Variansanalyse og lineær regresjon – notat til STK2120

Ørulf Borgan februar 2016

Formålet med dette notatet er å beskrive sammenhengen mellom variansanalyse med faste effekter og multippel lineær regresjon og å diskutere hvordan `R` kan brukes til å tilpasse modeller i variansanalysen. Notatet er et supplement til det som står om variansanalyse i kapittel 11 i boka til Devore & Berk (D&B). Når ikke annet er sagt bruker vi notasjonen i denne boka.

1. Modellene

Vi begynner med å oppsummere de modellene vi skal se nærmere på.

Enveis variansanalyse

Modellen for enveis variansanalyse kan gi som (jf. avsnitt 11.3 i D&B)

$$X_{ij} = \mu + \alpha_i + \epsilon_{ij}\,; \quad j = 1, 2, \ldots, J_i\,; \quad i = 1, 2, \ldots, I\,; \tag{1}$$

der $\sum_{i=1}^{I} \alpha_i = 0$ og ϵ_{ij}-ene er uavhengige og $N(0, \sigma^2)$-fordelte. Vi lar $n = \sum_{i=1}^{I} J_i$ betegne det totale antall observasjoner.

Toveis variansanalyse

Modellen for toveis variansanalyse kan gi som (jf. avsnitt 11.5 i D&B)

$$X_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \epsilon_{ijk}\,; \quad k = 1, \ldots, K_{ij}\,; \quad j = 1, \ldots, J\,; \quad i = 1, \ldots, I\,; \tag{2}$$

der $\sum_{i=1}^{I} \alpha_i = \sum_{j=1}^{J} \beta_j = \sum_{i=1}^{I} \gamma_{ij} = \sum_{j=1}^{J} \gamma_{ij} = 0$ og ϵ_{ijk}-ene er uavhengige og $N(0, \sigma^2)$-fordelte. Vi lar $n = \sum_{i=1}^{I} \sum_{j=1}^{J} K_{ij}$ betegne det totale antall observasjoner.

Hvis alle γ_{ij}-ene i (2) er lik null, har vi en modell uten interaksjon (samspill); jf. avsnitt 11.4 i D&B. Merk at vi ikke forutsetter at det er like mange observasjoner for hver kombinasjon av nivåene av faktorene A og B, slik tilfellet er i avsnitt 11.5 i D&B.

Multippel lineær regresjon

Modellen for multippel lineær regresjon er gitt ved (jf. avsnitt 12.7 i D&B)

$$Y_l = \theta_0 + \theta_1 x_{l1} + \cdots + \theta_k x_{lk} + \epsilon_l\,; \quad l = 1, 2, \ldots, n\,; \tag{3}$$

der $x_{l1}, x_{l2}, \ldots, x_{lk}$ er kjente tall og ϵ_l-ene er uavhengige og $N(0, \sigma^2)$-fordelte.

I avsnitt 12.7 i D&B kalles regresjonskoeffisientene $\beta_0, \beta_1, \ldots \beta_k$. I dette notetet har vi valgt å kalle regresjonskoeffisientene $\theta_0, \theta_1, \ldots \theta_k$ for bedre å kunne skille dem fra β_j-ene i (2). Videre bruker vi indeksen l for å angi individ i stedet for i som brukes i avsnitt 12.7 i D&B for å unngå sammenblanding med indeksen i i (1) og (2).

2. Enveis variansanalyse

Vi ser så på modellen (1) for enveis variansanalyse. I denne modellen angir X_{ij} observasjonen for individ j i gruppe i. Vi vil skrive modellen som en lineær regresjonsmodell, dvs. på formen (3). I (3) angir Y_l observasjonen for individ l, mens $x_{l1}, x_{l2}, \ldots, x_{lk}$ er forklaringvariable som

angir ulike egenskaper ved individet.

