Züchtungslehre - Varianzkomponentenschätzung

Peter von Rohr

2016-11-25

Einleitung

Im Kapitel BLUP-Zuchtwertschätzung haben wir die Varianzkomponenten σ_e^2 und σ_a^2 als bekannt angenommen. In der praktischen Zuchtarbeit sind diese unbekannt und müssen aus den Daten geschätzt werden. Der Prozess, welcher aus beobachteten Daten Varianzparameter für ein bestimmtes Modell liefert wird als **Varianzkomponentenschätzung** bezeichnet. In gewissen Studiengängen ist die Varianzkomponentenschätzung das Thema einer ganzen Vorlesung. Wir versuchen hier einen ersten Einblick in dieses Thema in einer Woche zu bekommen.

Regression und Least Squares

In einem klassischen Regressionsmodell (1) werden die fixen Effekte mit `Least Squares` geschätzt.

$$y = Xb + e \tag{1}$$

Unter der Annahme, dass die Matrix X vollen Kolonnenrang p hat, entspricht die Least Squares-Schätzung $\hat{b}$ für die fixen Effekte b

$$\hat{b} = (X^T X)^{-1} X^T y \tag{2}$$

Die Least-Squares Prozedur an sich liefert keine Schätzung $\hat{\sigma_e}^2$ für die Restvarianz σ_e^2. Häufig wird eine Schätzung $\hat{sigma_e}^2$ basierend auf den Residuen $r_i = y_i - x_i^T \hat{b}$ verwendet. Dieser Schätzer für σ_e^2 lautet

$$\hat{\sigma_e}^2 = \frac{1}{n-p} \sum_{i=1}^{n} r_i^2 \tag{3}$$

Die Residuen r_i sind plausible Schätzungen für die Reste e_i. Somit ist der Schätzer für die Restvarianz plausibel bis auf den Faktor $\frac{1}{n-p}$. Dieser Faktor macht den Schätzer in (3) erwartungstreu, was bedeutet, dass $E\left[\hat{\sigma_e}^2\right] = \sigma_e^2$.

Varianzanalyse

Ursprünglich wurde die Varianzanalyse entwickelt um globale Unterschiede zwischen fixen Effektstufen unter gewissen Unsicherheitsfakturen, wie Messfehler oder anderen Einflüssen zu testen. In einer späteren Entwicklung wurde die Varianzanalyse angepasst für die Schätzung von Varianzkomponenten in Modellen mit zufälligen Effekten.

Globale Tests von Effekten

Wollen wir zum Beispiel wissen ob das Geschlecht in unserem bekannten Datensatz mit den Zunahmen seit dem Absetzen überhaupt einen Einfluss hat, können wir das mit einer Varianzanalyse überprüfen. Wir schauen uns dazu den reduzierten Datensatz an und betrachten einmal nur einen allfälligen Einfluss des Geschlechts auf die Zunahmen.

Kalb	Geschlecht	WWG
4	M	4.5
5	F	2.9
6	F	3.9
7	M	3.5
8	M	5.0

Zum oben gezeigten reduzierten Datensatz betrachten wir das fixe Modell (d.h. ein Modell mit nur fixen Effekten), in welchem nur das Geschlecht als Einflussfaktor auf die Zunahmen (WWG) modelliert wird. Wir haben also

$$y = Xb + e \tag{4}$$

wobei der Vektor b die zwei Effektstufen für das Geschlecht enthält. Somit ist

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

Der globale Test, ob das Geschlecht überhaupt einen Einfluss hat, entspricht der Nullhypothese $H_0: \ b_F = b_M = 0$. Die geschätzten Effekte können wir mit Least-Squares berechnen. Angenommen, dass unsere Daten in einem Dataframe namens `dfWwgRed` gespeichert sind, sieht die Least-Squares-Schätzung In R folgendermassen aus.

```
##
## Call:
## lm(formula = WWG ~ -1 + Geschlecht, data = dfWwgRed)
##
## Residuals:
##       1       2       3       4       5
##  0.1667 -0.5000  0.5000 -0.8333  0.6667
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## GeschlechtF   3.4000     0.5270   6.451  0.00755 **
## GeschlechtM   4.3333     0.4303  10.070  0.00209 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7454 on 3 degrees of freedom
## Multiple R-squared:  0.9795, Adjusted R-squared:  0.9658
## F-statistic: 71.51 on 2 and 3 DF,  p-value: 0.002945
```

