Kirsebær: et eksempel på anvendelse af generaliserede lineære mixed modeller.

Rasmus Waagepetersen*

1 Introduktion

Kategoriske data eller tælletalsdata optræder hyppigt i jordbrugsforskningen. Generaliserede lineære modeller (Nelder and Wedderburn, 1972, McCullagh and Nelder, 1989) er en populær klasse af statistiske modeller for disse typer data. Modelklassen er dog mindre anvendelig, når der er undergrupper af data, hvor observationer tilhørende samme undergruppe er korrelerede.

Som et konkret eksempel betragtes i afsnit 2 et datasæt bestående af registreringer af tilstanden (frisk eller død) for kirsebærblomster. Dette datamateriale blev indsamlet for at undersøge betydningen af valget af grundstamme for kirsebærblomsternes overlevelsesevne i forbindelse med nattefrost. Observationerne kan her grupperes efter et hierarki af grupperingsvariable. Det nederste niveau i hierarkiet er knoppen, som en given blomst tilhører, det næste niveau er grenen hvorpå knoppen sidder, og endeligt er det sidste niveau givet ved træet, som grenen tilhører. Korrelationen mellem observationer tilhørende samme knop, gren eller træ bevirker, at betingelserne for at anvende en sædvanlig generaliseret lineær model til at analysere data ikke er opfyldt.

I afsnit 3 gives en kort beskrivelse af de såkaldte generaliserede lineære mixed modeller (Breslow and Clayton, 1993, Littell *et al.*, 1996), som muliggør en eksplicit modellering af korrelationen mellem observationer, der tilhører samme niveau for en grupperingsvariabel. Det vises endvidere, hvorledes kirsebærblomst-datasættet kan analyseres vha. en generaliseret lineær mixed model, og en model-kontrol udføres for at undersøge om data på tilfredsstillende vis kan beskrives vha. en model af denne type. Datasættet er venligst stillet til rådighed af Birka Falk Kühn, Afdeling for Vegetabilske Fødevarer, Forskningscenter Årslev.

*Biometry Research Unit, Department of Agricultural Systems, Danish Institute of Agricultural Sciences, P.O. Box 50, DK-8830 Tjele. email: `Rasmus.Waagepetersen@agrsci.dk`.

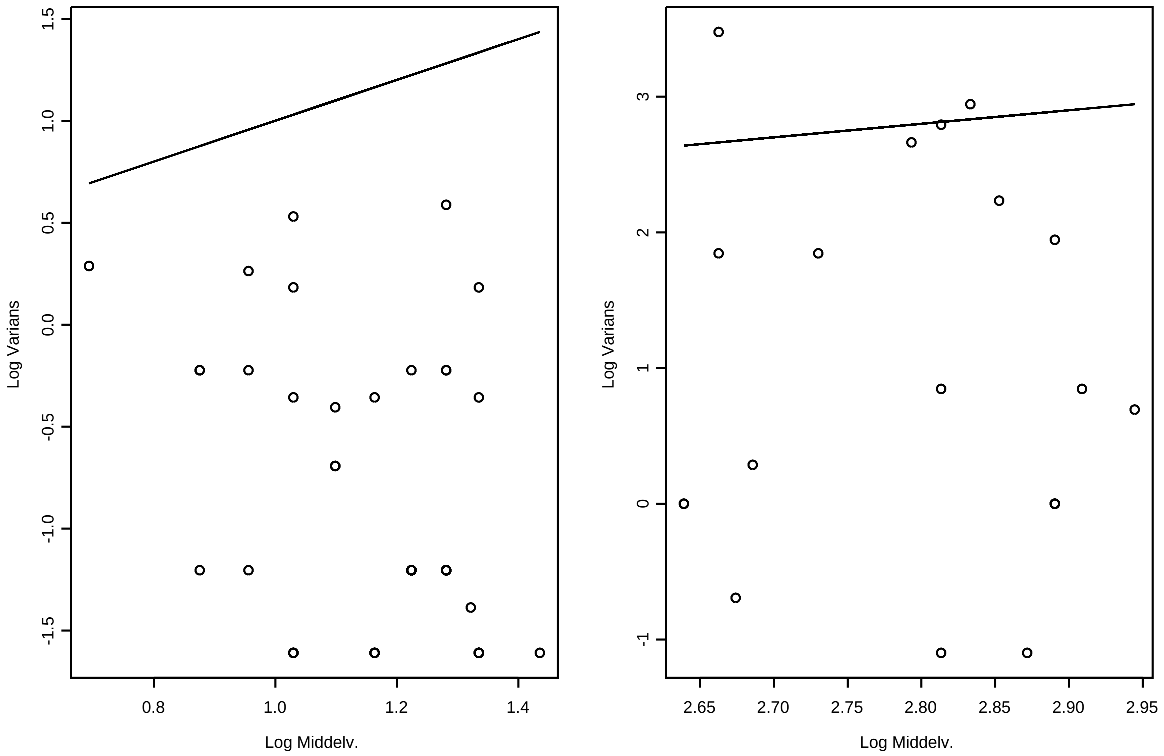
2 En indledende data-analyse

Lad den stokastiske variabel X_{ijklm} betegne tilstanden (0 for frisk, 1 for død) for den m'te blomst i den l'te knop på den k'te gren i det j'te træ tilhørende den i'te grundstamme, $m = 1, \ldots, n_{ijkl}$, $l = 1, \ldots, n_{ijk}$, $k = 1, \ldots, n_{ij}$, $j = 1, \ldots, 4$, og $i = 1, \ldots, 5$. Et udsnit af data er gengivet i Appendix D. Bemærk, at faktoren TRÆ, der nummererer træerne indenfor hver grundstamme, ikke optræder eksplicit i datamaterialet. Faktoren PARCEL er ækvivalent med krydsfaktoren GRST*TRÆ, således at de 20 niveauer af PARCEL svarer til de 20 træer hvorpå blomsterne er observeret.