For å skrive (1) på formen (3) ser vi først på sammenhengen mellom X_{ij} - ene og Y_l - ene. Vi får denne sammenhengen ved å la de J_1 første Y_l - ene være observasjonene fra gruppe 1, de neste J_2 Y_l - ene være observasjonen fra gruppe 2, osv. Mer presist har vi at (husk at $n = \sum_{i=1}^{I} J_i$):

$$\begin{array}{cccc} Y_1 = X_{11} & Y_2 = X_{12} & \cdots & Y_{J_1} = X_{1J_1} \\ Y_{J_1+1} = X_{21} & Y_{J_1+2} = X_{22} & \cdots & Y_{J_1+J_2} = X_{2J_2} \\ Y_{J_1+J_2+1} = X_{31} & Y_{J_1+J_2+2} = X_{32} & \cdots & Y_{J_1+J_2+J_3} = X_{3J_3} \\ \cdots & \cdots & \cdots & \cdots \\ Y_{n-J_I+1} = X_{I1} & Y_{n-J_I+2} = X_{I2} & \cdots & Y_n = X_{IJ_I} \end{array} \tag{4}$$

For å angi hvilken gruppe en observasjon Y_l hører til, må vi innføre passende forklaringsvariable. Før vi gjør det, merker vi oss at restriksjonen $\sum_{i=1}^{I} \alpha_i = 0$ gjør at bare $I - 1$ av α_i - ene kan variere fritt. Vi vil la de $I - 1$ første α_i - ene variere fritt, mens den siste er gitt ved

$$\alpha_I = -\sum_{i=1}^{I-1} \alpha_i \tag{5}$$

For gruppe I kan derfor (1) gis som

$$X_{Ij} = \mu - \sum_{i=1}^{I-1} \alpha_i + \epsilon_{Ij}\,; \quad j = 1, 2, \ldots, J_I \tag{6}$$

Vi innfører nå forklaringsvariablene

$$x_{li} = \begin{cases} 1 & \text{hvis individ } l \text{ hører til gruppe } i \\ -1 & \text{hvis individ } l \text{ hører til gruppe } I \\ 0 & \text{ellers} \end{cases} \tag{7}$$

for $i = 1, \ldots, I - 1$. Hvis vi setter disse inn i formelen (3) for lineær regresjon (med $k = I - 1$), får vi

$$Y_l = \theta_0 + \theta_i + \epsilon_l \tag{8}$$

hvis individ l hører til gruppe i der $i = 1, \ldots, I - 1$, mens

$$Y_l = \theta_0 - \theta_1 - \theta_2 - \cdots - \theta_{I-1} + \epsilon_l \tag{9}$$

hvis individ l hører til gruppe I. Ved å sammenholde (8) med (1) og (9) med (6), ser vi at de to modellene er like når vi lar parametrene $\mu, \alpha_1, \ldots, \alpha_{I-1}$ i (1) svare til parametrene $\theta_0, \theta_1, \ldots, \theta_{I-1}$ i (8) og (9).[Sammenhengen mellom ϵ_{ij} - ene i (1) og ϵ_l - ene i (8) og (9) er tilsvarende som sammenhengen mellom X_{ij} - ene og Y_j - ene i (4).]

Vi har dermed vist at ved å innføre forklaringsvariablene (7) kan modellen (1) for enveis variansanalyse skrives som en multippel lineær regresjonsmodell (3).

3. Bruk av R for enveis variansanalyse

For å vise hvordan vi kan bruke R for enveis variansanalyse, ser vi på eksempel 11.5 i D&B. Eksempelet gjelder en studie der $n = 20$ rotter ble delt inn i $I = 4$ grupper med $J = 5$ rotter i hver gruppe. Rottene fikk en viss mengde etanol avhengig av hvilken gruppe de var i, og en målte lengden av REM søvn de neste 24 timene. Se side 568 i D&B for en mer utførlig beskrivelse av eksempelet.

Vi leser dataene inn i en dataramme som vi kaller `exmp11.5` slik det er beskrevet på kurssiden. Dataene er som følger:

```
> exmp11.5
   sleep treat
1   88.6     1
2   73.2     1
3   91.4     1
4   68.0     1
5   75.2     1
6   63.0     2
7   53.9     2
8   69.2     2
9   50.1     2
10  71.5     2
11  44.9     3
12  59.5     3
13  40.2     3
14  56.3     3
15  38.7     3
16  31.0     4
17  39.6     4
18  45.3     4
19  25.2     4
20  22.7     4
```

Ovenfor og i det følgende skriver vi `>` først på en linje for å markere det vi selv skriver inn i R, mens linjer uten `>` angir utskrift fra R.

