Kapitel 13

Generaliserede lineære modeller

13.1. Introduktion.

Betegnelsen *generaliserede lineære modeller* anvendes bredt om statistiske modeller for uafhængige observationer $y_1, \ldots, y_n$, hvor middelværdien af den enkelte observation har formen

$$\mathrm{E}y_i = \mu_i = m(\eta_i)$$

hvor η_i står for et lineært udtryk af den slags, der kunne stå på højre side af lighedstegnet i specifikationen af middelværdien i en almindelig lineær normalfordelingsmodel. Funktionen m er i denne sammenhæng en fast, monoton (sædvanligvis voksende) funktion, som vi vil kalde *middelværdifunktionen.*

Af forskellige (blandt andet historiske) grunde foretrækker man ofte at formulere dette ved hjælp af funktionen m^{-1}, som i denne sammenhæng kaldes *linkfunktionen.* Formuleringen bliver så at

$$m^{-1}\left(\mathrm{E}y_i\right) = m^{-1}(\mu_i) = \eta_i (= \text{lineært udtryk}).$$

De generaliserede lineære modeller dannes således ud fra de lineære normalfordelingsmodeller ved følgende udvidelser:

(1) Fordelingerne af observationerne kan være af andre typer end den normale. Vi skal kun se på de to vigtigste tilfælde: Poissonfordelte observationer og binomialfordelte observationer. For begge disse fordelingstyper gælder, at de kan parametriseres ved deres middelværdi (idet man for binomialfordelingens vedkommende opfatter antalsparametrene som givne).

(2) Udtrykket for den enkelte observations middelværdi μ_i er ikke i sig selv lineært, det er middelværdien transformeret med linkfunktionen $\eta_i = m^{-1}(\mu_i)$, der kan skrives som et lineært udtryk (dvs. som sum af lineære effekter af baggrundsvariable, hovedvirkninger af faktorer, vekselvirkninger mellem faktorer osv.).

Vi vil endvidere holde os til følgende to valg af middelværdifunktion og linkfunktion:

For Poissonfordelte observationer betragtes middelværdifunktionen

$$\mu = m(\eta) = \exp(\eta)$$

svarende til linkfunktionen

$$\eta = m^{-1}(\mu) = \log(\mu).$$

Dette fører til klassen af *log–lineære* eller *multiplikative* Poissonmodeller. Disse modeller behandles i afsnit 13.2.

For binomialfordelte observationer — eller rettere: for observationer som er relative hyppigheder $y_i = y_i'/m_i$ svarende til binomialfordelte observationer y_i' med antalsparametre m_i — betragtes middelværdifunktionen

$$\mu = m(\eta) = \frac{\exp(\eta)}{1 + \exp(\eta)}$$

svarende til linkfunktionen

$$\eta = m^{-1}(\mu) = \log \frac{\mu}{1-\mu} = \text{logit}(\mu).$$

Dette fører til klassen af *logit–lineære* modeller, også kaldet de *logistiske regressionsmodeller.* Disse modeller behandles i afsnit 13.3.

Fordelen ved på denne måde at opfatte en række andre modeller som generaliseringer af de lineære normalfordelingsmodeller er, at en stor del af begrebsapparatet fra de lineære modeller kan overtages. Det gælder i særdeleshed hele maskineriet omkring specifikation af modellen ved hjælp af en modelmatrix, overparametrisering og udtynding til lineær uafhængighed. Men også begreber som vekselvirkning, additivitet, parallelle regressionslinier osv. får en mening i denne generelle ramme, selvom den i nogle tilfælde er lidt anderledes end den vi kender fra de lineære modeller. Endelig kan kodningen af et lineært udtryk ved en symbolsk modelformel (jvf. side 109–111) umiddelbart overtages fra de lineære modeller.

