Eksamen i Statistisk Dataanalyse 2, 10. april 2008

Vejledende besvarelse

Opgave 1

1. Modellen fittet som `model1` er den lineære model med vekselvirkning af jordtype og fredningsstatus samt højde som kovariat:

$$\texttt{model1}: y_i = \alpha(\texttt{fredet}_i) + \beta(\texttt{jord}_i) + \gamma(\texttt{jord}_i, \texttt{fredet}_i) + \delta \cdot \texttt{hojde}_i + e_i$$

   hvor $e_1, \ldots, e_{50}$ er uafhængige $N(0, \sigma^2)$-fordelte.

2. Først undersøges det om der er vekselvirkning mellem jordtype og fredningsstatus. Modellen uden vekselvirkning er `model2`:

$$\texttt{model2}: y_i = \alpha(\texttt{fredet}_i) + \beta(\texttt{jord}_i) + \delta \cdot \texttt{hojde}_i + e_i$$

   Modellen kan ikke afvises ($F = 0.74$, $p = 0.48$), dvs. der er ikke signifikant vekselvirkning.

   Dernæst undersøges om der er en effekt af fredningstypen. Hypotesen er, at $\alpha(\texttt{fredet}) = \alpha(\texttt{normal})$ svarende til

$$\texttt{model3}: y_i = \beta(\texttt{jord}_i) + \delta \cdot \texttt{hojde}_i + e_i.$$

   Hypotesen kan ikke afvises ($F = 2.80$, $p = 0.10$), så der er ikke signifikant forskel på fredede og ikke-fredede områder.

   Derefter undersøges om der er forskel mellem jordtyperne. Hypotesen er, at $\beta(\texttt{ler}) = \beta(\texttt{muld}) = \beta(\texttt{kalk})$ svarende til

$$\texttt{model4}: y_i = \beta + \delta \cdot \texttt{hojde} + e_i.$$

   Hypotesen kan afvises klart ($F = 22.6$, $p < 0.0001$). Der er således en klar effekt af jordtype.

   Endelig undersøges om biomassen ændres med højden. Hypotesen er at $\delta = 0$ svarende til

$$\texttt{model5}: y_i = \beta(\texttt{jord}_i) + e_i.$$

   Modellen testes mod `model3` og afvises klart ($F = 519.8$, $p < 0.0001$). Der er således en klar effekt af højden over havet.

3. Der er således ikke signifikant effekt af fredningsstatus på biomassen. Derimod er der klar signifikant effekt af både jordtype og højde over havet på biomassen. Den endelige model er `model3`.

Den forventede biomasse er højest for kalk, 0.086 lavere for ler og 0.236 lavere for muld. Begge forskelle er signifikante ($p = 0.02$ for ler og $p < 0.0001$ for muld). Også forskellen mellem muld og ler på 0.159 er signifikant, selvom det ikke direkte kan aflæses fra udskriften (standard error vil være cirka som for de andre forskelle).

Desuden falder den forventede biomasse med 0.0030 per meter over havet.

4. For et fredet jordstykke med muldjord beliggende 150 m over havet er den forventede biomasse

$$\hat{y} = \hat{\beta}(\texttt{muld}) + \hat{\delta} \cdot 150 = (2.3337 - 0.2359) - 0.002954 \cdot 150 = 1.655$$

Dette ses også af `estimable`-outputtet ud for `x3`. Herfra aflæses også konfidensintervallet, $[1.606, 1.704]$.

5. At effekten af højden over havet er forskellig for de tre jordtyper svarer til forskellige hældninger, dvs. en vekselvirkning mellem `jord` og `hojde`:

$$\texttt{model3} : y_i = \beta(\texttt{jord}_i) + \delta(\texttt{jord}_i) \cdot \texttt{hojde}_i + e_i.$$

Modellen fittes fx. med

```
lm(biomasse ~ hojde + jord + hojde:jord)
```

Opgave 2

1. Modellen fittet som `modelA` er modellen med en tilfældig effekt af `dag` og en systematisk vekselvirkning mellem faktoren `morgen` og kovariaten `puls`:

$$\texttt{modelA} : Y_i = \alpha(\texttt{morgen}_i) + \beta(\texttt{morgen}_i) \cdot \texttt{puls}_i + A(\texttt{dag}_i) + e_i,$$

hvor $A(1), \ldots, A(14)$ er uafhængige $N(0, \sigma_D^2)$-fordelte og $e_1, \ldots, e_{70}$ er uafhængige $N(0, \sigma^2)$-fordelte. Residualplottet over de standardiserede residualer viser, at der stort set er varianshomogenitet, idet variationen i "punktskyen" ikke afhænger systematisk af de forventede værdier. Punkterne på QQ-plottet afviger ikke systematisk fra en ret linje, hvorfor antagelsen om normalfordelte fejl med rimelighed kan antages at være opfyldt for datasættet.

2. Hypotesen om, at effekten af den tilfældige faktor `dag` kan ignoreres formuleres som $\sigma_D^2 = 0$. Et F-test for hypotesen kan fås ved at sammenligne `model0` og `model1` givet ved

$$\begin{aligned} \texttt{model0} : Y_i &= \alpha(\texttt{morgen}_i) + \beta(\texttt{morgen}_i) \cdot \texttt{puls}_i + \delta(\texttt{dag}_i) + e_i \\ \texttt{model1} : Y_i &= \alpha(\texttt{morgen}_i) + \beta(\texttt{morgen}_i) \cdot \texttt{puls}_i + e_i, \end{aligned}$$

som svarer til, at `dag` inddrages som en systematisk faktor. Af R-udskriften ses, at der er en signifikant effekt af `dag` ($F = 12.47, p < 0.0001$).