For å gjøre en enveis variansanalyse kan vi bruke kommandoen `aov`:

```
> fit.aov=aov(sleep~factor(treat),data=exmp11.5)
```

Da få vi variansanalysetabellen (jf. side 569 i D&B):

```
> summary(fit.aov)
             Df Sum Sq Mean Sq F value   Pr(>F)
factor(treat)  3   5882    1961   21.09 8.32e-06
Residuals     16   1487      93
```

For å få estimater for μ og α_i-ene i (1) kan vi ikke bruke `aov`-kommandoen. Da må vi i stedet bruke `lm`-kommandoen for lineær regresjon. Men før vi bruker denne kommandoen til å gjøre en enveis variansanalyse, må vi gi R beskjed om at α_i-ene i (1) skal tilfredsstille restriksjonen $\sum_{i=1}^{I} \alpha_i = 0$. Det gjør vi ved kommandoen:

```
> options(contrasts=c("contr.sum","contr.poly"))
```

Vi kan så tilpasse modellen (1) med kommandoen:

```
> fit.lm=lm(sleep~factor(treat),data=exmp11.5)
```

Da får vi resultatet:

```
> summary(fit.lm)
               Estimate Std. Error t value Pr(>|t|)
(Intercept)      55.375      2.156  25.685 1.96e-14
factor(treat)1   23.905      3.734   6.402 8.77e-06
factor(treat)2    6.165      3.734   1.651   0.1182
factor(treat)3   -7.455      3.734  -1.996   0.0632

Residual standard error: 9.642 on 16 degrees of freedom
Multiple R-squared: 0.7982,     Adjusted R-squared: 0.7603
F-statistic: 21.09 on 3 and 16 DF,  p-value: 8.325e-06
```

Merk at det ovenfor er gitt estimater for μ og de tre første α_i-ene. Estimatet for den siste α_i-en får vi ved å bruke relasjonen (5) med $I = 4$.

Etter at vi har tilpasset en variansanalysemodell med `lm`-kommandoen, kan vi få variansanalysetabellen med `anova`-kommandoen:

```
> anova(fit.lm)
              Df Sum Sq Mean Sq F value    Pr(>F)
factor(treat)  3 5882.4 1960.79  21.092 8.325e-06
Residuals     16 1487.4   92.96
```

Vi ser at dette er den samme tabellen som vi fikk med `aov`-kommandoen.

I `lm`-kommandoen (og `aov`-kommandoen) skrev vi `factor(treat)` for å gi `R` beskjed om at `treat` er en kategorisk variabel. Det som skjer når vi bruker `factor(treat)` i `lm`-kommandoen er at `R` lager 3 forklaringsvariable slik det er gitt i (7). Du kan gi kommandoen `model.matrix(fit.lm)` for å se at det er dette `R` gjør.

4. Toveis variansanalyse uten interaksjon

Vi tar så for oss modellen (2) for toveis variansanalyse. På lignende måte som i (4) skriver vi de $n = \sum_{i=1}^{I} \sum_{j=1}^{J} K_{ij}$ observasjonene X_{ijk} som $Y_1, \ldots, Y_n$. Vi kan for eksempel først ta de K_{11} observasjonen på første nivå av både faktor A og B, deretter de K_{12} observasjonen på første nivå faktor A og andre nivå av faktor B, osv.