De akkumulerede størrelser $N_{ijkl0} = \sum_{m=1}^{n_{ijkl}}(1 - X_{ijklm})$ og $N_{ijkl1} = \sum_{m=1}^{n_{ijkl}} X_{ijklm}$ betegner antal friske henholdsvis døde blomster i den $ijkl$'te knop. Tilsvarende betegner $N_{ijk0} = \sum_{l=1}^{n_{ijk}} N_{ijkl0}$ og $N_{ijk1} = \sum_{l=1}^{n_{ijk}} N_{ijkl1}$ antal henholdsvis friske og døde blomster på den ijk'te gren. Det totale antal blomster for en knop eller en gren betegnes $N_{ijkl\cdot}$ og $N_{ijk\cdot}$ henholdsvist. Vi vil først undersøge om disse totale antal kan beskrives ved Poisson fordelinger og beregner for hver kombination af i, j og k den empiriske middelværdi og varians for variablene $N_{ijkl\cdot}$, $l = 1, \ldots, n_{ijk}$. Logaritmen til de empiriske varianser er plottet mod logaritmen til de empiriske middelværdier i Figur 1 (højre plot). Hvis $N_{ijkl\cdot}$'erne for hver kombination af ijk var Poisson fordelte med fælles middelværdi λ_{ijk}, burde de plottede punkter ligge omkring identitetslinien. Dette er ikke tilfældet. En tilsvarende procedure kan anvendes på $N_{ijk\cdot}$'erne for at checke om $N_{ijk\cdot}$'erne skulle være Poisson-fordelte med middelværdi givet ved λ_{ij} for hver kombination ij, se Figur 1 (venstre plot). Heller ikke for disse tal er Poisson antagelsen holdbar, idet varianserne også her generelt er for små i forhold til middelværdierne. Faktisk gælder det, at knop-tallene $N_{ijkl\cdot}$ stort set alle er enten 3 eller 4, og der er derfor ikke meget grund til at inddrage variationen af de totale antal i den statistiske model for data. Istedet kan man vælge at modellere N_{ijkl0} som binomialfordelt med antalsparameter $N_{ijkl\cdot}$ og sandsynlighedsparameter p_{ijkl}. Imidlertid gælder endvidere, at værdierne for N_{ijkl0}'erne i overvejende grad (86 % af tilfældene) enten er lig 0 eller $N_{ijkl\cdot}$. Det vil sige, at for en knop vil der typisk gælde, at enten er alle blomster friske eller også er alle blomster døde. Det er ikke så interessant at modellere degenererede binomialfordelinger, så istedet kan man anskue afledte variable $Y_{ijkl} = 1(N_{ijkl1} > N_{ijkl0})$. Det vil sige, at Y_{ijkl} er en indikator for, at flertallet af blomster på den $ijkl$'te knop er døde. Man kan dernæst vælge at modellere $\tilde{N}_{ijk} = \sum_{l=1}^{n_{ijk}} Y_{ijkl}$ som binomialfordelt $b(n_{ijk}, p_{ijk})$ eller ækvivalent, Y_{ijkl} som binomialfordelt $b(1, p_{ijk})$ hvor p_{ijk} er sandsynligheden for, at den ijk'te knop er død.

På forhånd forventes det, at variationen i data ikke kun skal henføres til effekterne af de forskellige grundstammer, men at der også er en heterogenitet mellem grupper af observationer tilhørende samme gren eller samme træ, hvor denne heterogenitet skyldes korrelation mellem knopper på samme gren eller træ. For at undersøge dette kan man lave tests for homogenitet af andel friske/døde knopper over grene indenfor de enkelte parceller og over parceller indenfor hver grundstamme (de relevante SAS-kommandoer er vist i appendix A). Testene for homogenitet over grene indenfor hver parcel er ikke signifikante pånær for parcel nr. 54. Testene for homogenitet af parceller indenfor grundstamme er signifikante for grundstamme 1 og 4, hvor henholdsvis parcellerne 41 og 59 skiller sig ud fra de øvrige parceller. Laves et test for homogenitet over grene indenfor grundstamme fås, at hypotesen om homogenitet kun kan accepteres for grundstamme 2. Dette synes umiddelbart ikke at være i tråd med de øvrige homogenitets-test. Dog skal man huske på, at testene for homogenitet over grene indenfor parceller er baseret på små antal observationer, og at inhomogenitet på grenniveau kan sløres, når observationerne akkumuleres på parcel-niveau. En heterogenitet, der ikke kan henføres til en systematisk behandlingseffekt, modelleres sædvanligvis vha. tilfældige faktorer. Dette eksemplificeres i næste afsnit.

Endeligt skal det bemærkes, at den indledende analyse af data giver et indtryk af, at andelen af døde blomster på grene med værdien 1 for faktoren GRENNR generelt er lavere end for de øvrige grene. Dette kan skyldes en reel systematisk effekt af måden, hvorpå grenene er udvalgt for hver parcel.



Figur 1: Plot af log-transformerede empiriske varianser mod log-transformerede empiriske middelværdier.

