Eksamen i Statistisk Dataanalyse 2, 14. april 2007

Vejledende besvarelse

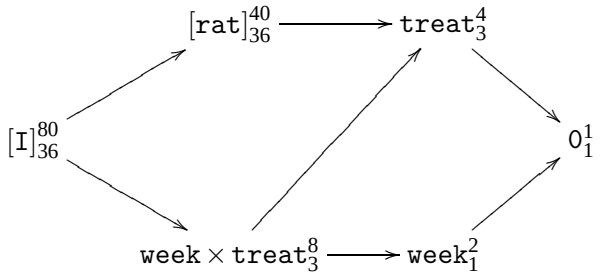
Opgave 1

1. De relevante faktorer er `rat` (rotte, 40 niveauer), `treat` (behandling, 4 niveauer) og `week` (uge, 2 niveauer). I den statistiske model bør `rat` indgå som tilfældig, mens `treat`, `week` og produktfaktoren $\texttt{week} \times \texttt{treat}$ bør indgå som systematiske faktorer:

   $$y_i = \mu + \gamma(\texttt{week}_i, \texttt{treat}_i) + A(\texttt{rat}_i) + e_i$$

   hvor $A(1), \ldots, A(40) \sim N(0, \sigma_A^2)$, $e_i \sim N(0, \sigma^2)$, alle uafhængige.

   Behandling er grovere end rotte, så faktordiagrammet bliver følgende:



2. Først testes hypotesen om at der ikke er nogen vekselvirkning:

   $$H_0 : \gamma(\texttt{week}_i, \texttt{treat}_i) = \alpha(\texttt{week}_i) + \beta(\texttt{treat}_i)$$

   svarende til den additive model

   $$y_i = \mu + \alpha(\texttt{week}_i) + \beta(\texttt{treat}_i) + A(\texttt{rat}_i) + e_i$$

   Fra R-udksiften ses at $F = 3.975$ der vurderet i $F(3,36)$-fordelingen giver $p = 0.24$. Vekselvirkningen er således ikke signifikant.

   Derefter testes hypoteserne om at der ikke er effekt af `treat` og `week`:

   $$H_0 : \alpha(7) = \alpha(13); \qquad H_0 : \beta(1) = \beta(2) = \beta(3) = \beta(4).$$

   Bemærk at effekten af `treat` skal testes mod `rat`, mens effekten af `week` skal testes mod `I`. Fra R-udskriften ses at begge effekter er signifikante ($F = 3.975$, $F(3,36)$, $p = 0.015$ for `treat` og $F = 841$, $F(1,39)$, $p < 0.0001$ for `week`), således at den additive model ikke kan reduceres yderligere.

   Forskellene mellem behandlingerne, dvs. forskel i forventet vægt, estimeres til:

   $$\hat{\alpha}(2) - \hat{\alpha}(1) = 28.1; \quad \hat{\alpha}(3) - \hat{\alpha}(1) = 26.5; \quad \hat{\alpha}(4) - \hat{\alpha}(1) = 3.2.$$

Fra de tilhørende SE'er (eller LSD-værdien der er cirka 2 gange *se*, dvs. cirka 21), ser vi at der ikke er signifikant forskel på behandling 1 og 4 og på behandling 2 og 3.

Effekten af uge er estimeret til

$$\hat{\beta}(13) - \hat{\beta}(7) = 66.7$$

således at rotterne forventes at veje 66.7 g mere i uge 13 end i uge 6.

3. Startmålingerne kan inddrages som kovariat i modellen:

$$y_i = \mu + \alpha(\texttt{week}_i) + \beta(\texttt{treat}_i) + \delta \cdot \texttt{start}_\texttt{i} + A(\texttt{rat}_i) + e_i$$

At denne model *ikke* giver en bedre beskrivelse af data svarer til hypotesen $\delta = 0$. Denne hypotese forkastes klart ($\hat{\delta} = 1.40$, $t = 3.08$, $p = 0.004$), så startværdierne bidrager signifikant til beskrivelsen af data.

Fra `anova(modelC)` ser vi at både `treat` og `week` stadig er signifikante, og fra `summary(modelC)` endvidere at der ikke er signifikant forskel på behandlingerne 1 og 4 og på behandlingerne 2 og 3. Estimaterne er sammenlignelige med dem fra modellen uden startværdier. Konklusionen er således uændret.

4. 6% bioprotein svarer til behandling 2. Vi er således interesseret i et estimat for

$$f = \mu + \alpha(13) + \beta(2) + \delta \cdot 100$$

Fra `summary(modelC)` får vi estimatet:

$$\begin{aligned}\hat{f} &= \hat{\mu} + \hat{\alpha}(13) + \hat{\beta}(2) + \hat{\delta} \cdot 100 \\ &= 163.528 + 66.675 + 29.033 + 100 \cdot 1.404 = 399.6\end{aligned}$$

svarende til linearkombinationen `(1,100,1,1,0,0)` af estimaterne. Fra `estimable(forv3)` får vi desuden konfidensintervallet $(384.9, 414.3)$.

