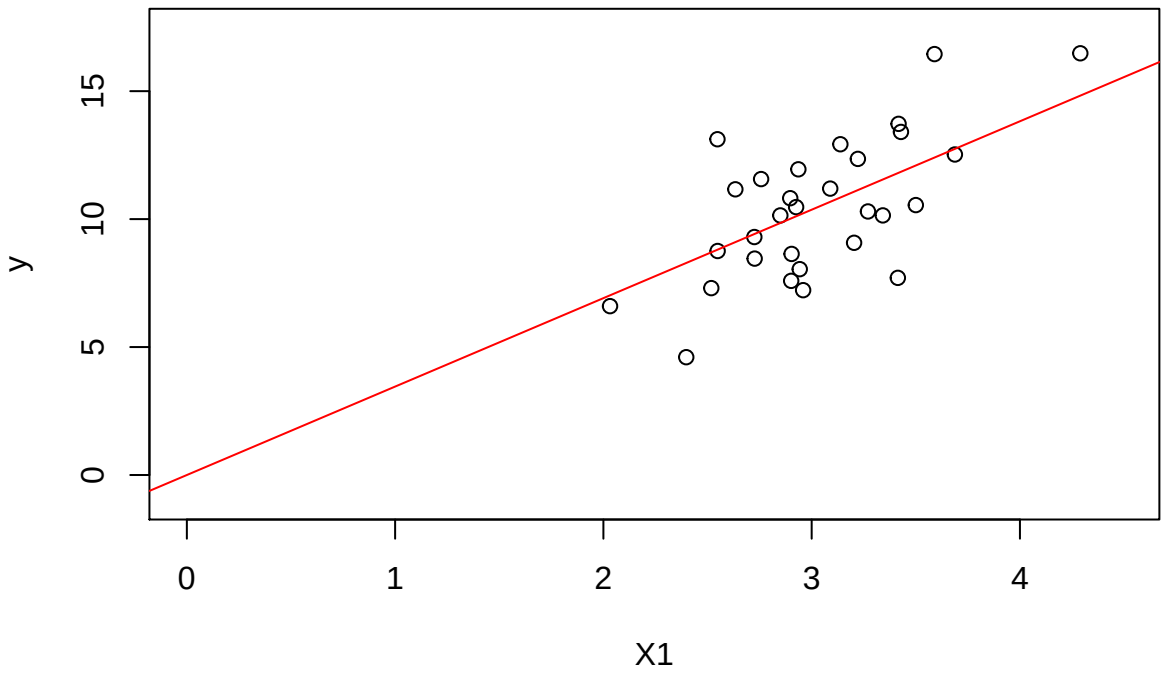
a) Geben Sie die Formeln der beiden statistischen Modelle an, welche zu Output A und Output B geführt haben. Wo liegt der hauptsächliche Unterschied zwischen den beiden Modellen

8

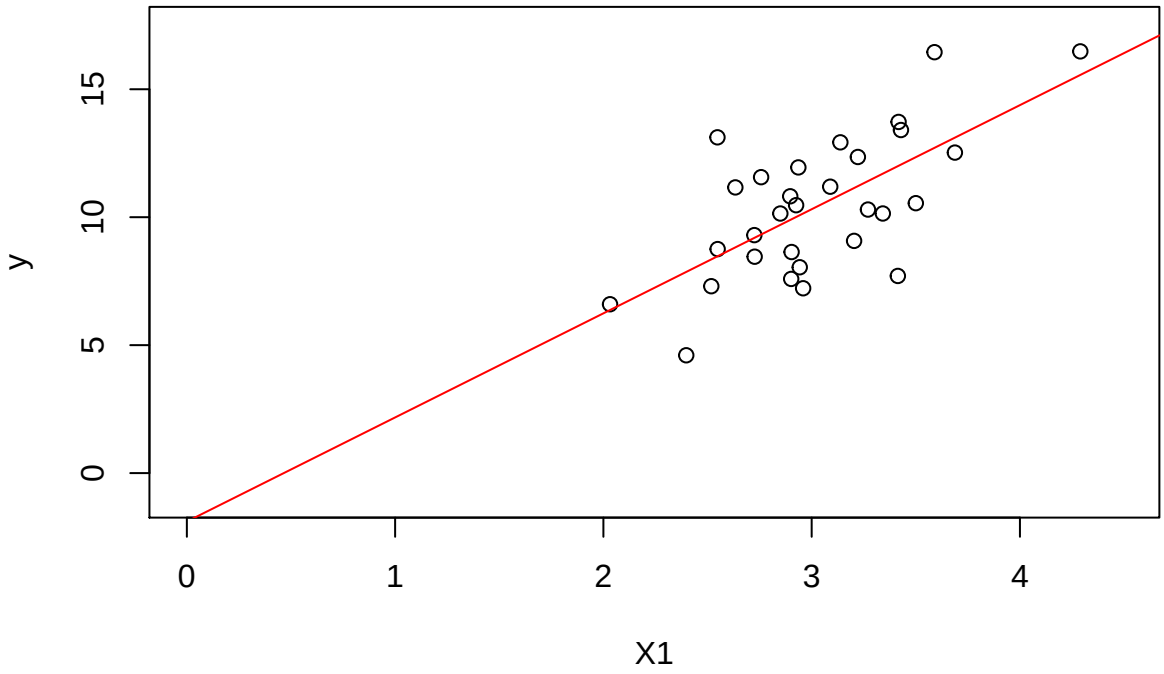
b) Für die zwei Analysen wurden auch zwei Plots gemacht. Ordnen Sie die Plots 1 und 2 den Outputs A und B zu.