3 Generaliserede lineære mixed modeller

Kirsebærblomst-datamaterialet blev indsamlet med henblik på at undersøge om valget af grundstamme influerede på blomsternes modstandsdygtighed overfor nattefrost. Antages i første omgang at Y_{ijk}'erne er uafhængige og ens fordelte for fastholdt i (dvs. $p_{ijk} = p_i$) kan hypotesen $p_i = p$ (dvs. ingen effekt af grundstamme) undersøges vha. et test for uafhængighed af inddelingskriterier, som let udføres vha PROC FREQ. Med denne simple fremgangsmåde er det dog ikke muligt nærmere at undersøge evt. forskelle mellem forskellige grundstammer. Alternativt kan man vælge eksplicit at modellere data som en generaliseret lineær model. I det konkrete tilfælde modellerer man logit til p_{ijk} som en lineær regression:

$$\text{logit}(p_{ijk}) = \log(\frac{p_{ijk}}{1 - p_{ijk}}) = \alpha + \beta_i. \tag{1}$$

En stor værdi af parameteren β_i svarer dermed til en stor værdi af p_{ijk}, altså, at en knop på en grundstamme af i'te type har en stor risiko for at dø.

Som nævnt forudsætter ovenstående model, at datamaterialet er homogent indenfor hver grundstamme gruppe. Dette er jvf. den indledende undersøgelse af data i afsnit 2 ikke tilfældet, idet der udover den mulige systematiske grundstamme effekt synes at være en del yderligere variation mellem grupper af observationer tilhørende forskellige grene og træer. Denne yderligere variation kan f.eks. tilskrives biologisk variation, skadedyrsangreb, omgivelserne o.a. For normalfordelte data modelleres denne type variation som oftest vha. tilfældige effekter. Dette kan for generaliserede lineære modeller gøres ved at tilføje tilfældige effekter på højresiden af (1), dvs.

$$\text{logit}(p_{ijk}) = \alpha + \beta_i + \sigma_{\text{gren}} U_{ijk} + \sigma_{\text{træ}} U_{ij}, \tag{2}$$

hvor variablene U_{ijk} og U_{ij} er uafhængige standard normal fordelte. Derved opnås en såkaldt generaliseret lineær mixed model. For den ijk'te gren beskriver U_{ijk} den tilfældige effekt associeret med den pågældende gren. Størrelsen af de såkaldte varianskomponenter σ_{gren} og $\sigma_{\text{træ}}$ bestemmer graden af heterogenitet mellem henholdsvis grene og træer (med samme grundstamme).

Generaliserede lineære modeller kan analyseres vha. af SAS macroen GLIMMIX (Littell *et al.*, 1996). I appendix B er gengivet et SAS-program, der benytter GLIMMIX til at analysere datasættet indeholdende variablene Y_{ijkl}. Bemærk, at GLIMMIX sætter $\beta_5 = 0$ og fastholder β_5 på denne værdi. Dermed er døds-sandsynligheden for en knop på 5'te grundstamme givet ved $\text{logit}(p_{5jk}) = \alpha + \sigma_{\text{gren}} U_{5jk} + \sigma_{\text{træ}} U_{5j}$. Dvs. værdierne af β_l, $l = 1, \ldots, 4$ bestemmer den systematiske forskel mellem de første 4 grundstammer og den femte, med hensyn til sandsynligheden for at en knop dør. Tabel 1 indeholder parameter-estimater, std. errors og testsandsynligheder for hypoteserne $\alpha = 0$ og $\beta_l = 0$, $l = 1, \ldots, 4$. Det ses, at kun grundstamme 2 er signifikant forskellig fra grundstamme 5. Udfra estimaterne for varianskomponenterne ses, at den største andel af den tilfældige variation forekommer på grenniveau. GLIMMIX beregner ikke std. errors og tests for varianskomponenterne. F-testet for effekten af grundstamme er ikke signifikant på 5 %-niveauet. Dette kan skyldes

Parameter	estimat	std. error	test ssh.
α	0.8279	0.5154	0.1291
β_1	-0.3137	0.7119	0.6599
β_2	1,7804	0.8254	0.0320
β_3	-0.0966	0.7164	0.8929
β_4	-0.3790	0.7485	0.6131
β_5	0	-	-
σ^2_{gren}	1.4769	-	-
$\sigma^2_{\text{træ}}$	0.1233	-	-

Tabel 1: Parameter estimater, std. errors og testsandsynligheder.

at effekten af grundstamme 2 “forsvinder” i det overordnede billede pga. ligheden mellem de øvrige 4 grundstammer. De beregnede test-sandsynligheder skal dog tages med et gran salt jvf. resultaterne af modelkontrollen i næste afsnit.

3.1 Modelkontrol

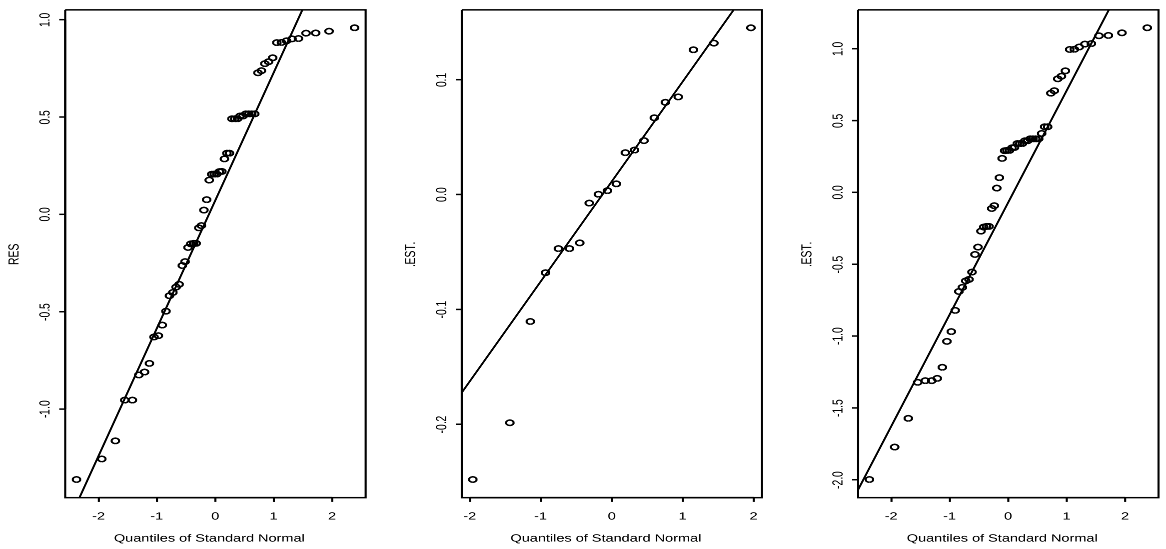
En modelkontrol kan udføres vha. probit-diagrammer for de estimerede tilfældige effekter og for Pearson-residualerne udregnet for den betingede fordeling af data givet de tilfældige effekter. Pearson-residualerne er udregnet for de observationer, der fremkommer ved at akkumulere data indenfor grupper med fælles værdi af systematiske og tilfældige effekter (dvs. de totale antal døde knopper på hver gren). Probit-diagrammerne tjener primært til at afsløre outliere, da man ikke skal regne for meget med, at de estimerede tilfældige effekter og residualerne er normal-fordelte. SAS-koden til beregning af de relevante datasæt er vist i appendix C, og plottene er vist i Figur 2. Det midterste probit-diagram i Figur 2 indeholder 2 outliere, der netop svarer til parcel 41 og 59, som også blev udpeget som atypiske i afsnit 2.

Modellen kan yderligere checkes ved at plotte de forventede antal af døde knopper for hver gren mod de forventede værdier under den fittede model. Afvigelserne mellem de observerede og de fittede værdier i venstre plot i Figur 3 udviser en klar systematik, og denne systematik går igen i det højre plot i Figur 3, der viser Pearson-residualerne mod de fittede værdier. Det ses, at residualerne er negative for små fittede værdier og positive for de store fittede værdier. Punkterne i højre plot i Figur 3 ligger i grupper svarende til de mulige observerede procent-andele. Den systematiske placering af punkterne indenfor hver af disse grupper er en naturlig følge af datas diskrete natur, men systematikken mht. residualernes fortegn viser, at modellen ikke giver et tilfredstillende fit til data.

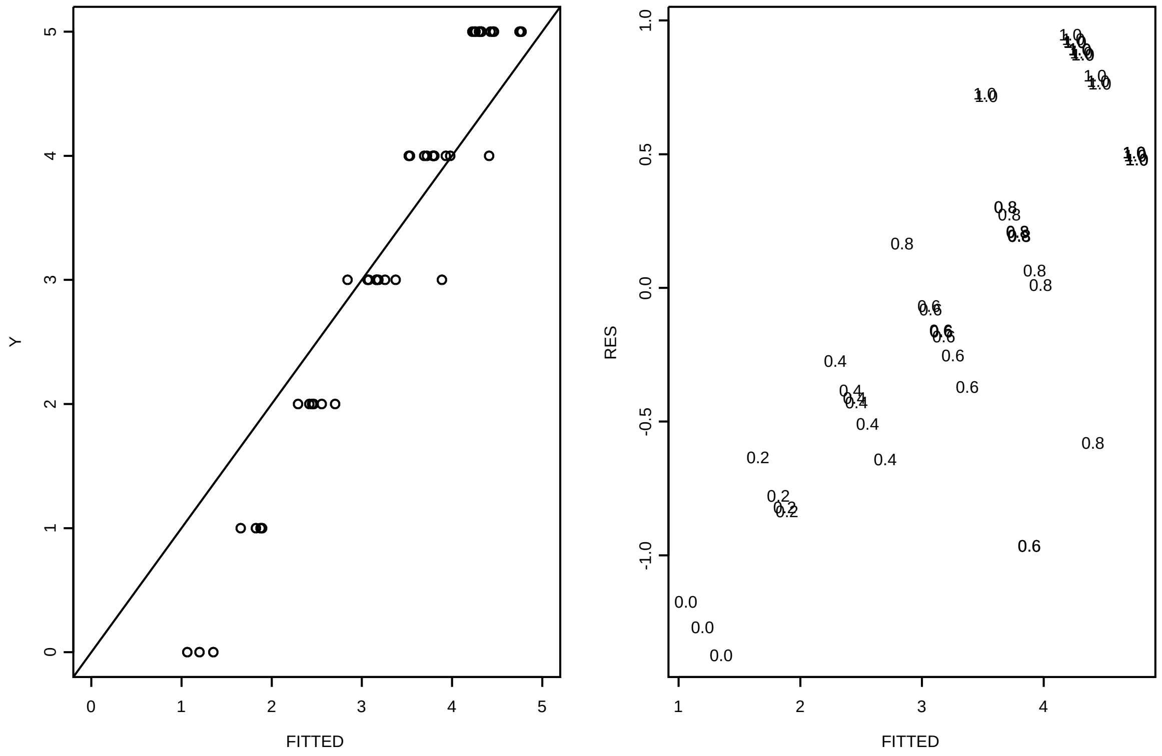
For at undersøge effekten af de to atypiske parceller 41 og 59 er GLIMMIX kørt igen men med observationerne fra disse parceller udeladt. Endvidere er faktoren GRENNR medtaget som en systematisk faktor (jvf. sidste bemærkning i afsnit 2). I dette tilfælde estimeres variankomponenterne for parcel og gren til at være henholdsvist 0 og 1.1228, mens faktoren GRENNR ikke var signifikant. Plottene svarende til Figur 3 men for den modificerede model er vist i Figur 4. Igen er der problemer med systematisk bias af residualerne. For at fitte de små og de store observerede procent-andele kræves meget små henholdsvis store værdier for gren-effekterne, og dermed en stor værdi af varianskomponenten. Omvendt er det almindeligt kendt, at varianskomponenten kan blive underestimeret, når antallet af binære observationer er lille for hvert niveau af den tilfældige faktor (se diskussionsbidragene til Lee and Nelder, 1996).

