Eksamen i Statistisk Dataanalyse 2, 4. april 2009

Vejledende besvarelse

Opgave 1

1. Modellen fittet som `modelA` er den additive model for tosidet variansanalyse med hovedeffekt af faktorerne `hest` og `halt`

$$Y_i = \alpha(\texttt{halt}_i) + \beta(\texttt{hest}_i) + e_i, \quad e_1, \ldots, e_{40} \sim N(0, \sigma^2).$$

2. Vi ønsker at teste hypotesen, $H_0 : \beta(1) = \ldots = \beta(8) = 0$, svarende til den ensidede variansanalysemodel med hovedeffekt af `halt`

$$Y_i = \alpha(\texttt{halt}_i) + e_i, \quad e_1, \ldots, e_{40} \sim N(0, \sigma^2).$$

   Denne model er anført som `modelB` i R-udskriften. Testet godkendes ($F = 0.932, p = 0.498$).

3. Dernæst undersøges om halthedfaktoren (`halt`) har betydning for halthedsgraden svarende til hypotesen,

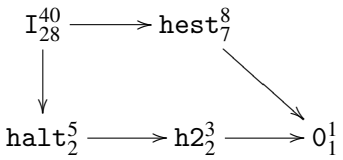
$$H_0 : \alpha(\texttt{normal}) = \alpha(\texttt{vforben}) = \alpha(\texttt{hforben}) = \alpha(\texttt{vbagben}) = \alpha(\texttt{hbagben}).$$

   Dette er modellen med fælles intercept

$$Y_i = \mu + e_i, \quad e_1, \ldots, e_{40} \sim N(0, \sigma^2),$$

   som er fittet som `modelE` i R-udskriften. Testet forkastes ($F = 18.76, p < 0.0001$).

4. Det væsentlige er at indse, at faktoren `h2` er grovere end halthedsfaktoren `halt`. Faktordiagrammet ser ud som følger



5. Den statistiske model svarende til `modelD` er

$$Y_i = \gamma(\texttt{h2}_i) + e_i, \quad e_1, \ldots, e_{40} \sim N(0, \sigma^2)$$

   og modellen skal testes imod `modelB`. Testet godkendes ($F = 0.335, p = 0.718$). Udtrykt ved parametriseringen i `modelB` svarer `modelD` til hypotesen,

$$H_0 : \alpha(\texttt{vforben}) = \alpha(\texttt{hforben}) \text{ og } \alpha(\texttt{vbagben}) = \alpha(\texttt{hbagben}).$$

I ord testes om halthedsgraden (y) kan bruges til at skelne mellem halthed i højre og venstre side. Konklusionen er, at der målt med halthedsscoren y ikke er forskel på de to sider.

Man kan eventuelt gå videre og teste hypotesen

$$H_0 : \gamma(\texttt{for}) = \gamma(\texttt{bag}) = \gamma(\texttt{normal})$$

mod modelD. Af R-udskriften ses at denne hypotese forkastes klart ($F = 38.6, p < 0.0001$). Slutmodellen er således modelD, som beskrevet ovenfor. Parameterestimaterne for den systematiske effekt af h2 bliver

$$\hat{\gamma}(\texttt{for}) = -2.7380 \quad \hat{\gamma}(\texttt{bag}) = -3.6407 \quad \hat{\gamma}(\texttt{norm}) = -5.7223$$

mens residualspredningen estimeres til $\hat{\sigma} = 0.7854$. Man kan eventuelt angive konfidensintervaller vha. formlen

$$\hat{\gamma}(j) \pm \texttt{qt}(0.975, 37) \cdot SE.$$

6. Slutmodellen fra spørgsmål **5.** er

$$Y_i = \gamma(\texttt{h2}_i) + e_i, \quad e_1, \ldots, e_{40} \sim N(0, \sigma^2)$$

svarende til modelD i R-udskriften. Testet for om der er forskel på forbens- og bagbenshalthed svarer til hypotesen

$$H_0 : \gamma(\texttt{for}) = \gamma(\texttt{bag}).$$

Ved at anvende estimable på modelD1 fra R-udskriften (- svarer til modelD uden intercept), kan vi få et estimat og et konfidensinterval for forskellen

$$\gamma(\texttt{bag}) - \gamma(\texttt{for}).$$

Af R-udskriften ved kommandoen

```
> estimable(modelD1,estD,cont.int=0.95)
```

kan man i linjen svarende til est4 aflæse, at forskellen estimeres til

$$\hat{\gamma}(\texttt{bag}) - \hat{\gamma}(\texttt{for}) = -0.9027 \quad [-1.465, -0.340].$$