På den beregningsmæssige side er der også en række fordele, som historisk set har vejet endnu tungere. Det viser sig, at maksimering af log–likelihood'en i disse modeller kan foretages ved en iterativ procedure, hvor hvert enkelt trin i alt væsentligt er ækvivalent med løsningen af et lineært regressionsproblem, givet ved den samme modelmatrix. Derfor, hvis man har udviklet et program, der kan håndtere den lineære normalfordelingsmodel, med alt hvad det involverer af oversættelse af modelformel til modelmatrix, udtynding til lineær uafhængighed og løsning af normalligningerne, udskrift af parameterestimater, dannelse af fittede værdier og residualer osv., så har man i virkeligheden udført 90% af det arbejde der skal til for at gøre præcis det samme for den meget større klasse af generaliserede lineære modeller. Dette var baggrunden for en gammel, idag næsten glemt, succeshistorie i forbindelse med programpakken GLIM, udviklet i 1970'erne og starten af 1980'erne af "The

Working Party on Statistical Computing of the Royal Statistical Society". Der er ikke mange som bruger GLIM idag, men selve algoritmen lever i bedste velgående i SAS (PROC GENMOD), GENSTAT (MODEL, FIT) og, i al beskedenhed, ISU (FITLOGLIN, FITLOGIT, FITNONLIN). Ved gennemgangen i det følgende illustreres modellerne ved gennemregning af eksempler i ISU.

13.2. Log–lineære modeller.

Log–lineære modeller er en samlebetegnelse som primært bruges om de multiplikative Poissonmodeller. Men som forklaret i kapitel 6 kan polynomialfordelingsmodeller fortolkes som Poissonmodeller, hvor der er betinget med totalsummen eller visse marginalsummer. Derfor er klassen af modeller, som kan håndteres inden for rammerne af de multiplikative Poissonmodeller, i virkeligheden større end det umiddelbart lyder. Men i beskrivelsen af modellerne holder vi os her til tilfældet, hvor observationerne er Poissonfordelte.

Lad $y_1, \ldots, y_n$ være uafhængige, Poissonfordelte med parametre (eller middelværdier, det er jo det samme) $\mu_1, \ldots, \mu_n$. Hvis middelværdivektoren $\mu = (\mu_1, \ldots, \mu_n)$ varierer frit i $]0, +\infty[^n$ taler vi om den *fulde* model. En multiplikativ eller log–lineær model defineres nu ved, at vi for en given $n \times p$ modelmatrix X antager $\eta = X\beta$, hvor η er $n \times 1$ (søjle–) vektoren med elementer $\eta_i = \log(\mu_i)$ og $\beta = (\beta_1, \ldots, \beta_p)$ er en ($p \times 1$ søjle–) vektor af ukendte parametre. Altså, på koordinatvis form,

$$\log \mu_i = \eta_i = \beta_1 x_{i1} + \cdots + \beta_p x_{ip}.$$

Eksempel 13.1.

Hvis data foreligger i form af en tosidet antalstabel

$$y_{rs}, \ r = 1, \ldots R, \ s = 1, \ldots S$$

er den ved

$$\eta_{rs} = \log(\mathrm{E} y_{rs}) = \alpha_r + \beta_s$$

bestemte model åbenbart (på nær ændring af notationen, idet de parametre som på side 55 hed α_r og β_s nu er $\exp(\alpha_r)$ og $\exp(\beta_s)$) ækvivalent med den multiplikative Poissonmodel vi indførte i kapitel 6. Den lineære struktur er præcis den samme som i den additive tosidede variansanalysemodel (side 106), og problemet med overparametrisering kan naturligvis håndteres på samme måde.