Udelades begge tilfældige effekter fra modellen fås en sædvanlig generaliseret lineær model. Fittes denne model (igen med udeladelse af parcel 41 og 59) fås, at både GRENNR og GRST er signifikante med test-sandsynligheder på henholdsvis 0.0068 og 0.021. Et grafisk check af denne model er givet i Figur 5, hvor residualer og observerede/forventede værdier er beregnet for data akkumuleret indenfor hvert niveau af krydsfaktoren GRENNR* GRST (Figur 5 er således ikke sammenlignelig med Figur 4).

Det ses, at testsandsynlighederne for GRENNR og GRST afhænger kraftigt af, om PARCEL*GRENNR medtages som en tilfældig faktor eller ej. Faktoren GRENNR er ikke signifikant, når PARCEL*GREN optræder som tilfældig faktor. På den anden side er det også muligt, at PARCEL*GREN ikke er signifikant, således, at man ved at udelade PARCEL*GREN ville nå frem til den generaliserede lineære model givet ved faktorerne GRENNR og GRST. Det er derfor beklageligt, at et test for effekten af PARCEL*GRENNR ikke er til rådighed. Tests for homogenitet af andel friske/døde knopper over grene med samme værdi af GRENNR indenfor grundstamme er insignifikante. Dette tyder på, at gren-heterogeniteten kan beskrives udfra faktoren GRENNR, således, at modellen uden tilfældige faktorer er at foretrække.



Figur 2: Probit-diagrammer for Pearson-residualer (venstre), tilfældige træ-effekter (midten) og tilfældige gren-effekter (højre).



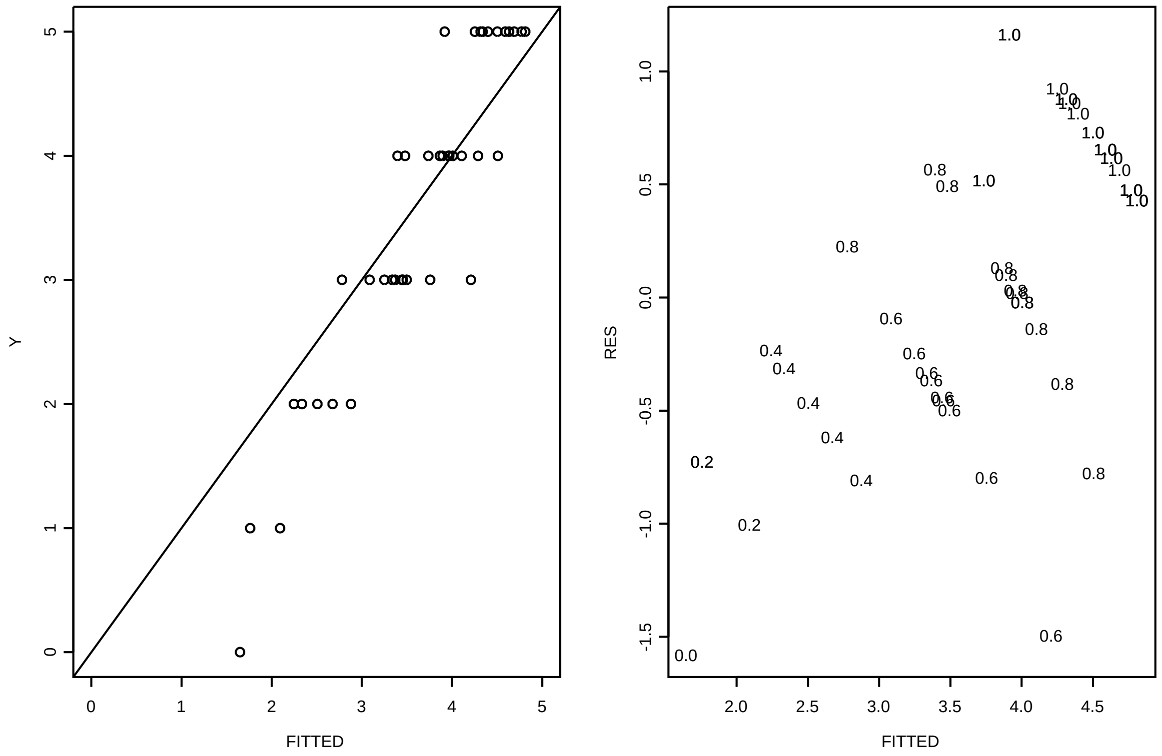
Figur 3: *Venstre:* observerede værdier mod fittede værdier. *Højre:* plot af Pearson-residualer mod fittede værdier. De observerede værdier er antal døde knopper på hver gren, og tallene i plottet angiver de tilsvarende procent-andele af døde knopper.

