ASMNW - Lösung 4

Peter von Rohr

2017-03-17

Aufgabe 1: Einfaches Beispiel mit nur einem SNP

Wir betrachten ein einfaches Beispiel mit nur 10 Tieren, welche für einen einzigen SNP typisiert sind. Die folgende Tabelle gibt eine Übersicht über die Daten mit den SNP-Allelen und den phänotypischen Beobachtungen.

Animal	Phentype	SnpAllele1	SnpAllele2
1	2.03	1	1
2	3.54	1	2
3	3.83	1	2
4	4.87	2	2
5	3.41	1	2
6	2.34	1	1
7	2.65	1	1
8	3.76	1	2
9	3.69	1	2
10	3.69	1	2

Wir nehmen an die Tiere seien nicht verwandt miteinander. Somit können wir die Beziehung zwischen dem einen SNP und den phänotypischen Beobachtungen mit einem einfachen Regressionsmodell testen. Unser Modell lautet:

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

wobei

- y Vektor der Länge n mit phänotypischen Beobachtungen
- μ allemeines Mittel, welches fixe Effekte repräsentiert
- 1_n Vektor der Länge n mit lauter Einsen
- g additiver Effekt des Marker-SNP
- W Inzidenzmatrix, welche die Beobachtungen zum Marker-Effekt verbindet
- e Vektor der zufälligen Resteffekte

Die Inzidenzmatrix W hat n Zeilen und so viele Kolonnen, wie SNP-Marker. Für unser Beispiel hat die Matrix W somit 1 Kolonne. Die Elemente der Matrix W zählen die Anzahl Allele mit positiver Wirkung. In diesem Beispiel sei das Allel "2".

Ihre Aufgabe

Stellen Sie das Modell aus Gleichung (1) für den gegebenen Datensatz auf und bestimmen Sie welche Modellkomponenten bekannt und welche unbekannt sind.

Lösung

Die Komponenten des Modells in Gleichung (1) lauten wie folgt

- Parameter μ und g sind unbekannt und müssen aus den Daten geschätzt werden.
- Die Resteffekte e sind unbekannt. Deren Varianz σ^2 muss aus den Daten geschätzt werden
- Die anderen Modellkomponenten y und W sind bekannt und sind wie folgt definiert

$$y = \begin{bmatrix} 2.03 \\ 3.54 \\ 3.83 \\ 4.87 \\ 3.41 \\ 2.34 \\ 2.65 \\ 3.76 \\ 3.69 \\ 3.69 \end{bmatrix}$$

W ist eigentlich eine Matrix, aber da wir nur einen SNP anschauen, reduziert sie sich auf einen Vektor.

$$W = \begin{bmatrix} 0 \\ 1 \\ 1 \\ 2 \\ 1 \\ 0 \\ 0 \\ 1 \\ 1 \\ 1 \end{bmatrix}$$

Aufgabe 2: Least Squares Lösungen

Da wir 10 Beobachtungen und nur einen SNP betrachten ist die Bedingung für die Schätzung der unbekannten Parameter mit Least Squares erfüllt. Somit können wir die unbekannten Parameter μ und g mit der folgenden Gleichung schätzen.

$$\begin{bmatrix} \hat{\mu} \\ \hat{g} \end{bmatrix} = \begin{bmatrix} 1_n^T 1_n & 1_n^T W \\ W^T 1_n & W^T W \end{bmatrix}^{-1} \begin{bmatrix} 1_n^T y \\ W^T y \end{bmatrix} \tag{2}$$

Berechnen Sie aufgrund der Gleichungen in (2) die Lösungen für $\hat{\mu}$ und $\hat{g}$.

Lösung

Mit den oben angegebenen Komponenten y und W lautet das Gleichungssystem (2)

$$\begin{bmatrix} \hat{\mu} \\ \hat{g} \end{bmatrix} = \begin{bmatrix} 10.00 & 8.00 \\ 8.00 & 10.00 \end{bmatrix}^{-1} \begin{bmatrix} 33.81 \\ 31.66 \end{bmatrix} = \begin{bmatrix} 2.36 \\ 1.28 \end{bmatrix}$$

Aufgabe 3

Überprüfen Sie die unter Aufgabe 2 erhaltenen Least Squares Lösungen für μ und g mit der Funktion `lm()` R.

Hinweise

- Lesen Sie die Daten aus der in Aufgabe 1 gezeigten Tabelle in den Dataframe namens `dfAufgabe3` ein.

```
nAnzTiere <- 10
dfAufgabe3 <- data.frame(
  Animal = c(1:nAnzTiere),
  Phentype = c(2.03, 3.54,3.83,4.87,3.41,2.34,2.65,3.76,3.69,3.69),
  SnpAllele1 = c(1,1,1,2,1,1,1,1,1,1),
  SnpAllele2 = c(1,2,2,2,2,1,1,2,2,2))
```

- Fügen Sie dem Dataframe eine zusätzliche Kolonne namens `Genotype`hinzu, welche die Genotypen-Codes enthält. Diese Codes entsprechen den Anzahl an "2" Allelen mit positiver Wirkung.
- Verwenden Sie die phänotypischen Beobachtungen in `dfAufgabe3$Phenotype` als Zielgrösse und `dfAufgabe3$Genotype` als erklärende Variable und passen Sie das Regressionsmodell mit der Funktion `lm()` an.

Lösung

Aus den Alleleinformationen im gegebenen Datenframe, bilden wir zuerst den Vektor der Genotypen-Codes und fügen diese zum Dataframe als zusätzliche Kolonne hinzu.

```
Genotype <- dfAufgabe3$SnpAllele1 + dfAufgabe3$SnpAllele2 - 2
dfAufgabe3 <- cbind(dfAufgabe3, Genotype)
```

Jetzt passen wir das Regressionsmodell an und berechnen die Least Squares Schätzungen

```
lmSnp <- lm(dfAufgabe3$Phentype ~ dfAufgabe3$Genotype, data = dfAufgabe3)
summary(lmSnp)
```

```
##
## Call:
## lm(formula = dfAufgabe3$Phentype ~ dfAufgabe3$Genotype, data = dfAufgabe3)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.32611 -0.08500  0.01833  0.10528  0.29389
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)           2.3561     0.1045   22.56 1.58e-08 ***
## dfAufgabe3$Genotype   1.2811     0.1045   12.27 1.81e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1982 on 8 degrees of freedom
## Multiple R-squared:  0.9495, Adjusted R-squared:  0.9432
## F-statistic: 150.4 on 1 and 8 DF,  p-value: 1.815e-06
```