For å se hvordan modellen for toveis variansanalyse kan skrives som en lineær regresjonsmodell, ser vi for enkelhets skyld først på situasjonen uten interaksjon. Når det ikke er interaksjon, er $\gamma_{ij} = 0$ i (2) for alle i og j. Modellen for toveis variansanalyse blir da:

$$X_{ijk} = \mu + \alpha_i + \beta_j + \epsilon_{ijk}\,; \quad k = 1, \ldots, K_{ij}\,; \quad j = 1, \ldots, J\,; \quad i = 1, \ldots, I\,; \tag{10}$$

For å skrive (10) som en lineær regresjonsmodell innfører vi $I-1$ forklaringsvariable for faktor A:

$$x_{li} = \begin{cases} 1 & \text{hvis individ } l \text{ har nivå } i \text{ for faktor } A \\ -1 & \text{hvis individ } l \text{ har nivå } I \text{ for faktor } A \\ 0 & \text{ellers} \end{cases} \qquad (i = 1, \ldots, I-1) \tag{11}$$

og $J-1$ forklaringsvariable for faktor B:

$$x_{l,I-1+j} = \begin{cases} 1 & \text{hvis individ } l \text{ har nivå } j \text{ for faktor } B \\ -1 & \text{hvis individ } l \text{ har nivå } J \text{ for faktor } B \\ 0 & \text{ellers} \end{cases} \qquad (j = 1, \ldots, J-1) \qquad (12)$$

Hvis vi setter disse forklaringsvariablene inn i (3) med $k = I + J - 2$, får vi ved et tilsvarende resonnement som for enveis variansanalyse [jf. (8) og (9)] at den lineære regresjonsmodellen (3) er lik variansanalysemodellen (10). Sammenhengen mellom parametrene i variansanalysemodellen og den lineære regresjonsmodellen er som følger:

$$\begin{aligned} \mu &= \theta_0 \\ \alpha_i &= \theta_i \ ; \qquad i = 1, \ldots, I-1 \\ \beta_j &= \theta_{I-1+j} \ ; \qquad j = 1, \ldots, J-1 \end{aligned}$$

Vi har dermed vist at ved å innføre forklaringsvariablene (11) og (12) kan modellen (10) for toveis variansanalyse uten interaksjon skrives som en multippel lineær regresjonsmodell (3).

5. Bruk av `R` for toveis variansanalyse uten interaksjon

For å vise hvordan vi kan bruke `R` for toveis variansanalyse uten interaksjon, ser vi på eksempel 11.11 i D&B. Eksempelet gjelder en studie der en har undersøkt hvor lett det er å fjerne merker på tøy fra $I = 3$ ulike merkepenner og hvordan dette avhenger av $J = 4$ forskjellige måter å vaske på. Hver penn ble prøvd $K = 1$ gang for hver vaskemåte og det ble målt hvor mye av merket som ble igjen etter vasken (lav verdi svarer til at lite av merket ble igjen). Se side 583 i D&B for en mer utførlig beskrivelse av eksempelet.

Vi leser dataene inn i en dataramme som vi kaller `exmp11.11` slik det er beskrevet på kurssiden. Dataene er som følger:

```
> exmp11.11
   Response Treat Brand
1      0.97     1     1
2      0.77     1     2
3      0.67     1     3
4      0.48     2     1
5      0.14     2     2
6      0.39     2     3
7      0.48     3     1
8      0.22     3     2
9      0.57     3     3
10     0.46     4     1
11     0.25     4     2
12     0.19     4     3
```

For å få estimater for μ, α_i-ene og β_j-ene i (10) bruker vi `lm`-kommandoen. Vi antar at vi alt har gitt kommandoen `options(contrasts=c("contr.sum","contr.poly"))` slik det er beskrevet i avsnitt 3. (Denne kommandoen trenger vi bare å gi én gang per `R`-sesjon.) Vi får da følgende resultat:

```
> fit2.lm=lm(Response~factor(Brand)+factor(Treat),data=exmp11.11)
> summary(fit2.lm)
               Estimate Std. Error t value Pr(>|t|)
(Intercept)     0.46583    0.03472  13.418 1.06e-05
factor(Brand)1  0.13167    0.04910   2.682  0.03645
factor(Brand)2 -0.12083    0.04910  -2.461  0.04905
factor(Treat)1  0.33750    0.06013   5.613  0.00137
factor(Treat)2 -0.12917    0.06013  -2.148  0.07531
factor(Treat)3 -0.04250    0.06013  -0.707  0.50622

Residual standard error: 0.1203 on 6 degrees of freedom
Multiple R-squared: 0.8751,     Adjusted R-squared: 0.771
F-statistic: 8.406 on 5 and 6 DF,  p-value: 0.01104
```

Merk at vi får estimater for μ, de to første α_i-ene og de tre første β_j-ene. Estimatet for den siste α_i-en får vi ved å bruke relasjonen (5) med $I = 3$. Tilsvarende får vi estimatet for den siste β_j-en.