I ISU estimeres log–lineære modeller ved hjælp af kommandoen `FITLOGLINEAR`. Vi illustrerer ved at gennemføre den analyse af en tabel fra Falck datasættet som findes på side 40–44. Vi antager at indlæsningen har fundet sted, således at vi har rådighed over en variate `ANTAL` af længde 16 som indeholder de 16 tal i tabellen, og to faktorer ANC og HOLDN af længde 16 på 4 niveauer, der fortæller hvilke celler disse tal hører til i. Således at kommandoen

```
LIST ANTAL::0 ANC HOLDN
```

ville producere et output af følgende udseende:

```
ANTAL  ANC HOLDN
  546    1     1
  488    1     2
   40    1     3
    3    1     4
  495    2     1
  790    2     2
   93    2     3
    7    2     4
  527    3     1
  880    3     2
   85    3     3
    9    3     4
  563    4     1
  666    4     2
   55    4     3
    6    4     4
```

Estimation af den log–lineære model foretages nu ved

```
FITLOGLINEAR antal=1+anc+holdn
```

der producerer følgende output, som blandt andet indeholder testet mod den fulde model (både i likelihood–ratio og Pearson versionen):

```
Convergence after 8 iterations.
16 observations, 7 parameters estimated.
-2log(Likelihood) =            88.4635
Likelihood ratio test against full model
P[ ChiSquare(9) > -2log(Likelihood) ] =  0.000000
Pearsons's chi-square test for goodness of fit
P[ ChiSquare(9) >      88.8797 ] =    0.000000
```

Modellen kan altså ikke godkendes. Der er — som vi nu kan tillade os at kalde det — vekselvirkning mellem anciennitet og holdning. Eller, sagt mere forståeligt, svarmønstret er ikke det samme i de fire anciennitetsgrupper. Alligevel udskriver vi parameterestimaterne for illustrationens skyld. Kommandoen

```
LISTPARAMETERS
```

giver følgende output:

```
                  Estimate       Std.dev.          U          P
CONSTANT             1.815         0.2015      9.008   0.000000
  ANC[1]           -0.1805        0.04128     -4.372   0.000012
  ANC[2]            0.0711        0.03869      1.836   0.066297
  ANC[3]            0.1515        0.03797      3.990   0.000066
  ANC[4]       set to zero
HOLDN[1]             4.445         0.2012     22.098   0.000000
HOLDN[2]             4.727         0.2009     23.531   0.000000
HOLDN[3]             2.391         0.2090     11.441   0.000000
HOLDN[4]       set to zero
```

Konventionerne for hvilke parametre der sættes til 0 er de samme som vi kender fra kapitel 10. Til illustration kan vi udregne den estimerede middelværdi af antallet af personer i anciennitetsgruppe 3 som er "meget utilfredse":

$$\exp(1.815 + 0.1515 + 0) = 7.15.$$

Vi kan trække fittede værdier, residualer og normerede residualer ud af modellen med kommandoen

```
SAVEFITTED fit res nres
```

Tabellen med de normerede residualer nederst side 41, som viser noget om hvor afvigelserne fra den multiplikative model ligger, kan herefter produceres ved hjælp af kommandoen

```
TWOWAYTABLE nres anc=,0-5,6-10,11-20,>20 holdn=,--,-,+,++
```

EKSEMPEL 13.2.

Den model vi i kapitel 6 behandlede under overskriften "proportionalitet med en given baggrundsvariabel" (side 51–54) hører også hjemme blandt de log–lineære modeller. Men det kræver at vi indfører et nyt begreb. Ud over modelmatricen (eller modelformlen) kan man i specifikationen af en generaliseret lineær model angive et *afsæt* (eng. *offset*). Ved et afsæt forstås en vektor af givne tal, der skal adderes til de lineære udtryk η_i på forhånd. I en lineær normalfordelingsmodel, f.eks. den simple regressionsmodel

$$\mathrm{E}y_i = \mu_i = \alpha + \beta x_i$$

kunne vi indføre et afsæt $(a_1, \ldots, a_n)$ ved i stedet at skrive

$$\mathrm{E}y_i = \mu_i = a_i + \alpha + \beta x_i.$$

Man kan tænke på vektoren a som en almindelig regressionsvariabel (ligesom x), der bare har sin koefficient "fastfrosset" (=1) i stedet for ukendt. I forbindelse med de lineære normalfordelingsmodeller er denne konstruktion helt overflødig, fordi man kan opnå præcis det samme ved at erstatte observationerne y_i med de "afsætkorrigerede" observationer $y_i - a_i$. Men i generaliserede lineære modeller spiller begrebet en rolle.