Die Tabelle der Varianzanalyse zur Überprüfung der globalen Nullhypothese $H_0: \ b_F = b_M = 0$ erhalten wir mit

```
##            Df Sum Sq Mean Sq F value  Pr(>F)
## Geschlecht  2  79.45   39.73   71.51 0.00294 **
## Residuals   3   1.67    0.56
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Die Test-Statistik für unseren globalen Test entnehmen wir der Spalte, welche mit `F-value` überschrieben ist. Den gleichen Wert hatten wir schon bei den Resultaten der Funktion `lm()` gefunden. Die sehr tiefe Irrtumswahrscheinlichkeit ($Pr(> |t|) = 0.00294$) bedeutet, dass wir bei einer Ablehnung der globalen Nullhypothese $H_0 : \ b_F = b_M = 0$ nur mit einer sehr tiefen Wahrscheinlichkeit einen Fehler erster Art begehen würden. Die Summenquadrate (Sum Sq) berechnen wir gemäss

$$SSQ_T = \sum_{i=1}^{n} y_i^2 \tag{5}$$

Die Summenquadrate der Residuen (Residuals) entspricht der Summe der quadrierten Residuen.

$$SSQ_R = \sum_{i=1}^{n} r_i^2 \tag{6}$$

wobei $r_i = y_i - \hat{y}_i = y_i - x_i^T b$. Die Summenquadrate des Geschlechts SSQ_b entsprechen der Summe der quadrierten gefitteten Werte $\hat{y}_i = x_i^T b$. Die Summenquadrate SSQ_b sind auch gleich der Differenz zwischen SSQ_T und SSQ_R.

$$SSQ_b = \sum_{i=1}^{n} \hat{y}_i^2 = \sum_{i=1}^{n} (x_i^T b)^2 \tag{7}$$

Die mittleren Summenquadrate (abgekürzt MSQ, wird im R-output mit `Mean Sq` bezeichnet) berechnen sich aus dem Verhältnis der Summenquadrate durch die Anzahl Freiheitsgrade (df). In diesem einfachen Bespiel entspricht die totale Anzahl an Freiheitsgraden (df_T) der Anzahl Beobachtungen (n) minus 1. Die Anzahl Freiheitsgrade df_b für das Geschlecht entspricht den Anzahl Faktorstufen, somit ist $df_b = 2$. Die Anzahl Freiheitsgrade der Residuen df_e ist dann die Differenz zwischen df_T und df_b.

Das Verhältnis der mittleren Summenquadrate des Modells (Geschlecht) und der mittleren Summenquadrate der Residuen (Residuals) definiert eine Teststatistik F. Unter der globalen Nullhypothese folgt die Teststatistik F einer F-Verteilung mit df_b und df_R Freiheitsgraden. Aus dieser Verteilung lässt sich dann die Irrtumswahrscheinlichkeit $Pr(> |t|)$ ableiten.

Schätzung einer Varianzkomponente

Lineare Modelle, welche neben den Resteffekten auch noch weitere zufällige Effekte aufweisen werden häufig als "zufällige Modelle" (random models) bezeichnet. Als Beispiel eines zufälligen Models können wir wieder den Datensatz mit den Zunahmen anschauen. In dieser Variante betrachten wir aber nur den Effekt der Väter auf die Zunahmen und ignorieren den Geschlechtseinfluss. Wir haben also einen reduzierten Datensatz, wo nur die Vätereffekte vorkommen. Damit wir die später geforderte Unabhängigkeit der Vatereffekte in unserem Datensatz nicht verletzen, weisen wir Kalb 7 den Vater 3 an Stelle vom aktuellen Vater 4 zu.

Kalb	Vater	WWG
4	1	4.5
5	3	2.9
6	1	3.9
7	3	3.5
8	3	5.0

Das Modell (8), welches die Beobachtungen als Funktion der zufälligen Vatereffekte darstellt sieht algebraisch ähnlich aus wie das fixe Modell in (4). Beim fixen Modell (4) sind die einzelnen Stufen b_i fix und es wurden

alle möglichen Faktorstufen des Geschlechts berücksichtigt. Hingegen in (8) steht u_i für den Effekt vom Vater i und Vater i ist einfach ein Vater aus einer sehr grossen Population von Vätern. Die Väter, welche im Vektor u berücksichtigt sind, entsprechen einer zufälligen Auswahl aus der Population von Vätern.

$$y = Zu + e \tag{8}$$

Die Väter in (8) sind also charakteristisch für zufällige Effekte. Trotzdem, dass wir nur eine zufällige Stichprobe an Vätern kennen, möchten wir doch Aussagen zur ganzen Population machen. Da die Beiträge der Väter als zufällige Effekte modelliert werden, entspricht der einzelne Vatereffekt u_i einer Zufallsvariablen, welcher wir eine Dichteverteilung zuordnen. Zwei Eigenschaften von Dichteverteilungen, welche wir im Zusammenhang mit zufälligen Effekten häufig postulieren sind

1. die zufälligen Effekte u_i sind unabhängig voneinander. In unserem Beispiel trifft das nur zu, wenn die Väter nicht verwandt sind miteinander.
2. der Erwartungswert der zufälligen Effekte u_i ist 0 und die zufälligen Effekte haben alle die gleiche Varianz σ_u^2.

Für das zufällige Modell müssen wir also abgesehen von den Effekten auch noch die Varianzkomponenten σ_u^2 und σ_e^2 schätzen. Eine Möglichkeit zu Schätzungen für die Varianzkomponenten zu gelangen ist über die Varianzanalyse. Es kann gezeigt werden, dass die Erwartungswerte der mittleren Summenquadrate als Funktionen der unbekannten Varianzkomponenten dargestellt werden können. Spezifisch für unser Modell kann gezeigt werden, dass der Erwartungswert der mittleren Summenquadrate der Resteffekte gleich der Restvarianz ist. Es gilt also

$$E\left[MSQ_e\right] = \sigma_e^2 \tag{9}$$

Für die Varianzkomponente der Vatereffekte können wir die erwarteten mittleren Summenquadrate einer Funktion aus σ_u^2 und σ_e^2 gleichsetzen.

$$E\left[MSQ_u\right] = n\sigma_u^2 + \sigma_e^2 \tag{10}$$

Wir setzen nun die empirischen Werte der mittleren Summenquadrate gleich den Erwartungswerten und verwenden diese als Schätzer für die Varianzkomponenten. Somit erhalten wir

$$MSQ_e = \widehat{\sigma_e^2} \tag{11}$$

und

$$MSQ_u = n\widehat{\sigma_u^2} + \widehat{\sigma_e^2} \tag{12}$$

Lösen wir (12) nach $\widehat{\sigma_u^2}$ auf, so erhalten wir als Schätzer für die Varianzkomponenten der Vatereffekte

$$\widehat{\sigma_u^2} = \frac{MSQ_u - MSQ_e}{n} \tag{13}$$

Unser Beispiel

Die folgenden Anweisungen in `R` zeigen, wie die Varianzanalysentabelle für unser zufälliges Modell aufgestellt wird.