3. Først undersøges om der er vekselvirkning mellem `morgen` og `puls`, dvs. om hældningerne svarende til $\texttt{morgen} = \texttt{ja}$ og $\texttt{morgen} = \texttt{nej}$ er ens. Dette svarer til modellen:

$$\texttt{modelB} : Y_i = \alpha(\texttt{morgen}_i) + \gamma \cdot \texttt{puls}_i + A(\texttt{dag}_i) + e_i,$$

hvor $A(1),\ldots,A(14)$ er uafhængige $N(0,\sigma_D^2)$-fordelte og $e_1,\ldots,e_{70}$ er uafhængige $N(0,\sigma^2)$-fordelte. Hypotesen godkendes ($LR = 0.16, p = 0.69$), så der er ingen vekselvirkning mellem `morgen` og `puls`.

Dernæst testes om der overhovedet er en effekt af `morgen` dvs. om $\alpha(\texttt{ja}) = \alpha(\texttt{nej})$. Dette svarer til modellen:

$$\texttt{modelD}: Y_i = \mu + \gamma \cdot \texttt{puls}_i + A(\texttt{dag}_i) + e_i,$$

hvor $A(1),\ldots,A(14)$ er uafhængige $N(0,\sigma_D^2)$-fordelte og $e_1,\ldots,e_{70}$ er uafhængige $N(0,\sigma^2)$-fordelte. Hypotesen godkendes ($LR = 0.60, p = 0.44$), så der er ingen effekt af faktoren `morgen` på omgangstiderne.

Det viser sig, at modellen ikke kan reduceres yderligere til den rene interceptmodel:

$$\texttt{modelE}: Y_i = \mu + A(\texttt{dag}_i) + e_i,$$

hvor $A(1),\ldots,A(14)$ er uafhængige $N(0,\sigma_D^2)$-fordelte og $e_1,\ldots,e_{70}$ er uafhængige $N(0,\sigma^2)$-fordelte ($LR = 0.0079, p < 0.0001$).

Slutmodellen er derfor `modelD`, og parameterestimaterne bliver

$$\hat{\mu} = 901.80[849.18, 954.42] \qquad \hat{\gamma} = -3.05[-3.37, -2.74] \qquad \hat{\sigma}_d = 14.17 \qquad \hat{\sigma} = 9.38.$$

4. Baseret på slutmodellen, `modelD`, ses at den forventede omgangstid om morgenen ændres med 15γ, når pulsen øges med 15 slag per minut. Et estimat og et konfidensinterval for denne forbedring (målt i sekunder) kan fås ud fra parameterestimaterne i delspørgsmål 3, og man får

$$45.82 \quad [41.08, 50.56].$$

Opgave 3

1. Forsøget er et 2^3-forsøg udført på to blokke (personer). Det ses, f.eks. ved at benytte lige/ulige reglen som i nedenstående tabel, at vekselvirkningen $\texttt{A} \times \texttt{C}$ er konfunderet med person.

A	B	C	A+C	person 1	person 2
1	1	1	2	x	
1	1	2	3		x
1	2	1	2	x	
1	2	2	3		x
2	1	1	3		x
2	1	2	4	x	
2	2	1	3		x
2	2	2	4	x	

2. Spørgmålet går på om der findes et balanceret ufuldstændigt blokforsøg (BIBD) med $v_T = 9$ behandlinger, $v_B = 15$ blokke (personer) og blokstørrelse $r_B = 3$ (tre behandlinger per person). Ifølge kompendiets Theorem 9.6 kræver dette, at hver behandling skal afprøves

$$r_T = \frac{r_B \cdot v_B}{v_T} = \frac{3 \cdot 15}{9} = 5$$

gange i forsøgsplanen. Endvidere skal hvert par af behandlinger forekomme

$$\lambda = \frac{r_T \cdot (r_B - 1)}{\nu_T - 1} = \frac{5 \cdot (3-1)}{8} = \frac{10}{8} = 1.25.$$

Da λ ikke er et helt tal, kan vi ihvert tilfælde konkludere, at det ikke er muligt at konstruere det ønskede forsøgsdesign.

3. Det vigtige er at indse, at `dag` er finere end måltid `M`, samt at `person` er finere end kosttilskud `K`. I modellen skal $\mathtt{K} \times \mathtt{M}$, `K` og `M` indgå som systematiske faktorer. Den relevante statistiske model er:

$$Y_i = \gamma(\mathtt{K} \times \mathtt{M}_i) + A(\mathtt{dag}_i) + B(\mathtt{person}_i) + e_i,$$

hvor

- $A(1), \ldots, A(9)$ er uafhængige $N(0, \sigma_D^2)$-fordelte
- $B(1), \ldots, B(9)$ er uafhængige $N(0, \sigma_P^2)$-fordelte
- $e_1, \ldots, e_{81}$ er uafhængige $N(0, \sigma^2)$-fordelte.

Faktordiagrammet ser ud som følger: