Peter von Rohr

Institute of Agricultural Sciences
D-USYS
ETH Zurich

751-6212-00L V
Solutions to Exam
Applied Genetic Evaluation
SS 2019

Date: 27th May 2019

Name:

Legi-Nr:

Problem	Maximum Number of Points	Number of Points Reached
1	25	
2	26	
3	21	
4	39	
Total	111	

Questions in German are in italics

Problem 1: Model Selection

a) Model selection can be done using **forward** selection or **backward** selection. Describe the four steps that are needed in both model selection procedures.

*Modellselektion kann mit **Forward** Selektion oder **Backward** Selektion gemacht werden. Beschreiben Sie die vier Schritte, welche es für beide Modellselektionsprozeduren braucht.*

16

Solution

Step	Forward	Backward
1	Start with smallest model $\mathcal{M}_0$	Start with full model $\mathcal{M}_0$
2	Include predictor reducing RSS the most	Eliminate predictor increasing RSS the least
3	Continue step 2 until all predictors chosen	Continue step 2 until all predictors eliminated
4	From sequence of submodels $\mathcal{M}_0 \subseteq \mathcal{M}_1 \subseteq \mathcal{M}_2 \subseteq \dots$ choose model with smallest C_p	From sequence of submodels $\mathcal{M}_0 \supseteq \mathcal{M}_1 \supseteq \mathcal{M}_2 \supseteq \dots$ choose model with smallest C_p

b) We are given the following data set on body weight, breast circumference and shoulder height. Furthermore, the results of fitting the full regression model and a reduced model with only breast circumference as predictor. Compute the Mallow C_p value for both models and decide which of the two models is better based on C_p

Gegeben ist der folgende Datensatz zu Körpergewicht, Brustumfang und Widerristhöhe. Weiter sind auch die Resultate des Fits des vollen Modells und eines reduzierten Models mit nur Brustumfang als beschreibende Variable gegeben. Berechnen Sie den Mallow C_p-Wert für beide Modelle und begründen Sie, welches der beiden Modelle besser ist.

6

Animal	BreastCircumference	BodyWeight	ShoulderHeight
1	176	471	136
2	177	463	147
3	178	481	147
4	179	470	148
5	179	496	151
6	180	491	149
7	181	518	151
8	182	511	154
9	183	510	147
10	184	541	148

The results of the full model

Die Resultate des vollen Modells

```
##
## Call:
## lm(formula = BodyWeight ~ BreastCircumference + ShoulderHeight,
##     data = tbl_reg_withsh)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -16.9724  -5.8398   0.9169   8.2185  14.0487
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)         -1078.2804   272.5420  -3.956  0.00549 **
## BreastCircumference     9.0576     1.8103   5.003  0.00156 **
## ShoulderHeight         -0.3788     0.9950  -0.381  0.71473
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.73 on 7 degrees of freedom
## Multiple R-squared:  0.827,  Adjusted R-squared:  0.7776
## F-statistic: 16.73 on 2 and 7 DF,  p-value: 0.002153
```

The results of the reduced model

Die Resultate des reduzierten Modells

```
##
## Call:
## lm(formula = BodyWeight ~ BreastCircumference, data = tbl_reg_withsh)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -17.3941  -6.5525  -0.0673   9.3707  13.2594
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)         -1065.115    255.483  -4.169 0.003126 **
## BreastCircumference     8.673      1.420   6.108 0.000287 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.08 on 8 degrees of freedom
## Multiple R-squared:  0.8234, Adjusted R-squared:  0.8014
## F-statistic: 37.31 on 1 and 8 DF,  p-value: 0.000287
```