Etter at vi har tilpasset en variansanalysemodell med `lm`-kommandoen, får vi variansanalysetabellen med `anova`-kommandoen (jf. side 588 i D&B):

```
> anova(fit2.lm)
              Df  Sum Sq  Mean Sq F value   Pr(>F)
factor(Brand)  2 0.12822 0.064108  4.4323 0.065765
factor(Treat)  3 0.47969 0.159897 11.0549 0.007399
Residuals      6 0.08678 0.014464
```

I `lm`-kommandoen skrev vi `factor(Brand)+factor(Treat)` for å gi `R` beskjed om at `Brand` og `Treat` er kategoriske variable. Du kan gi kommandoen `model.matrix(fit2.lm)` for å se at det `R` gjør er å lage $3-1=2$ forklaringsvariable for `Brand` og $4-1=3$ forklaringsvariable for `Treat` slik det er gitt i (11) og (12).

6. Toveis variansanalyse med interaksjon

Vi ser så på modell (2) for toveis variansanalyse med interaksjon. For å skrive denne modellen som en lineær regresjonsmodell av formen (3) innfører vi de $I + J - 2$ forklaringsvariablene gitt ved (11) og (12) i avnitt 4. I tillegg innfører vi $(I-1)(J-1)$ forklaringsvariable for interaksjonen. For $i = 1, \ldots, I-1$ og $j = 1, \ldots, J-1$ er disse gitt ved:

$$x_{l,(I-1)j+J-1+i} = \begin{cases} 1 & \text{hvis individ } l \text{ har nivå } i \text{ for faktor } A \text{ og nivå } j \text{ for faktor } B \\ -1 & \text{hvis individ } l \text{ har nivå } i \text{ for faktor } A \text{ og nivå } J \text{ for faktor } B \\ -1 & \text{hvis individ } l \text{ har nivå } I \text{ for faktor } A \text{ og nivå } j \text{ for faktor } B \\ 1 & \text{hvis individ } l \text{ har nivå } I \text{ for faktor } A \text{ og nivå } J \text{ for faktor } B \end{cases} \tag{13}$$

Hvis vi setter forklaringsvariablene (11), (12) og (13) inn i formelen (3) for lineær regresjon med $k = I + J - 2 + (I-1)(J-1)$, får vi at den lineære regresjonsmodellen er lik variansanalysemodellen (2). Sammenhengen mellom parameterene i (2) og (3) er som følger:

$$\begin{aligned} \mu &= \theta_0 \\ \alpha_i &= \theta_i \ ; && i = 1, \ldots, I-1 \\ \beta_j &= \theta_{I-1+j} \ ; && j = 1, \ldots, J-1 \\ \gamma_{ij} &= \theta_{(I-1)j+J-1+i} \ ; && i = 1, \ldots, I-1 \ ; \quad j = 1, \ldots, J-1 \end{aligned}$$

Vi har dermed vist at ved å innføre forklaringsvariablene (11), (12) og (13) kan modellen (2) for toveis variansanalyse skrives som en multippel lineær regresjonsmodell (3).

7. Bruk av R for toveis variansanalyse

For å vise hvordan vi kan bruke R for toveis variansanalyse, ser vi på eksempel 11.16 i D&B. Eksempelet gjelder en studie der en har undersøkt hvordan avlingen blir for $I = 3$ ulike tomatsorter som er plantet med $J = 4$ ulike tettheter og der det er $K = 3$ observasjoner per tomatsort og tetthet. Se side 599 i D&B for en mer utførlig beskrivelse av eksempelet.