3.2 Klassen af generaliserede lineære mixed modeller

Varianskomponentmodellen beskrevet i det foregående afsnit er et eksempel på en generaliseret lineær mixed model (GLMM). En GLMM defineres generelt udfra den såkaldte lineære prediktor, link-funktionen og endelig den betingede fordeling af data givet værdierne af de tilfældige effekter. For en GLMM og observationer $Y_1, \ldots, Y_n$ er den lineære prediktor η_i for Y_i på formen

$$\eta_i = \sum_{l=1}^{p} \beta_l x_{il} + \sum_{k=1}^{q} \sigma_k U_k z_{ik} \tag{3}$$

hvor x_{il} og z_{ik} er de forklarende variable for Y_i mht. henholdsvis de systematiske og de tilfældige effekter. For fastholdte værdier af de tilfældige variable U_k er middelværdien af Y_i givet ved $E(Y_i) = g^{-1}(\eta_i)$ hvor g er link-funktionen. Den betingede fordeling af Y_i givet værdierne af U_k'erne er en generaliseret lineær model. Dermed kan Y_i givet U_k, $k = 1, \ldots, q$, eksempelvis være normal, Poisson, binomial, gamma eller invers Gauss fordelt. De såkaldte



Figur 4: *Venstre:* observerede værdier mod fittede værdier for den modificerede model og reducerede data-sæt. *Højre:* plot af Pearson-residualer mod fittede værdier for den modificerede model og reducerede data-sæt. De observerede værdier er antal døde knopper på hver gren, og tallene i plottet angiver de tilsvarende procent-andele af døde knopper.

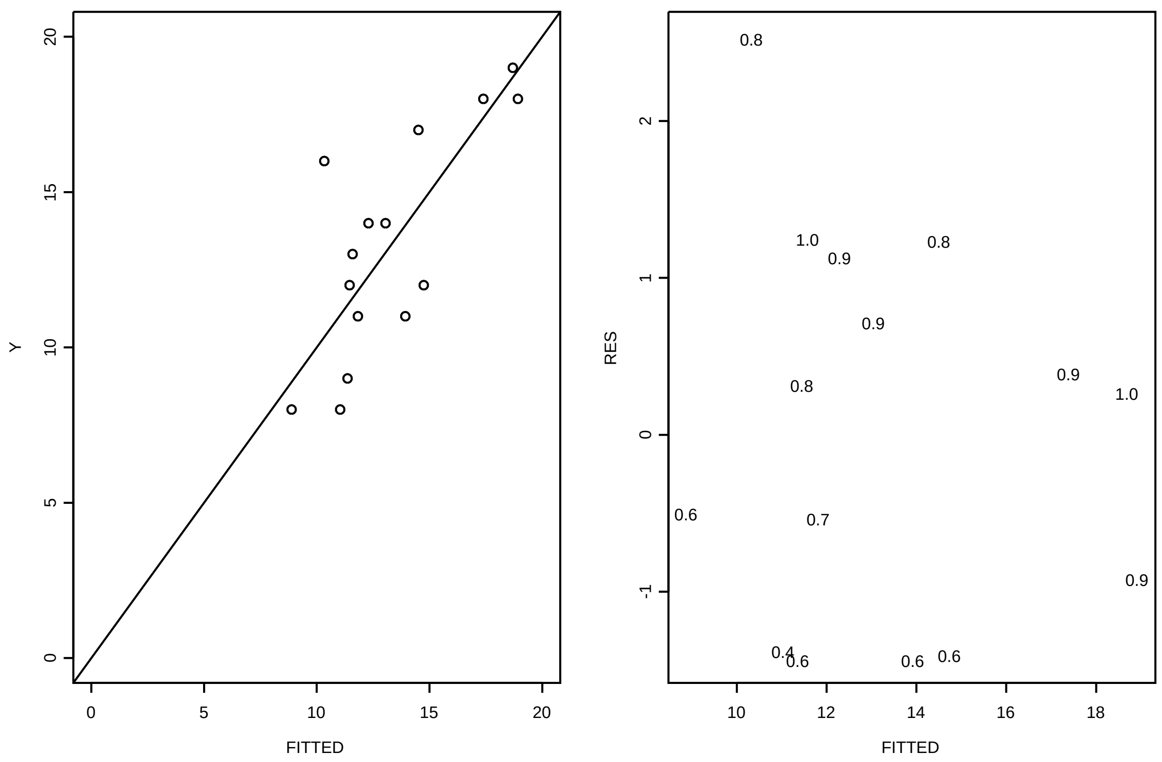
kanoniske link-funktioner for normal, Poisson og binomial fordelingen er henholdsvis identiteten ($g(x) = x$), log og logit. En sædvanlig normal mixed model opnås, når Y_i givet U_k'erne er normalfordelt, og den kanoniske link-funktion anvendes.

GLMM'er er en substantiel generalisering af de normale mixed modeller pga. den store fleksibilitet mht. valg af link-funktion og valg af fordelingen for Y_i'erne givet de tilfældige effekter.