Solution

Mit `olsrr` erhalten wir

```
olsrr::ols_step_best_subset(lm_bw_bc_sh)
```

```
##             Best Subsets Regression
## ---------------------------------------------------
## Model Index    Predictors
## ---------------------------------------------------
##      1         BreastCircumference
##      2         BreastCircumference ShoulderHeight
## ---------------------------------------------------
##
##                                                Subsets Regression Summary
## ------------------------------------------------------------------------------------------------------------------------------------------
##                        Adj.        Pred
## Model    R-Square    R-Square    R-Square     C(p)       AIC       SBIC        SBC        MSEP        FPE          HSP       APC
## ------------------------------------------------------------------------------------------------------------------------------------------
##   1        0.8234      0.8014      0.7137    1.1449    80.2529    53.1531    81.1606    154.3767    147.3596    17.5428    0.2649
##   2        0.8270      0.7776      0.5105    3.0000    82.0480    55.8733    83.2583    201.6608    178.7448    22.9160    0.3213
## ------------------------------------------------------------------------------------------------------------------------------------------
## AIC: Akaike Information Criteria
##  SBIC: Sawa's Bayesian Information Criteria
##  SBC: Schwarz Bayesian Criteria
##  MSEP: Estimated error of prediction, assuming multivariate normality
##  FPE: Final Prediction Error
##  HSP: Hocking's Sp
##  APC: Amemiya Prediction Criteria
```

Das reduzierte Model erhält einen tieferen C_p-Wert und ist somit das bessere Modell.

c) What is the reason why we are doing model selection, why should we not take always the full model? Please explain the underlying phenomenon in your own words.

Was ist der Grund, dass wir Modellselektion machen? Weshalb können wir nicht einfach immer das volle Modell mit allen erklärenden Variablen zur Anpassung verwenden? Bitte erklären Sie das unterliegende Phänomen in Ihren eigenen Worten.

3

Solution

The problem that makes model selection necessary is called Bias-Variance trade-off. The Bias is reduced, the more predictors we include. But on the other hand, the variance is increased. Hence we have to find the model that balances this trade-off. This is done with a criterion such as the Mallow C_p statistic.

Problem 2: Breeding Programs

The breeding goal 2021 (`https://homepage.braunvieh.ch/xml_1/internet/de/application/d2/d9/f464.cfm`) for the `Brown Swiss` cattle breed contains the following information

Das Zuchtziel 2021 (`https://homepage.braunvieh.ch/xml_1/internet/de/application/d2/d9/f464.cfm` *) für das Braunvieh enthält die folgenden Informationen*

Merkmal	Zielwert für Braunvieh
Jährlicher Zuchtfortschritt Milch	+60 kg
Jährlicher Zuchtfortschritt Eiweiss	+2 kg
Milchleistung Ausgewachsene Kuh Talzone	8500 kg
Milchgehalt Eiweiss	3.5%
Laktationspersistenz	85% und höher
Nutzungsdauer (ZWS)	positiver genetischer Trend
Lebensleistung bei Abgang (Talzone)	35000 kg
Eutergesundheit (ZW Zellzahl)	positiver genetischer Trend
ZW Rastzeit	positiver genetischer Trend
ZW Non Return Rate	positiver genetischer Trend
Serviceperiode (phänotypisch)	unter 120 Tagen
Rahmen, Becken, Fundament, Euter, Zitzen	positiver genetischer Trend
Grösse ausgewachsene Kühe	140 - 152 cm

The aggregate genotype of the Swiss `Edelschein` dam lines has the following weights for the different trait groups `Conformation`, `Production` and `Reproduction`.

Im Gesamtzuchtwert der Edelschwein Mutterlinie werden die verschiedenen Merkmalsgruppen 'Exterieur', 'Produktion' und 'Reproduktion' wie folgt gewichtet.

Das Zuchtziel 2021 (`https://homepage.braunvieh.ch/xml_1/internet/de/application/d2/d9/f464.cfm` *) für das Braunvieh enthält die folgenden Informationen*

Merkmalsgruppe	Gewicht im Gesamtzuchtwert
Exterieur	19
Produktion	30
Reproduktion	51

a) What are the names of the two different formulation of the breeding goals for 'Brown Swiss' cattle and for 'Edelschein' pigs? Please complete the following table and in doing so, indicate both an advantage and a disadvantage of both formulations.