Vi leser dataene inn i en dataramme som vi kaller `exmp11.16` slik det er beskrevet på kurssiden. Dataene er som følger:

```
> exmp11.16
   Yield Density Variety
1   10.5   10000       H
2    9.2   10000       H
3    7.9   10000       H
4    8.1   10000     Ife
5    8.6   10000     Ife
6   10.1   10000     Ife
7   16.1   10000       P
8   15.3   10000       P
9   17.5   10000       P
10  12.8   20000       H
11  11.2   20000       H
12  13.3   20000       H
13  12.7   20000     Ife
14  13.7   20000     Ife
15  11.5   20000     Ife
16  16.6   20000       P
17  19.2   20000       P
18  18.5   20000       P
19  12.1   30000       H
20  12.6   30000       H
21  14.0   30000       H
22  14.4   30000     Ife
23  15.4   30000     Ife
24  13.7   30000     Ife
25  20.8   30000       P
26  18.0   30000       P
27  21.0   30000       P
28  10.8   40000       H
29   9.1   40000       H
30  12.5   40000       H
31  11.3   40000     Ife
32  12.5   40000     Ife
33  14.5   40000     Ife
34  18.4   40000       P
35  18.9   40000       P
36  17.2   40000       P
```

For å få estimater for μ, α_i - ene, β_j - ene og γ_{ij} - ene i (2) bruker `lm` - kommandoen:

```
> fit3.lm=lm(Yield~factor(Variety)+factor(Density)
                  +factor(Variety):factor(Density),data=exmp11.16)
> summary(fit3.lm)
                                  Estimate Std. Error t value Pr(>|t|)
(Intercept)                        13.8889     0.2098  66.192  < 2e-16
factor(Variety)1                   -2.5556     0.2967  -8.612 8.35e-09
factor(Variety)2                   -1.6806     0.2967  -5.663 7.84e-06
factor(Density)1                   -2.4111     0.3634  -6.634 7.32e-07
factor(Density)2                    0.5000     0.3634   1.376    0.182
factor(Density)3                    1.8889     0.3634   5.197 2.52e-05
factor(Variety)1:factor(Density)1   0.2778     0.5140   0.540    0.594
factor(Variety)2:factor(Density)1  -0.8639     0.5140  -1.681    0.106
factor(Variety)1:factor(Density)2   0.6000     0.5140   1.167    0.255
factor(Variety)2:factor(Density)2  -0.0750     0.5140  -0.146    0.885
factor(Variety)1:factor(Density)3  -0.3222     0.5140  -0.627    0.537
factor(Variety)2:factor(Density)3   0.4028     0.5140   0.784    0.441

Residual standard error: 1.259 on 24 degrees of freedom
Multiple R-squared: 0.9174,     Adjusted R-squared: 0.8795
F-statistic: 24.22 on 11 and 24 DF,  p-value: 2.423e-10
```

Merk at vi ikke får estimater for α_i, β_j og γ_{ij} for $i = I = 3$ og/eller $j = J = 4$. Det kommer av restriksjonene $\sum_{i=1}^{I} \alpha_i = \sum_{j=1}^{J} \beta_j = \sum_{i=1}^{I} \gamma_{ij} = \sum_{j=1}^{J} \gamma_{ij} = 0$.

Vi får variansanalysetabellen med `anova` - kommandoen (jf. side 601 i D&B):

```
> anova(fit3.lm)
                                Df Sum Sq Mean Sq  F value    Pr(>F)
factor(Variety)                  2 327.60 163.799 103.3430 1.608e-12
factor(Density)                  3  86.69  28.896  18.2306 2.212e-06
factor(Variety):factor(Density)  6   8.03   1.339   0.8445    0.5484
Residuals                       24  38.04   1.585
```

Ovenfor skrev vi `factor(Variety)+factor(Density)+factor(Variety):factor(Density)` for å gi `R` beskjed om at `Variety` og `Density` er kategoriske variable og at vi vil estimere interaksjonen mellom dem. Du kan gi kommandoen `model.matrix(fit3.lm)` for å se at det `R` gjør er å lage $3 - 1 = 2$ forklaringsvariable for `Variety`, $4 - 1 = 3$ forklaringsvariable for `Density` og $(3 - 1)(4 - 1) = 6$ forklaringsvariable for interaksjonen slik det er gitt i (11), (12) og (13).