2

Plot 1



Plot 2

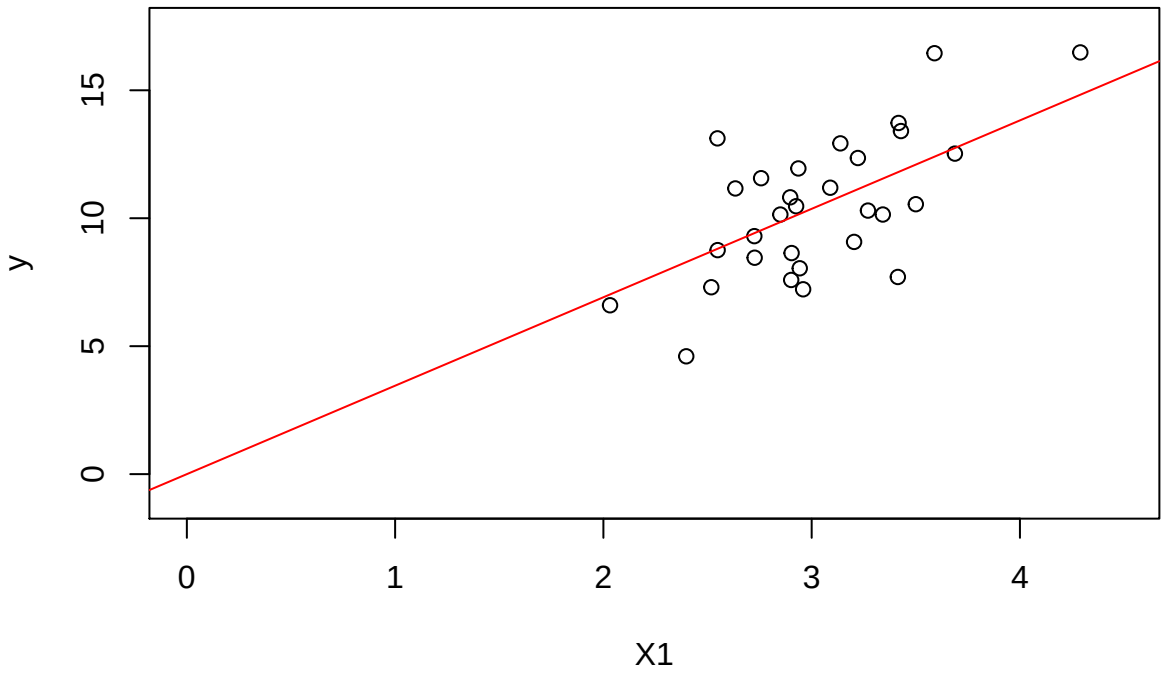


- Plot 1 gehört zu Output
- Plot 2 gehört zu Output

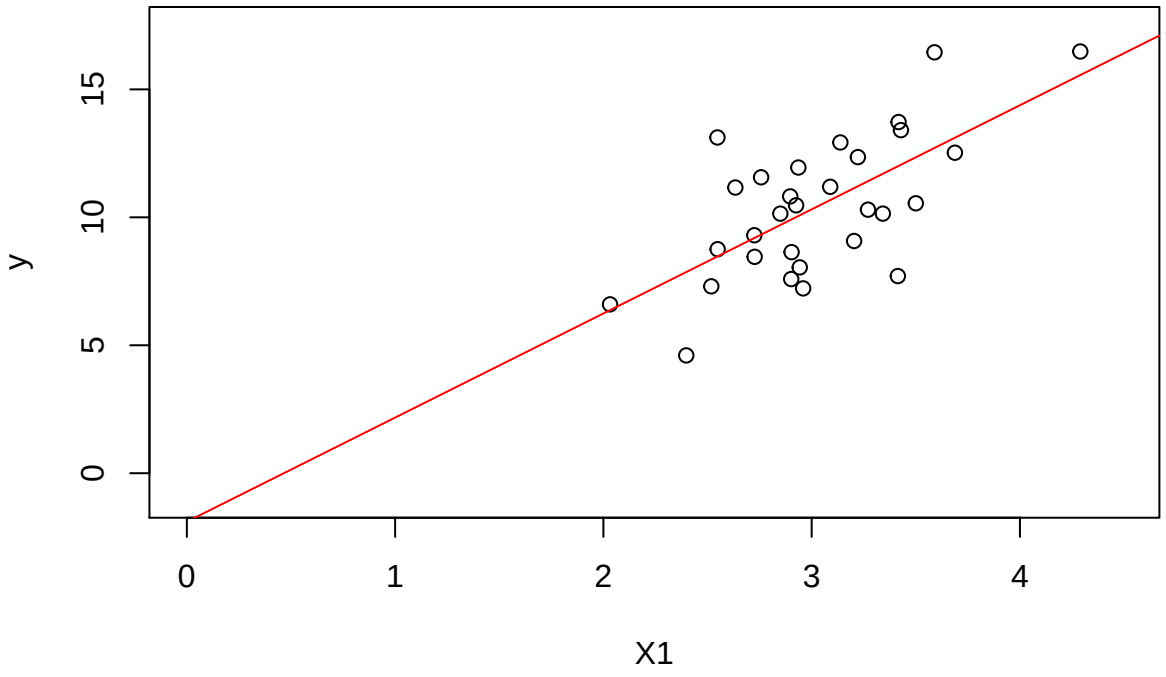
c) Zeichnen Sie die geschätzten Parameter (Estimate) aus den Outputs A und B in die Plots 1 und 2 ein

6

Plot 1



Plot 2



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

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

4

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

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

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

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

c) Gegeben ist der folgende Datensatz. Bei allen SNPs nehmen wir an, dass G_1 das Allel mit der positiven Wirkung sei. Stellen Sie die Modelle 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

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