Wie heissen die beiden verschiedenen Arten der Formulierung eines Zuchtziels beim Braunvieh und beim Edelschwein? Bitte füllen Sie die nachfolgende Tabelle aus und geben Sie je einen Vorteil und einen Nachteil der beiden Formulierungen an.

12

Solution

	Braunvieh	Edelschwein
Formulation	Political breeding goal with extensive description of mostly phenotypic properties of desired model breeding animal	Scientific breeding goal with aggregate genotype as mathematical form of breeding goal
Advantage	Easy to understand and easy to picture the model breeding animal	Based on solid theory, can be applied for all breeding animals, success can be verified
Disadvantage	Not clear what should be done with over-achiever animals. Goals can be contradictory. Success cannot be verified.	Not easy to understand. Difficult to get clear picture for single breeding animals

b) Genomic Selection has the potential to increase the genetic gain per year significantly. Currently beef cattle breeders start to introduce genomic selection for carcass traits. Let us assume the following selection parameters for the trait average daily gain. Compute the selection response per year for both scenarios with and without genomic selection. Based on the selection response per year which scenario is better? The trait is assumed to be normally distributed.

Die genomische Selektion hat das Potential den Zuchtfortschritt pro Jahr signifikant zu steigern. Aktuell beginnen die Fleischrinderzüchter mit der Einführung der genomischen Selektion für Fleischleistungsmerkmale. Für das Merkmal Tageszunahme gehen wir von den folgenden Selektionsparametern aus. Berechnen Sie den Selektionserfolg pro Jahr für die beiden Szenarien mit und ohne genomische Selektion. Welches der beiden Szenarien ist besser basierend auf dem Selektionsfortschritt pro Jahr? Für die Verteilung des Merkmals nehmen wir eine Normalverteilung an.

8

Parameter	No Genomic Selection	With Genomic Selection
proportion of animals selected	0.05	0.05
accuracy of predicted breeding values	0.85	0.65
additive genetic standard deviation	10.00	10.00
generation interval	7.00	2.00

Solution

The proportion animals selected has to be converted into a selection intensity i.

$$i = \frac{z}{p} = \frac{0.103}{0.05} = 2.063$$

The selection response is computed as

$$\Delta G = \frac{i * r_{TI} * \sigma_T}{L}$$

No Genomic Selection:

$$\Delta G_{ng} = \frac{2.063 * 0.85 * 10}{7} = 2.505$$

With Genomic Selection:

$$\Delta G_{gs} = \frac{2.063 * 0.65 * 10}{2} = 6.704$$

Based on the selection response per year, the scenario with genomic selection is better.

c) In cattle and in pigs, the structure of the breeding programs are different. What is the name of the structure of the different breeding programs. Please, specify a reason for the association of the structure of the breeding program to the different species.

In der Rinderzucht und der Schweinezucht ist die Struktur der Zuchtprogramme verschieden. Wie heissen diese Strukturen der Zuchtprogramme? Geben Sie je einen Grund an weshalb die beiden Spezies die entsprechende Struktur des Zuchtprogramms gewählt haben.

6

Solution

	Cattle	Pigs
Structure of breeding program	monolitic	hierarchical
Reason	low reproduction rate, single animal is valuable, sex-limited traits make expensive off-spring tests necessary, hence large herdbooks with many member farms are required	high reproduction rate, single animal not so valuable. Split between breeding and production can be used to increase efficiency

Problem 3: Variance Components Estimation

The following data set with average amounts of methane gas emission per year for three offspring per bull are given in the following table.

Im folgenden Datensatz sind die mittleren Mengen an Methanemissionen pro Jahr für je drei Nachkommen für fünf verschiedene Bullen gegeben.

Offspring	Bull	Methane
1	1	85
2	1	104
3	1	102
4	2	93
5	2	104
6	2	104
7	3	100
8	3	89
9	3	98
10	4	108
11	4	112
12	4	105
13	5	108
14	5	101
15	5	83

a) Given that we want to reduce the amount of methane gas emitted in cattle using the tools of livestock breeding, we first have to do a variance components estimation. Why is it important to have a certain variability in a given trait and which variance component is important when we want to improve a certain trait with livestock breeding tools?