Betragt nu en model for uafhængige Poissonfordelte variable, hvor parametrene for de enkelte fordelinger er givet på formen

$$\mu_i = \beta c_i,$$

hvor $c_1, \ldots, c_n$ er givne tal (som i testikel–kræft eksemplet side 52). Logaritmen til middelværdien får så formen

$$\eta_i = \log c_i + \log(\beta)$$

eller, hvis vi sætter $a_i = \log c_i$ og omdøber $\log \beta$ til β,

$$\eta_i = a_i + \beta.$$

Altså en model hvor den lineære struktur alene indeholder et konstantled (β), men hvor der yderligere er angivet et afsæt $a_i = \log c_i$.

Syntaksen for specifikation af en afsætvariabel i ISU består i at tilføje afsætvariablen med minustegn på modelformlens venstre side (idet en let gennemskuelig analogi med det lineære tilfælde udnyttes). Vi illustrerer ved at gennemregne testikelkræfteksemplet (side 52–54):

```
COMPUTE antal=[4 12 9 9]
COMPUTE popu=[10900 7336 27671 15499]
COMPUTE afs=ln(popu)
FITLOGLINEAR antal-afs=1
```

Den sidste kommando giver anledning til følgende output:

```
Convergence after 12 iterations.
4 observations, 1 parameters estimated.
-2log(Likelihood) =             13.9882
Likelihood ratio test against full model
P[ ChiSquare(3) > -2log(Likelihood) ] =  0.002921
Pearsons's chi-square test for goodness of fit
P[ ChiSquare(3) >      18.8281 ] =    0.000297
```

Vi får altså modellen forkastet. For at se hvad der er gået galt kan vi prøve at LISTe observationerne sammen med fittede værdier og normerede residualer:

```
SAVEFITTED fit * nres
LIST antal:5:0 fit:4:1 nres:4:1
```

Her fås følgende output:

```
ANTAL  FIT NRES
    4  6.0 -0.8
   12  4.1  3.9
    9 15.3 -1.6
    9  8.6  0.1
```

som stærkt antyder, at det er observationen 12 fra Frederikssund der får modellen til at bryde sammen.

Den i opgave 6.5.2 foreslåede analyse kan udføres ved at vi udvider modellen med en "dummy", der tillader Frederikssund at have særstatus:

```
COMPUTE frs=[0 1 0 0]
FITLOGLINEAR antal-afs=1+frs
TESTMODELCHANGE
```

Her får vi følgende output fra FITLOGLINEAR, som viser at de tre andre kommuner kan antages at have samme testikelkræfthyppighed:

```
Convergence after 12 iterations.
4 observations, 2 parameters estimated.
-2log(Likelihood) =               1.5459
```

```
Likelihood ratio test against full model
P[ ChiSquare(2) > -2log(Likelihood) ] =  0.461654
Pearsons's chi-square test for goodness of fit
P[ ChiSquare(2) >       1.6465 ] =    0.439004
```

Af output fra `TESTMODELCHANGE` fremgår det herefter, at reduktion til den foregående model (simpel proportionalitet) er helt umulig:

```
1 parameters added
-2log(Q) =               12.4423
P[ ChiSquare(1) > -2log(Q) ] =  0.000420
```

Eksempel 13.3.

I antalstabeller af orden højere end to optræder et væld af hypoteser med mere eller mindre komplicerede fortolkninger. Vi kan ikke præsentere hele denne teori her. Vi nøjes med at forklare et af dens vigtigste begreber ved et eksempel på en tresidet antalstabel.

Fra Falck–datasættet (hvorfra blandt andet tabellen side 40 er udtrukket) er hentet tre faktorer på fire niveauer: `ANC`, `HOLDN` og `FUNK`. De to første kender vi i forvejen (se side 40), den tredje `FUNK` er svaret på det mere specifikke spørgsmål "Hvor tilfreds er du med din nuværende funktion i Falck?". Svarkategorierne er de samme som for `HOLDN`.

