ASMNW - Lösung 5

Peter von Rohr

2017-03-24

Aufgabe 1: RR-BLUP

Im Paper von (Clark and van der Werf 2013) basierend auf der Arbeit von (Meuwissen, Hayes, and Goddard 2001) wurde das sogenannten RR-BLUP Modell vorgeschlagen. In diesem Modell werden die SNP-Effekte an jeder Position für jedes Tier explizit modelliert.

In Matrix-Vektorschreibweise sieht das RR-BLUP Modell wie folgt aus

$$y = 1_n\mu + Wq + e \tag{1}$$

wobei

y	Vektor der Länge n mit phänotypischen Beobachtungen
μ	allgemeines Mittel
q	Vektor der zufälligen additiven Effekte aller SNPs
W	Inzidenzmatrix, welche Genotypen für die SNPscodiert
e	Vektor der Resteffekte

Die SNP-Genotypen in Matrix W sind mit 0, 1 oder 2 codiert in Abhängigkeit der Anzahl an positiven Allelen. Da es sich bei RR-BLUP um eine Methode mit den BLUP-Eigenschaften handelt ist das Modell in (1) ein gemischtes lineares Modell und somit ist q ein Vektor von zufälligen Effekten. Wir müssen für q also auch eine Covarianzmatrix angeben. Diese beträgt $Var(g) = WW^T\sigma_q^2$, wobei σ_q^2 die additiv genetische Varianz der SNPs ist. Bei diesem einfachen RR-BLUP Modell nehmen wir an, dass die Varianz unter den SNPs konstant ist. In der Praxis ist σ_q^2 unbekannt und muss aus den Daten geschätzt werden.

Ihre Aufgabe

Wir nehmen einen kleinen Datensatz mit drei Tieren, welche je eine Beobachtung y haben und fünf SNPs pro Tier. Dieser Datensatz ist in der folgenden Tabelle gezeigt.

	Tier 1	Tier 2	Tier 3
SNP1	G_0G_0	G_1G_1	G_0G_1
SNP2	G_0G_1	G_1G_1	G_0G_1
SNP3	G_0G_0	G_0G_1	G_0G_1
SNP4	G_1G_1	G_0G_1	G_0G_1
SNP5	G_0G_1	G_0G_1	G_0G_0
y	33.43	12.11	29.7

Stellen Sie für den gegebenen Datensatz das RR-Modell aus Gleichung (1) auf. In diesem Datensatz hat das Allel G_0 die positive Wirkung.

Lösung

Die Komponenten μ und q sind unbekannt und müssen aus den Daten geschätzt werden. Da wir jetzt nicht mehr nur einen SNP-Locus sondern 5 Loci betrachten ist q ein Vektor mit fünf Komponenten.

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

Die Matrix W verbindet die genetischen Effekte mit den Tieren. Die Elemente sind codiert als 0, 1 oder 2 je nachdem wie viele Allele mit positiver Wirkung in einem Genotyp enthalten sind. Die Matrix W lautet somit

$$W = \begin{bmatrix} 2.00 & 1.00 & 2.00 & 0.00 & 1.00 \\ 0.00 & 0.00 & 1.00 & 1.00 & 1.00 \\ 1.00 & 1.00 & 1.00 & 1.00 & 2.00 \end{bmatrix}$$

Der Vektor der Beobachtungen y ist definiert als

$$y = \begin{bmatrix} 33.43 \\ 12.11 \\ 29.70 \end{bmatrix}$$

Der Vektor der unbekannten Resteffekte ϵ ist definiert als

$$\epsilon = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \end{bmatrix}$$

Das Gleichungssystem als Ganzes lautet somit

$$\begin{bmatrix} 33.43 \\ 12.11 \\ 29.70 \end{bmatrix} = \begin{bmatrix} 1.00 \\ 1.00 \\ 1.00 \end{bmatrix} \mu + \begin{bmatrix} 2.00 & 1.00 & 2.00 & 0.00 & 1.00 \\ 0.00 & 0.00 & 1.00 & 1.00 & 1.00 \\ 1.00 & 1.00 & 1.00 & 1.00 & 2.00 \end{bmatrix} \begin{bmatrix} q_1 \\ q_2 \\ q_3 \\ q_4 \\ q_5 \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \end{bmatrix}$$

Aufgabe 2: Least Squares nicht möglich

Im Datensatz aus Aufgabe 1 ist die Bedingung für Least Squares nicht erfüllt. Überzeugen Sie sich, dass Least Squares nicht funktioniert, indem Sie versuchen den Datensatz aus Aufgabe 1 mit der Funktion `lm()` in R analysieren.