Wir möchten die Methanemission beim Rind mit den Werkzeugen der Tierzucht senken. Weshalb brauchen wir für die züchterische Bearbeitung eines Merkmals ein gewisse Variabilität im Merkmal und welche Varianzkomponente interessiert uns besonders?

2

Solution

One of the main livestock breeding tools is selection. Selection means that we choose from a given population the best individuals to be parents of the next generation. Hence selection can only happen, when a certain level of variation can be observed in a population. Because livestock breeding means improvement of a population at the genetic level, the genetic variance component or in our case the sire variance component is important.

b) What is the linear model that would be used to allows us a separation of the different variance components in the given dataset on methane gas emission?

Wie sieht das lineare Modell aus, welches eine Aufteilung der Varianz in ihre verschiedenen Komponenten im Methandatensatz erlaubt?

11

Solution

The model is

$$y_{ij} = \mu + s_i + e_{ij}$$

where

y_{ij}	measurement j of animal i
μ	expected value of y
s_i	deviation of y_{ij} from μ attributed to bull i
e_{ij}	measurement error

The expected values are: $E[s_i] = 0$, $E[e_{ij}] = 0$, $E[y_{ij}] = \mu$

The variances: $var(s_i) = \sigma_s^2$, $var(e_{ij}) = \sigma_e^2$, $var(y_{ij}) = \sigma_s^2 + \sigma_e^2$

c) Use an analysis of variance (ANOVA) to estimate the variance components given in the above data set.

Schätzen Sie mit einer Varianzanalyse (ANOVA) die Varianzkomponenten aus dem oben angegebenen Datensatz zu den Methanemissionen.

8

Solution

The sires are first converted into factors to prevent R to fit them as covariables.

```
tbl_met$Bull <- as.factor(tbl_met$Bull)
```

Use `aov()` to get the ANOVA-table

```
aov_meth <- aov(Methane ~ Bull, data = tbl_met)
(sy_aov_meth <- summary(aov_meth))
```

```
##             Df Sum Sq Mean Sq F value Pr(>F)
## Bull         4  312.3   78.07   1.077  0.418
## Residuals   10  724.7   72.47
```

From this the estimate of the residual variance corresponds to

```
(n_hat_sigma_e2 <- sy_aov_meth[[1]]$`Mean Sq`[2])
```

```
## [1] 72.46667
```

```
(n_hat_sigma_s2 <- (sy_aov_meth[[1]]$`Mean Sq`[1] - sy_aov_meth[[1]]$`Mean Sq`[2]) / n_nr_dau)
```

```
## [1] 1.866667
```

Problem 4: Prediction of Breeding Values

We are using the same dataset as in Problem 3. The results of the variance components estimation were convincing that it is possible to reduce methane gas emission via selection. The next step is to predict breeding values.

Wir verwenden nochmals den gleichen Datensatz wie in Aufgabe 3. Die Resultate der Varianzkomponenten waren positiv in dem Sinne, dass eine züchterische Bearbeitung des Merkmals Methanemission möglich erscheint.

Offspring	Bull	Methane
1	1	85
2	1	104
3	1	102
4	2	93
5	2	104
6	2	104
7	3	100
8	3	89
9	3	98
10	4	108
11	4	112
12	4	105
13	5	108
14	5	101
15	5	83

a) Specify a sire model by writing down the model formula, describing the meaning of all model components and indicating the expected values and the variances of all random effects in the model. We are using a general mean μ as the only fixed effect. We assume that all bulls are unrelated. Input the numeric information from the dataset into the model components

Spezifizieren Sie für die Zuchtwertschätzung ein Vatermodell. Geben Sie die Modellformel, beschreiben Sie die Bedeutung aller Modellkomponenten und geben Sie für alle zufälligen Komponenten im Modell den Erwartungswert und die Varianz an. Als einziger fixer Effekt wird ein allgemeines Mittel μ angenommen. Die Väter sind nicht miteinander verwandt. Setzen Sie die numerische Information aus dem Datensatz in die Modellkomponenten ein.