Da konfidensintervallet for forskellen ikke indeholder 0 forkastes hypotesen, og vi konkluderer, at der er forskel på forben- og bagbenshalthed. Af R-udskriften fremgår desuden direkte, at p-værdien for testet er $p = 0.002$.

Opgave 2

1. Den statistiske model svarende til model1 er modellen med systematisk effekt af vekselvirkningen mellem faktorerne dyrkningsbetingelse (treat) og temperatur (temp) samt en tilfældig effekt af tube

$$Y_i = \gamma(\texttt{treat} \times \texttt{temp}_i) + A(\texttt{tube}_i) + e_i,$$

hvor $A(1), \ldots, A(18)$ er uafhængige $\sim N(0, \sigma_A^2)$, og $e_1, \ldots, e_{126}$ er uafhængige $\sim B(0, \sigma^2)$.

2. I resten af opgaven indgår temperaturen som en kovariat gennem den reciprokke Kelvin temperatur `tempreci`. Den statistiske model som beskriver en retlinet sammenhæng mellem `tempreci` og ATPase aktivitet svarer til `model3` i R-udskriften og kan skrives som

$$Y_i = \alpha(\texttt{treat}_i) + \beta(\texttt{treat}_i) \cdot \texttt{tempreci}_i + A(\texttt{tube}_i) + e_i,$$

hvor $A(1),\ldots,A(18)$ er uafhængige $\sim N(0,\sigma_A^2)$, og $e_1,\ldots,e_{126}$ er uafhængige $\sim B(0,\sigma^2)$.

3. Test for Arrhenius ligning svarer til at sammenligne modellerne `model3` og `model1` fra de to foregående spørgsmål. Testet godkendes ($LR = 31.27, p = 0.402$), så datasættet bekræfter tilsyneladende Arrhenius ligning.

4. Med udgangspunkt i den lineære model (`model3`) fra spørgsmål **2.** bør man teste hypotesen om, at hældningen er uafhængig af `treat`. Dette svarer til modellen

$$Y_i = \alpha(\texttt{treat}_i) + \beta \cdot \texttt{tempreci}_i + A(\texttt{tube}_i) + e_i,$$

hvor $A(1),\ldots,A(18)$ er uafhængige $\sim N(0,\sigma_A^2)$, og $e_1,\ldots,e_{126}$ er uafhængige $\sim B(0,\sigma^2)$, svarende til `model4` i R-udskriften. Testet forkastes ($LR = 110.6, p < 0.0001$).

Man kunne også teste om skæringen er uafhængig af `treat`, men denne hypotese er ret uinteresserant i praksis. Denne svarer til sammenligningen af modellerne `model5` og `model3` i R-udskriften ($F = 400, p < 0.0001$).

Slutmodellen bliver således `model3` svarende til modellen i spørgsmål **2.** ovenfor. Parameterestimaterne fås af R-udskriften svarende til `model3ny`, hvor R er blevet tvunget til at benytte en parametrisering, hvoraf de 6 hældninger og de 6 skæringer direkte kan aflæses. Estimaterne for varianskomponenterne bliver

$$\hat{\sigma}_A = 0.395 \quad \hat{\sigma} = 0.118.$$

Man kan heraf f.eks. beregne, at 91.8% af variansen stammer fra variationen mellem reagensglas (`tube`).

Svarende til `treat=CaCl2:1` og `tempreci`=0.00341 fås en forventet ATPase aktivitet på

$$\hat{\alpha}(\texttt{CaCl2:1}) + \hat{\beta}(\texttt{CaCl2:1}) \cdot 0.00341 = 21.897 - 5546.09 \cdot 0.00341 = 2.98.$$

5. I stedet for at teste Arrhenius ligning ved at sammenligne modellerne `model1` og `model3`, hvor temperaturen benyttes som hhv. en faktor og en kovariat, kunne man teste `model3` mod en model, hvor man tilføjer et kvadratisk led i kovariaten `tempreci`. Dette svarer til middelværdistrukturen i spørgsmål **5.** Testet for Arrhenius ligning svarer til at undersøge, om koefficienten til det kvadratiske led er nul, dvs. $H_0 : \gamma = 0$.