A Analyse for homogenitet mellem grene

Indledningsvist konstrueres et datasæt med variablene Y_{ijkl} (det oprindelige data-sæt gengivet i appendix D hedder viki3):

```
/* danner indikator variabel Y_ijkl  */
data viki4;
```



Figur 5: *Venstre:* observerede værdier mod fittede værdier for model uden tilfældige effekter. *Højre:* plot af Pearson-residualer mod fittede værdier for model uden tilfældige effekter. De observerede værdier er antal døde knopper for hvert niveau af faktoren GRENNR*GRST, og tallene i plottet angiver de tilsvarende procent-andele af døde knopper.

```
set viki3;
pargrekn=parcel*1000+grst*100+grennr*10+knopnr;

proc means data=viki4;
var karakter;
by pargrekn notsorted;
output out=summary5 mean=mean;
run;

proc freq data=viki4;
tables grst*parcel*grennr*knopnr / out=temp;
run;
```

```
data summary5;
set summary5;
merge temp;

data y;
set summary5;
if mean > 0.5 then y=1;
if mean <= 0.5 then y=0;
keep grst parcel grennr knopnr y;
run;
```

Diverse tests for homogenitet beregnes (eksakt test benyttes pga. de lave antal observationer i visse af tabellernes indgange):

```
/* undersoeger homogenitet mellem grene indenfor parcel */
proc sort data=y out=y;
by parcel;
run;

proc freq data=y;
tables grennr*y / exact;
by parcel;
run;

/* undersoeger homogenitet mellem parceller indenfor grst */
proc sort data=y out=y;
by grst;

proc freq data=y;
tables parcel*y /exact;
by grst;
run;

/* undersoeger homogenitet mellem grene indenfor grst */
data y;
set y;
parcgren=parcel*10+grennr;
run;

proc freq  data=y;
tables parcgren*y / exact;
by grst;
run;
```

B Generaliseret lineær model analyse af effekt af grundstamme

Nedenstående SAS-program analyserer data $(Y_{ijkl})_{ijkl}$ vha. en generaliseret lineær mixed model. Specifikationerne for CLASS, MODEL og RANDOM sætningerne er helt som for PROC MIXED. PARMS sætningen bevirker, at en såkaldt ekstra-dispersions parameter ikke estimeres. ERROR og LINK angiver henholdsvis den betingede fordeling af data givet de tilfældige effekter og den valgte link-funktion (se afsnit 3.2). Det er også muligt at indføre ESTIMATE og CONTRAST sætninger på samme måde som i PROC MIXED. Valget af METHOD=REML har man i reglen ikke brug for at ændre på. Bemærk, at krydsfaktoren PARCEL*GRENNR har et niveau for hver enkelt gren i datamaterialet.

```
\%include 'd:\sas612\stat\sample\glimmix.txt';
\%glimmix(data=y,
         procopt=method=reml,
         stmts=\%str(
              class y parcel grst grennr;
              model y=grst / solution;
              random parcel parcel*grennr / solution;
              parms  (1) (1) (1) / eqcons=3;),
         error=binomial,
         link=logit,
         maxit=50,
         output= / out=outstat pred=eta xbeta=mu reschi=res);
run;
```

Outputtet fra ovenstående program er som følger:

```
Class       Levels

Y              2
PARCEL        20
GRST           5
PARCGREN      58
GRENNR         3

Values

0 1
36 38 41 45 47 48 50 53 54 56 58 59 63 134 140 145 149 154 254 354
1 2 3 4 5
361 362 381 382 383 411 412 451 452 453 471 472 473 481 482 483 501
502 503 531 532 533 541 54
1 2 3
```

Covariance Parameter Estimates

Cov Parm	Estimate
PARCEL	0.12325925
PARCEL*GRENNR	1.47688158

GLIMMIX Model Statistics

Description	Value
Deviance	225.8749
Scaled Deviance	225.8749
Pearson Chi-Square	177.5865
Scaled Pearson Chi-Square	177.5865
Extra-Dispersion Scale	1.0000

Parameter Estimates

Effect	GRST	Estimate	Std Error	DF	t	Pr > |t|
INTERCEPT		0.8279	0.5154	15	1.61	0.1291
GRST	1	-0.3137	0.7119	229	-0.44	0.6599
GRST	2	1.7804	0.8254	229	2.16	0.0320
GRST	3	-0.0966	0.7164	229	-0.13	0.8929
GRST	4	0.3790	0.7485	229	0.51	0.6131
GRST	5	0.0000	.	.	.	

.

Random Effects Estimates

Effect	PARCEL	GRENNR	Estimate	SE Pred	DF	t	Pr > |t|
PARCEL	36		0.0033	0.3379	229	0.01	0.9921
PARCEL	38		0.0668	0.3335	229	0.20	0.8415
PARCEL	41		-0.1987	0.3382	229	-0.59	0.5574
PARCEL	45		-0.0421	0.3345	229	-0.13	0.8999
PARCEL	47		0.1259	0.3374	229	0.37	0.7093

(estimater er udeladt for at spare plads)

PARCEL*GRENNR	254	2	0.9957	0.9262	229	1.08	0.2835
PARCEL*GRENNR	254	3	0.2920	0.8537	229	0.34	0.7326

```
  PARCEL*GRENNR  354     1    1.0354     0.9204    229     1.12  0.2618
  PARCEL*GRENNR  354     2   -0.2370     0.8136    229    -0.29  0.7711
  PARCEL*GRENNR  354     3   -0.2370     0.8136    229    -0.29  0.7711

                             Tests of Fixed Effects

                  Source    NDF    DDF    Type III F    Pr > F

                   GRST       4    229          1.93    0.1066
```