8. En alternativ parameterisering av variansanalysemodeller

I modellen for envis variansanalyse:

$$X_{ij} = \mu + \alpha_i + \epsilon_{ij}\,; \quad j = 1, 2, \ldots, J_i\,; \quad i = 1, 2, \ldots, I\,; \tag{1}$$

er der vanlig å bruke restriksjonen

$$\sum_{i=1}^{I} \alpha_i = 0 \tag{14}$$

slik vi har gjort i avsnittene 2 og 3 ovenfor og slik det er beskrevet i avsnitt 11.3 i D&B. Et alternativ til denne restriksjonen er å sette

$$\alpha_1 = 0 \tag{15}$$

Når vi bruker restriksjonen (15) vil μ i (1) være forventningsverdien for gruppe 1, mens α_i for $i = 2, \ldots, I$ vil være forventet forskjell mellom gruppene $2, \ldots, I$ og gruppe 1 (som da kalles referansegruppen). Om vi velger restriksjonen (14) eller restriksjonen (15) har ingen betydning for variansanalysetabellen. Men estimatene for μ og α_i-ene blir ikke de samme for de to restriksjonene.

Svarende til restriksjonen (15) har vi forklaringsvariablene

$$x_{l,i-1} = \begin{cases} 1 & \text{hvis individ } l \text{ hører til gruppe } i \\ 0 & \text{ellers} \end{cases} \tag{16}$$

for $i = 2, \ldots, I$. Merk at dette svarer til det som står om koding av kategoriske variable ved lineær regresjon på sidene 696-699 i D&B.

Hvis vi setter forklaringsvariablene (16) inn i (3) med $k = I - 1$, ser vi at den lineære regresjonsmodellen er lik variansanalysemodellen (1) når vi bruker restriksjonen (15).

For å vise hvordan vi kan bruke `R` for å få estimater for μ og α_i-ene når vi bruker restriksjonen (15), ser vi på eksempelet i avsnitt 3. Vi må først gi `R` beskjed om at vi vil bruke restriksjonen (15). Det gjør vi ved kommandoen:

```
> options(contrasts=c("contr.treatment","contr.poly"))
```

Merk at opsjonen `options(contrasts=c("contr.treatment","contr.poly"))` er "default", så vi behøver bare gi kommandoen hvis vi tidligere i samme `R`-sesjon har gitt kommandoen `options(contrasts=c("contr.sum","contr.poly"))`.

Vi kan så tilpasse modellen (1) på tilsvarende måte som i avsnitt 3:

```
> fit.lm=lm(sleep~factor(treat),data=exmp11.5)
> summary(fit.lm)
               Estimate Std. Error t value Pr(>|t|)
(Intercept)      79.280      4.312  18.386 3.49e-12
factor(treat)2  -17.740      6.098  -2.909   0.0102
factor(treat)3  -31.360      6.098  -5.143 9.83e-05
factor(treat)4  -46.520      6.098  -7.629 1.02e-06

Residual standard error: 9.642 on 16 degrees of freedom
Multiple R-squared: 0.7982,     Adjusted R-squared: 0.7603
F-statistic: 21.09 on 3 and 16 DF,  p-value: 8.325e-06
```

Merk at vi nå får estimater for μ og de tre siste α_i-ene, mens den første α_i-en er satt til null ved restriksjonen (15).

Vi ser så på modellen for toveis variansanalyse:

$$X_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \epsilon_{ijk}\,; \quad k = 1, \ldots, K_{ij}\,; \quad j = 1, \ldots, J\,; \quad i = 1, \ldots, I\,; \tag{2}$$

Her angir γ_{ij} - ene interaksjon mellom faktor A og faktor B. For en modell uten interaksjon utelater vi γ_{ij} - ene i (2) og i restriksjonene (17) og (18).