```r
aovWwgSire <- aov(formula = WWG ~ Vater, data = dfWwgSire)
summary(aovWwgSire)
```

```
##             Df Sum Sq Mean Sq F value Pr(>F)
## Vater        1  0.192   0.192   0.229  0.665
## Residuals    3  2.520   0.840
```

Aufgrund von Gleichung (11) erhalten wir eine Schätzung für die Restvarianz als

$$\widehat{\sigma_e^2} = 0.84$$

Setzen wir diese Schätzung in Gleichung (13) ein, dann erhalten wir

$$\widehat{\sigma_u^2} = -0.344$$

Negative Schätzwerte

Der Schätzwert für die Varianzkomponente σ_u^2 ist negativ. Dies ist durch die spezielle Datenkonstellation verursacht. Aufgrund von nur 3 Beobachtungen können keine zuverlässigen Varianzkomponenten geschätzt werden. Die Methode der Varianzanalyse hat keinen Mechanismus zur Verfügung, welcher negative Schätzwerte verhindern könnte.

Varianzkomponenten sind als Quadrate definiert und somit können diese nicht negativ sein. Da aber aufgrund von Datenkonstellationen die Schätzungen negativ sein können, ist die Methode der Varianzanalyse für die Varianzkomponentenschätzung nicht sehr beliebt.

Im nächsten Kapitel werden wir uns alternative Verfahren zur Schätzung von Varianzkomponenten anschauen.