C Modelkontrol

Nedenstående kode genererer data-sæt med henholdsvis Pearson-residualerne og de estimerede tilfældige effekter. Pearson-residualerne er udregnet for data akkumuleret indenfor grupper med fælles værdi af systematiske og tilfældige effekter. Data-sættene `outstat` og `_SolnR` blev genereret af GLIMMIX-kørslen i appendix B.

```
/* udregner Pearson-residualer for akkumulerede data */
proc means data=outstat;
var y;
by parcgren notsorted;
output out=sumy sum=;
run;

proc means data=outstat;
var mu;
by parcgren  notsorted;
output out=mu mean=;
run;

data pearson;
set sumy;
merge mu;
N=_freq_;
res=(y-N*mu)/sqrt(N*mu*(1-mu));
run;

/* tilfaeldige effekter paa trae-niveau */
data randparc;
set _SolnR;
if _effect_='PARCEL';
run;
```

```
/* tilfaeldige effekter paa gren-niveau */
data randpagr;
set _SolnR;
if _effect_='PARCEL*GRENNR';
run;
```

D Data

Nedenfor er angivet de første 110 observationer i data. Det fuldstændige datasæt indeholder 953 observationer.

OBS	GRST	PARCEL	KNOPNR	GRENNR	KARAKTER
1	1	38	1	1	1
2	1	38	1	1	1
3	1	38	1	1	1
4	1	38	1	1	1
5	1	38	2	1	0
6	1	38	2	1	0
7	1	38	2	1	0
8	1	38	2	1	1
9	1	38	3	1	1
10	1	38	3	1	1
11	1	38	3	1	1
12	1	38	4	1	0
13	1	38	4	1	1
14	1	38	4	1	1
15	1	38	4	1	1
16	1	38	5	1	0
17	1	38	5	1	0
18	1	38	5	1	1
19	1	38	1	2	1
20	1	38	1	2	1
21	1	38	1	2	1
22	1	38	1	2	1
23	1	38	2	2	0
24	1	38	2	2	1
25	1	38	2	2	1
26	1	38	3	2	1
27	1	38	3	2	1
28	1	38	3	2	1
29	1	38	3	2	1
30	1	38	4	2	0

31	1	38	4	2	0
32	1	38	4	2	0
33	1	38	5	2	1
34	1	38	5	2	1
35	1	38	5	2	1
36	1	38	1	3	1
37	1	38	1	3	1
38	1	38	1	3	1
39	1	38	1	3	1
40	1	38	2	3	1
41	1	38	2	3	1
42	1	38	2	3	1
43	1	38	3	3	1
44	1	38	3	3	1
45	1	38	3	3	1
46	1	38	3	3	1
47	1	38	4	3	1
48	1	38	4	3	1
49	1	38	4	3	1
50	1	38	4	3	1
51	1	38	5	3	0
52	1	38	5	3	0
53	1	38	5	3	1
54	1	48	2	1	0
55	1	48	2	1	0
56	1	48	2	1	0
57	1	48	2	1	0
58	1	48	3	1	1
59	1	48	3	1	1
60	1	48	3	1	1
61	1	48	4	1	1
62	1	48	4	1	1
63	1	48	5	1	1
64	1	48	5	1	1
65	1	48	5	1	1
66	1	48	1	2	1
67	1	48	1	2	0
68	1	48	1	2	0
69	1	48	1	2	0
70	1	48	2	2	1
71	1	48	2	2	1
72	1	48	2	2	1
73	1	48	2	2	1

74	1	48	3	2	1
75	1	48	3	2	1
76	1	48	3	2	1
77	1	48	4	2	1
78	1	48	4	2	1
79	1	48	4	2	1
80	1	48	4	2	1
81	1	48	5	2	1
82	1	48	5	2	1
83	1	48	5	2	1
84	1	48	5	2	1
85	1	48	1	3	1
86	1	48	1	3	1
87	1	48	1	3	1
88	1	48	2	3	1
89	1	48	2	3	1
90	1	48	2	3	1
91	1	48	2	3	1
92	1	48	3	3	1
93	1	48	3	3	1
94	1	48	3	3	1
95	1	48	3	3	1
96	1	48	4	3	1
97	1	48	4	3	1
98	1	48	4	3	1
99	1	48	4	3	1
100	1	48	5	3	0
101	1	48	5	3	1
102	1	48	5	3	1
103	1	50	1	1	0
104	1	50	1	1	0
105	1	50	1	1	0
106	1	50	2	1	0
107	1	50	2	1	0
108	1	50	2	1	0
109	1	50	2	1	0
110	1	50	3	1	0

Litteratur

Breslow, N. E. and Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *J. Amer. Statist. Assoc.* , 9–25.

Lee, Y. and Nelder, J. A. (1996). Hierarchical generalized linear models. *J. R. Statist. B* **58**, 619–678.

Littell, R. C., Milliken, G. A., Stroup, W. W. and Wolfinger, R. D. (1996). *SAS system for mixed models*. SAS Institute Inc., Cary, NC.

McCullagh, P. and Nelder, J. A. (1989). *Generalized linear models*. Chapman and Hall, London, 2nd edition.

Nelder, J. A. and Wedderburn, R. W. M. (1972). Generalized linear models. *J. R. Statist. Soc. A* **135**, 370–384.