For toveis variansanalyse er der vanlig å bruke restriksjonene

$$\sum_{i=1}^{I} \alpha_i = \sum_{j=1}^{J} \beta_j = \sum_{i=1}^{I} \gamma_{ij} = \sum_{j=1}^{J} \gamma_{ij} = 0 \tag{17}$$

slik vi har gjort i avsnittene 4 - 7 ovenfor og slik det er beskrevet i avsnitt 11.5 i D&B. Et alternativ til disse restriksjonen er å sette

$$\alpha_1 = \beta_1 = \gamma_{1j} = \gamma_{i1} = 0 \tag{18}$$

Hvis vi bruker opsjonen `options(contrasts=c("contr.treatment","contr.poly"))` for toveis variansanalyse, får vi tilpasset modellen (2) med restriksjonene (18). Prøv dette selv for eksemplene i avsnittene 5 og 7 og sammenlign med det vi fikk der.

9. Toveis variansanalyse når K_{ij} - ene ikke er like

I eksempelet i avsnitt 7 hadde vi like mange observasjoner for hver kombinasjon av nivåene for faktorene A og B. Da er det en entydig oppdeling av den totale kvadratsummen SST i en kvadratsum SSA for faktor A, en kvadratsum SSB for faktor B, en kvadratsum $SSAB$ for interaksjon og en residualkvadratsum SSE; jf. side 599 i D&B. Hvis det ikke er like mange observasjoner for hver kombinasjon av nivåene for faktorene A og B, har vi ikke en slik entydig oppdeling.

For å illustrere dette ser vi på data fra en studie der en undersøkte effekten av et syntetisk veksthormon for barn av kort vekst. Spesielt var en interessert i å studere effekten av barnets kjønn (1 = gutt, 2 = jente) og graden av veksthemming (1 = alvorlig veksthemming, 2 = moderat veksthemming og 3 = mild veksthemming). Barna ble behandlet med veksthormon i ett år, og en målte økningen i veksthastigheten (cm per måned). I utgangspunktet var det tre gutter og tre jenter for hver grad av veksthemming. Men siden noen av barna ikke fullførte behandlingen, har en ikke data for så mange barn.

Vi leser dataene inn i en dataramme som vi kaller `vekst` slik det er beskrevet på kurssiden. Dataene er som følger:

```
> vekst
   kjonn hemming vekst
1      1       1   1.4
2      1       1   2.4
3      1       1   2.2
4      1       2   2.1
5      1       2   1.7
6      1       3   0.7
7      1       3   1.1
```

```
8      2       1   2.4
9      2       2   2.5
10     2       2   1.8
11     2       2   2.0
12     2       3   0.5
13     2       3   0.9
14     2       3   1.3
```

Vi gjør først en variansalalyse der vi gir faktoren `kjonn` først i modellformelen:

```
> fit4.lm=lm(vekst~factor(kjonn)+factor(hemming)
             +factor(kjonn):factor(hemming),data=vekst)
> anova(fit4.lm)
                             Df Sum Sq Mean Sq F value   Pr(>F)
factor(kjonn)                 1 0.0029 0.00286  0.0176 0.897785
factor(hemming)               2 4.3960 2.19800 13.5262 0.002713
factor(kjonn):factor(hemming) 2 0.0754 0.03771  0.2321 0.798034
Residuals                     8 1.3000 0.16250
```

Så gjør vi en variansalalyse der vi gir faktoren `hemming` først i modellformelen:

```
> fit4b.lm=lm(vekst~factor(hemming)+factor(kjonn)
              +factor(hemming):factor(kjonn),data=vekst)
> anova(fit4b.lm)
                             Df Sum Sq Mean Sq F value   Pr(>F)
factor(hemming)               2 4.3063 2.15314 13.2501 0.002891
factor(kjonn)                 1 0.0926 0.09257  0.5697 0.472022
factor(hemming):factor(kjonn) 2 0.0754 0.03771  0.2321 0.798034
Residuals                     8 1.3000 0.16250
```

Vi ser at variansanalysetabellene er forskjellige. Det illusterer at oppdelingen av den totale kvadratsummen ikke er entydig når det ikke er like mange observasjoner for hver kombinasjon av nivåene for de to faktorene. Toveis variansanalyse der K_{ij}-ene er forskjellige krever derfor noe mere omtanke enn når K_{ij}-ene er like.