19

Solution

The sire model is

$$y = Xb + Zs + e$$

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
s	vector of length t with random sire effects
Z	incidence matrix linking elements in s to observations
e	vector of length n of random error terms

Inserting the data into the model

$$y = \begin{bmatrix} 85 \\ 104 \\ 102 \\ 93 \\ 104 \\ 104 \\ 100 \\ 89 \\ 98 \\ 108 \\ 112 \\ 105 \\ 108 \\ 101 \\ 83 \end{bmatrix}, b = \begin{bmatrix} \mu \end{bmatrix}, X = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}, s = \begin{bmatrix} s_1 \\ s_2 \\ s_3 \\ s_4 \\ s_5 \end{bmatrix}, Z = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}, e = \begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ e_4 \\ e_5 \\ e_6 \\ e_7 \\ e_8 \\ e_9 \\ e_{10} \\ e_{11} \\ e_{12} \\ e_{13} \\ e_{14} \\ e_{15} \end{bmatrix}$$

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

The variances are $var(s) = G = A_s * \sigma_s^2$ where A_s is the numerator relationship matrix for the sires. Because the sires are unrelated $A_s = I$. The variance of the errors is given by $var(e) = R = I * \sigma_e^2$ and $var(y) = ZGZ^T + R$.

b) Set up the mixed model equations for the model specified in 4a) to estimate the fixed effect and to predict the sire effects. We assume the sire variance to be $\sigma_s^2 = 4$ and the error variance to be $\sigma_e^2 = 72$.

Stellen Sie die Mischmodellgleichungen für das Modell aus 4a) auf und berechnen Sie die Schätzung des fixen Effekts und der Vatereffekte.

20

Solution

The mixed model equations using the simplifications from 4a) are given by

$$\begin{bmatrix} X^TX & X^TZ \\ Z^TX & Z^TZ + \lambda * I \end{bmatrix} \begin{bmatrix} \hat{\mu} \\ \hat{s} \end{bmatrix} = \begin{bmatrix} X^Ty \\ Z^Ty \end{bmatrix}$$

where $\lambda = \frac{\sigma_e^2}{\sigma_s^2}$

Inserting the numbers leads to

$$\begin{bmatrix} 15 & 3 & 3 & 3 & 3 & 3 \\ 3 & 21 & 0 & 0 & 0 & 0 \\ 3 & 0 & 21 & 0 & 0 & 0 \\ 3 & 0 & 0 & 21 & 0 & 0 \\ 3 & 0 & 0 & 0 & 21 & 0 \\ 3 & 0 & 0 & 0 & 0 & 21 \end{bmatrix} \begin{bmatrix} \hat{\mu} \\ \hat{s} \end{bmatrix} = \begin{bmatrix} 1496 \\ 291 \\ 301 \\ 287 \\ 325 \\ 292 \end{bmatrix}$$

Solving for the vector of unknowns leads to

$$\begin{bmatrix} \hat{\mu} \\ \hat{s} \end{bmatrix} = \begin{bmatrix} 15 & 3 & 3 & 3 & 3 & 3 \\ 3 & 21 & 0 & 0 & 0 & 0 \\ 3 & 0 & 21 & 0 & 0 & 0 \\ 3 & 0 & 0 & 21 & 0 & 0 \\ 3 & 0 & 0 & 0 & 21 & 0 \\ 3 & 0 & 0 & 0 & 0 & 21 \end{bmatrix}^{-1} \begin{bmatrix} 1496 \\ 291 \\ 301 \\ 287 \\ 325 \\ 292 \end{bmatrix} = \begin{bmatrix} 99.733 \\ -0.390 \\ 0.086 \\ -0.581 \\ 1.229 \\ -0.343 \end{bmatrix}$$