Opgave 2

1. Opstil f.eks. følgende tabeller:

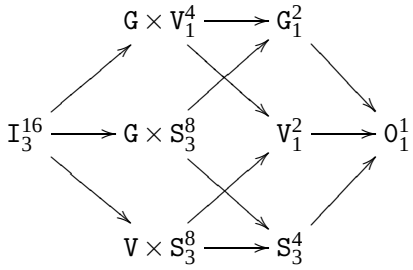
G \ V	1	2
1	4	4
2	4	4

S \ V	1	2
1	2	2
2	2	2
3	2	2
4	2	2

S \ G	1	2
1	4	0
2	0	4
3	4	0
4	0	4

Heraf fremgår, at `G` $\times$ `V` og `S` $\times$ `V` er balancerede, mens `S` $\times$ `G` *ikke* er balanceret. Vi ser også, at `G` er grovere end `S`.

2. Faktordiagram:



Hver kombination af G × V × S optræder netop en gang, hvorfor vi ikke kan teste for effekt af pågældende trefaktorvekselvirkning.

3. Forsøget er et 2^n-te forsøg med 3 faktorer udført på to lige store blokke, hvor trefaktorvekselvirkningen G × V × S er konfunderet med Blok. Dette ses f.eks. ved at benytte lige/ulige reglen som i nedenstående tabel og konstatere, at det netop er forsøgsplanen fra opgaveformuleringen som fremkommer.

G	V	S	G+V+S	Blok 1	Blok 2
1	1	1	3	x	
1	1	2	4		x
1	2	1	4		x
1	2	2	5	x	
2	1	1	4		x
2	1	2	5	x	
2	2	1	5	x	
2	2	2	6		x

4. Forsøgsplan 2 er her fremkommet ved at benytte planen fra spm 3. og dublere nogle af behandlingerne inden for de to blokke. Dette ændrer ikke på det faktum, at G × V × S er konfunderet med Blok. Dette ses af R-udskriften ved, at `anova(model2blok)` ikke giver et test for G × V × S, når Blok inddrages i modellen som systematisk faktor. Derimod viser R-udskriften, at `anova(model1blok)` giver et test for G × V × S (og alle øvrige hoved- og vekselvirkninger), selvom Blok inddrages som systematisk faktor i modellen. Da formålet med forsøget er at undersøge alle hoved- og vekselvirkninger mellem G, V og S, bør man således foretrække forsøgsplan 1.

Opgave 3

1. Modellerne A og B svarer til følgende modeller, hvor y angiver løbstiden, dag dagen og både grp og m antal mænd på holdet (som faktor hhv. numerisk variabel):

$$\texttt{modelA}: \quad y_i = \mu + \alpha(\texttt{dag}_i) + \beta(\texttt{grp}_i) + e_i$$
$$\texttt{modelB}: \quad y_i = \mu + \alpha(\texttt{dag}_i) + \delta \cdot m_i + e_i$$

Forskellen mellem modellerne er således om antallet af mænd indgår som faktor eller som kovariat. modelA er den sædvanlige additive model med dimension dim(modelA)=4+6-1=9, mens dimensionen af modelB er dim(modelB)=4+1=5 (en parameter per dag samt en hældning).

2. Figuren til venstre kan bruges til at vurdere om der er vekselvirkning mellem `dag` og `grp`. De fire kurver ser rimeligt parallelle ud så der synes ikke at være grund til bekymring om vekselvirkning. Residualplottet til højre kan bruges til at vurdere om der er varianshomogenitet, og giver ikke umiddelbart anledning til bekymring.

   I `modelB` antages at der er en lineær sammenhæng mellem antal mænd på holdet og løbstiden. Test for reduktion fra `modelA` til `modelB` testes på $F = 2.65$ der vurderet i $F(4,15)$-fordelingen giver en p-værdi på 0.075. Faktormodellen giver således ikke en signifikant bedre beskrivelse af data (omend de 7.5% er tæt på!), så vi bruger `modelB` i det følgende.

3. Parameteren δ er netop ændringen i løbstiden når antallet af mænd øges med en. Et estimat og et konfidensinterval for forskellen mellem kvinders og mænds femkilometertid, målt i sekunder, er derfor $-\hat{\delta} = 242\ (221,262)$.

4. Forskellene mellem dagene estimeres til

$$\hat{\alpha}(\texttt{ti}) - \hat{\alpha}(\texttt{ma}) = 15.8, \quad \hat{\alpha}(\texttt{on}) - \hat{\alpha}(\texttt{ma}) = -205.2, \quad \hat{\alpha}(\texttt{to}) - \hat{\alpha}(\texttt{ma}) = -248.5$$

   LSD-værdien er cirka 98 (beregnet som halvdelen af længden af konfidensintervallerne eller som $2.093 \cdot 46.824$). Vi ser derfor at der ikke er signifikant forskel på løbstiderne mandag og tirsdag, og ikke signifikant forskel mellem onsdag og torsdag, men til gengæld voldsomt signifikant forskel på tiderne mandag-tirsdag og onsdag-torsdag. Ruten var altså hurtigere de to sidste dage.

   NB. Hvis man skulle udføre et egentligt test for ovenstående i stedet for at basere det på parvise sammenligninger kunne man gøre følgende: først undersøge om faktoren `dag` kunne erstattet med en faktor der slog mandag og tirsdag hhv. onsdag og torsdag sammen; og dernæst teste om der var en effekt af denne faktor. Dette kan ikke udføres vha. udskrifterne i opgaven og forventes ikke angivet i besvarelsen.