Hinweise

Die folgenden Vorbereitungen sind für die Analyse mit `lm` nötig. Wir definieren den folgenden Dataframe mit allen SNP und den phänotypischen Beobachtungen

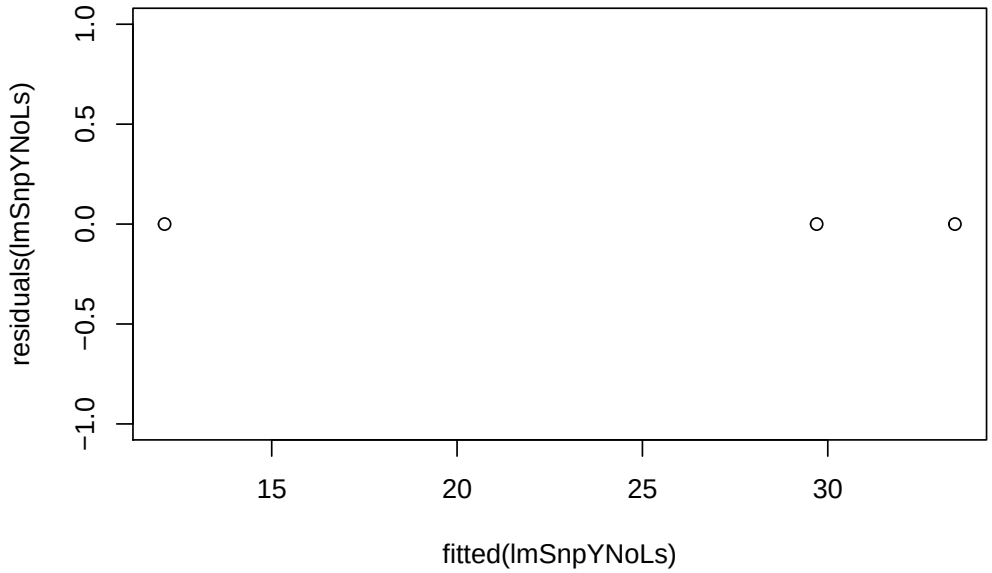
Lösung

Das lineare Modell mit `lm()`

```
##
## Call:
## lm(formula = y ~ ., data = dfSnpYdata)
##
## Residuals:
## ALL 3 residuals are 0: no residual degrees of freedom!
##
## Coefficients: (3 not defined because of singularities)
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    12.11         NA      NA       NA
## snp1            3.73         NA      NA       NA
## snp2           13.86         NA      NA       NA
## snp3              NA         NA      NA       NA
## snp4              NA         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 2 and 0 DF,  p-value: NA
```

In der Zusammenfassung der Resultate gibt R den Hinweise, dass nicht alle Parameter geschätzt werden konnten. Dies wird als `Singularity` bezeichnet. Dies bedeutet, dass die Bedingung $n > p$ für Least Squares nicht erfüllt ist. R gibt aber trotzdem noch für zwei erklärende Variablen Schätzwerte an. Diese haben aber keine biologische Bedeutung und sollten nicht berücksichtigt werden.

Auch der in einer früheren Übung eingeführte Tukey-Anscomb-Plot als Modelldiagnose zeigt, dass mit dem hier angepassten Modell etwas nicht stimmt. Die Residuen sollten zufällig verteilt sein. Hier sind diese aber auf einer Geraden $r_i = 0$ angeordnet. Dies ist ein definitives Zeichen, dass etwas nicht stimmt mit dem angepassten Modell.