I vedhæftede R-udskrift svarer `modelalt` til modellen

$$Y_i = \alpha(\texttt{treat}_i) + \beta(\texttt{treat}_i) \cdot \texttt{tempreci}_i + \gamma \cdot \texttt{tempreci}_i^2 + A(\texttt{tube}_i) + e_i,$$

hvor $A(1),\ldots,A(18)$ er uafhængige $\sim N(0,\sigma_A^2)$, og $e_1,\ldots,e_{126}$ er uafhængige $\sim B(0,\sigma^2)$. Bemærk, at middelværdistrukturen i `modelalt` er identisk med den, som er anført under spørgsmål **5.** i opgaveformuleringen.

Af kommandoen

`summary(modelalt)`

findes estimatet for koefficienten til de kvadratiske led

$$\hat{\gamma} = -3.65 \cdot 10^6,$$

og det anføres, at γ er signifikant forskellig fra nul ($p < 0.0001$.)

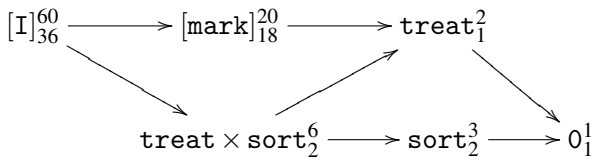
Opgave 3

1. Da produktfaktoren `treat` $\times$ `sort` optræder på 6 niveauer, vil det være oplagt at lave et fuldstændigt randomiseret blokforsøg, hvor hver af de 6 kombinationer allokeres ved randomisering til netop en parcel (forsøgsenhed) inden for hver af de 20 marker. Dette er praktisk muligt, da hver mark ifølge opgaveformuleringen kan opdeles i op til 8 mindre parceller. Det totale antal forsøgenheder ved et fuldstændigt randomiseret blokforsøg bliver $6 \cdot 20 = 120$.

2. Den skitserede forsøgsplan er et splitplot forsøg med

   - marker (`mark`) som helplot
   - behandling (`treat`) som helplotfaktor
   - parceller som delplot
   - `sort` som delplotfaktor

   Faktordiagrammet ser ud som følger



   Randomiseringen foretages i to trin

   - Først udvælges ved randomisering de 10 marker, som skal behandles (`treat=T`). De øvrige 10 marker behandles ikke (`treat=U`).
   - Inden for hver af de 20 marker randomiseres de 3 sorter (`sort=1,2,3`) ud på de 3 parceller på pågældende mark.

3. Produktfaktoren `treat` $\times$ `sort` optræder på 6 niveauer, så da blokkene (`mark`) kun indeholder 3 parceller hver, er det ikke muligt at udføre et fuldstændigt blokforsøg. Den anførte forsøgsplan er imidlertid et balanceret ufuldstændigt blokforsøg (BIBD), hvor hvert par af behandlinger givet ved `treat` $\times$ `sort` optræder præcis 4 gange inden for samme mark. For at overbevise sig om dette, kan man f.eks. lave en incidens-matrix

	T1	T2	T3	U1	U2	U3
T1	10	4	4	4	4	4
T2	4	10	4	4	4	4
T3	4	4	10	4	4	4
U1	4	4	4	10	4	4
U2	4	4	4	4	10	4
U3	4	4	4	4	4	10

Udfærdigelsen af forsøgsplanen foregår som følger

- Først laves 20 grupper hver bestående af 3 af de 6 kombinationer givet ved `treat` $\times$ `sort`, således at betingelserne for et fuldstændigt balanceret blokforsøg er opfyldt.
- Dernæst allokeres de 20 grupper ud på de 20 blokke (`mark`) ved randomisering.
- Endelig randomiseres de 3 behandlinger (-givet ved `treat` $\times$ `sort`) ud på de 3 parceller. Denne randomisering foretages for hver af de 20 blokke.