Aufgabe 3: GBLUP

Bei RR-BLUP in Aufgabe 1 werden die Genotypen aller SNP für jedes Individuum im Modell (1) berücksichtigt. Im Gegensatz dazu werden bei GBLUP die additiven Effekte aller SNP pro Individuum mit nur einem Effekt modelliert. Unter der Annahme, dass jedes Tier sowohl eine Beobachtung als auch SNP-Informationen aufweist, führt dies zum folgenden Modell

$$y = 1_n\mu + Zg + \epsilon \tag{2}$$

wobei

- y Vektor der Länge n mit phänotypischen Beobachtungen
- μ allgemeines Mittel
- g Vektor der Länge n mit zufälligen additiven SNP-Effekten pro Individuum
- Z Inzidenzmatrix, welche SNP-Effekte mit Beobachtungen verknüpft
- ϵ Vektor von zufälligen Resteffekten

Das Modell (2) ist ein gemischtes lineares Modell mit den zufälligen Effekten g. Die Varianz $Var(g) = G\sigma_g^2$ wobei G der **genomischen Verwandtschaftsmatrix** entspricht und σ_g^2 der genetisch additiven Varianz entspricht. Die genomische Verwandtschaftsmatrix G baisert auf SNP-Informationen der typisierten Individuen und ersetzt die addtiv genetische Verwandtschaftsmatrix A im BLUP-Tiermodell.

Eine Art der Berechnung der genomischen Verwandtschaftsmatrix G ist im Artikel von (Clark and van der Werf 2013) beschrieben und diese soll hier noch etwas genauer erklärt werden.

Hier folgt nun der Abschnitt des R-Codes aus (Clark and van der Werf 2013), welcher die genomische Verwandtschaftsmatrix berechnet. Wir nehmen an, dass das Objekt `data` die SNP-Datenmatrix enthält. Diese entspricht der mit 0, 1, und 2 codierten Genotypen analog zur Matrix W aus dem RR-BLUP Modell in (1).

```
sumpq <- 0
freq <- dim(data)[1]
P <- freq
for(i in 1:ncol(data)){
  (freq[i] <- ((mean(data[,i])/2)))
  (P[i] <- (2*(freq[i]-0.5)))
  (sumpq <- sumpq+(freq[i]*(1-freq[i])))
}
Z <- data
for(i in 1:nrow(data)){
  for(j in 1:ncol(data)){
    (Z[i,j] <- ((data[i,j]-1)-(P[j])))
  }
}
Zt <- t(Z)
ZtZ <- Z%*%Zt
G <- ZtZ/(2*sumpq)
```

Im folgenden Abschnitt wollen wir das obige R-Programm zur Berechnung der genomischen Verwandtschaftsmatrix noch etwas genauer erklären. Das Programm ist mit Ausnahme der Formatierung und dem Ersetzen des Zuweisungsoperators unverändert aus (Clark and van der Werf 2013) übernommen worden. Im Bezug auf den Programmierstil ist das gezeigte Programm nicht optimal, aber das soll hier nicht das Thema sein.

Bei der Berechnung der genomischen Verwandtschaftsmatrix G mit dem obigen R-Programm werden vor dem `for()`-Loop die Variablen `sumpq`, `freq` und `P` initialisiert. Die Werte, welche für die Initialisierung verwendet werden sind bei `freq` und `P` für den weiteren Verlauf nicht wichtig. Die Variable `sumpq` muss zwingend mit 0 initialisiert werden.

Im ersten `for()`-loop werden die Werte der Variablen `freq`, `P` und `sumpq` berechnet. Die Variable `freq` referenziert ein Vektor der Länge der Anzahl SNPs, in welchem die Frequenzen der Allele mit positiver Wirkung gespeichert sind. `P` ist ein Vektor der gleichen Länge wie `freq` und enhält die doppelten Differenzen der Frequenzen in `freq` minus 0.5. Somit sind die Einträge in `P` positive für alle SNP, welche eine Frequenz des Allels mit positiver Wirkung von > 0.5 haben. Bei allen anderen SNPs ist der Eintrag in `P` negativ. In der Variablen `sumpq` werden die Produkte der Frequenzen der beiden Allele an allen SNPs aufaddiert.

Im zweiten `for`-loop wird von der Matrix W aus dem Modell (1), welche unter der Variablen `data` gespeichert ist, bei allen Elementen 1 abgezogen und der Wert aus dem Vektor `P` des entsprechenden SNPs abgezogen. Die so korrigierte Matrix wird mit `Z` bezeichnet. Die genomische Verwandtschaftsmatrix G wird zum Schluss berechnet als $Z * Z^T/(2\sum_j p_j q_j)$ wobei die Variable `sumpq` die summierten Produkte der Allelfrequenzen aller SNPs gespeichert hat.

Ihre Aufgabe

1. Stellen Sie das GBLUP-Modell mit dem Datensatz aus Aufgabe 1 auf
2. Berechnen Sie die genomische Verwandtschaftsmatrix G anhand des oben gezeigten R-Programms aus (Clark and van der Werf 2013)
3. Stellen Sie die Mischmodellgleichungen für das GBLUP-Modell auf und berechnen Sie die Lösungen mit Hilfe des folgenden R-Programms

```
lamda <- ncol(data)
matG <- G
for(i in 1:nrow(G)){
  (matG[i,i] <- ((matG[i,i]+0.01)))
}
# matrix X
matX <- matrix(1,nrow=nAnzTiere,1)
matXtX <- crossprod(matX)
matZ <- diag(1,nAnzTiere)
matXtZ <- crossprod(matX,matZ)
matZtZ <- crossprod(matZ)
matCoeff <- cbind(rbind(matXtX,t(matXtZ)),rbind(matXtZ,matZtZ + lamda * solve(matG)))
vecRhs <- rbind(crossprod(matX,y),crossprod(matZ,y))
vecSol <- solve(matCoeff,vecRhs)
```

Lösung

1. Die Elemente y, 1_n, μ und ϵ sind gleich wie im RR-BLUP Modell ((1)). Die Matrix Z und der Vektor g sind wie folgt definiert.

$$Z = \begin{bmatrix} 1.00 & 0.00 & 0.00 \\ 0.00 & 1.00 & 0.00 \\ 0.00 & 0.00 & 1.00 \end{bmatrix}$$

Der Vektor g enthält die genetischen Effekte pro Individuum über alle SNP.

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

2. Die Genomische Verwandtschaftsmatrix, wie sie nach dem R-Programm aus (Clark and van der Werf 2013) berechnet wird lautet

```r
cat(" * Genomische Verwandtschaftsmatrix G:\n")
```

```
##  * Genomische Verwandtschaftsmatrix G:
```

```r
print(G)
```

```
##            [,1]        [,2]        [,3]
## [1,]  0.9268293 -0.68292683 -0.24390244
## [2,] -0.6829268  0.78048780 -0.09756098
## [3,] -0.2439024 -0.09756098  0.34146341
```

3. Zur Berechnung der Lösung mit GBLUP müssen wir die entsprechenden Mischmodellgleichungen aufstellen. Diese lauten

$$\begin{bmatrix} 3.00 & 1.00 & 1.00 & 1.00 \\ 1.00 & 170.48 & 166.65 & 163.87 \\ 1.00 & 166.65 & 171.41 & 162.95 \\ 1.00 & 163.87 & 162.95 & 174.18 \end{bmatrix} \begin{bmatrix} \hat{\mu}_1 \\ \hat{g}_1 \\ \hat{g}_2 \\ \hat{g}_3 \end{bmatrix} = \begin{bmatrix} 75.24 \\ 33.43 \\ 12.11 \\ 29.70 \end{bmatrix}$$

Der Lösungsvektor lautet

$$\begin{bmatrix} \hat{\mu}_1 \\ \hat{g}_1 \\ \hat{g}_2 \\ \hat{g}_3 \end{bmatrix} = \begin{bmatrix} 25.08 \\ 2.34 \\ -2.56 \\ 0.22 \end{bmatrix}$$

Abkürzungen

Abbreviation	Meaning
BLUP	Best Linear Unbiased Prediction
RR	Ridge Regression (Random Regression)
GBLUP	Genomic Best Linear Unbiased Prediction

References

Clark, Samuel A., and Julius van der Werf. 2013. "Genomic Best Linear Unbiased Prediction (GBLUP) for the Estimation of Genomic Breeding Values." In *Genome-Wide Association Studies and Genomic Prediction, Methods in Molecular Biology, Vol 1019*, edited by Cedric Gondro, Julius van der Werf, and Ben Hayes. Springer. doi:10.1007/978-1-62703-447-0_13.

Meuwissen, Theo HE, Ben J Hayes, and Mike E Goddard. 2001. "Prediction of Total Genetic Value Using Genome-Wide Dense Marker Maps." *Genetics*, no. 157: 1819–29.