Vi prøver at udskrive den tresidede antalstabel:

```
THREEWAYTABLE anc holdn funk
```

Resultatet vil ikke blive gengivet her, for det viser sig at denne tabel indeholder for mange små tal. Både for `HOLDN` og `FUNK` er der kun meget få som har svaret "meget utilfreds". Vi vil derfor slå de to sidste niveauer sammen til ét for begge disse faktorer. Samtidig danner vi et datasæt af den relevante længde $4 \times 3 \times 3 = 36$, med en variabel `ANTAL` til at indeholde antallene i tabellens celler. Disse komplicerede datatransformationer vil vi nøjes med at gengive uden kommentarer:

```
RENAME anc a
RENAME holdn h
RENAME funk f
FACTOR h3 f3 5418 3
COMPUTE h3=[1 2 3 3](h)
COMPUTE f3=[1 2 3 3](f)
TABULATE antal a*h3*f3 anc*holdn*funk
SAVEDATA falck3 antal anc holdn funk
DELETE
GETDATA falck3
THREEWAYTABLE antal anc holdn funk
```

Fra den sidste kommando fås følgende output, som viser de tal vi vil arbejde videre med:

```
Table of antal sums.

ANC=1:

    FUNK      1    2    3   SUM
HOLDN   |---------------|-----
       1|  390  141    6|  537
       2|  126  326   28|  480
       3|    0   20   21|   41
        |---------------|-----
     SUM|  516  487   55| 1058

ANC=2:

    FUNK      1    2    3   SUM
HOLDN   |---------------|-----
       1|  338  138   12|  488
       2|  172  548   58|  778
       3|    8   29   61|   98
        |---------------|-----
     SUM|  518  715  131| 1364

ANC=3:

    FUNK      1    2    3   SUM
HOLDN   |---------------|-----
       1|  369  142    6|  517
       2|  221  584   65|  870
       3|   10   43   41|   94
        |---------------|-----
     SUM|  600  769  112| 1481

ANC=4:

    FUNK      1    2    3   SUM
HOLDN   |---------------|-----
       1|  383  150   16|  549
       2|  159  450   38|  647
       3|    3   20   35|   58
        |---------------|-----
     SUM|  545  620   89| 1254

All levels of ANC:

    FUNK      1    2    3   SUM
HOLDN   |---------------|-----
       1| 1480  571   40| 2091
       2|  678 1908  189| 2775
       3|   21  112  158|  291
        |---------------|-----
     SUM| 2179 2591  387| 5157
```

Man bemærker at det totale antal 5157 er lidt mindre end i tabellen på side 40. Det skyldes naturligvis, at dem der ikke svarede på spørgsmålet `FUNK` (kodet som niveau 0) er udgået.

Vi prøver først at fitte den største model som ligger under den fulde model, modellen uden tre–faktor vekselvirkning:

```
FITLOGLINEAR antal=1+anc*holdn+anc*funk+holdn*funk
```

Dette fører til følgende output:

```
Convergence after 9 iterations.
36 observations, 24 parameters estimated.
-2log(Likelihood) =            21.1526
Likelihood ratio test against full model
P[ ChiSquare(12) > -2log(Likelihood) ] =  0.048190
Pearsons's chi-square test for goodness of fit
P[ ChiSquare(12) >      17.8046 ] =    0.121755
```

Modellen bliver altså godkendt med et meget lille forbehold. Desværre er den svær at fortolke. Det eneste man kan bruge den til er sådan set at få nogle fittede værdier som er lidt mere "rensede for støj" end selve observationerne. Vi kunne forsøge at kikke på dem:

```
SAVEFITTED fit
THREEWAYTABLE fit anc holdn funk
```

Men vi vil ikke bruge en side mere på at gengive resultatet. Vi nøjes med at referere hvad denne udskrift viser: Den mindste fittede værdi (som findes i den celle hvor der står et nul i den oprindelige tabel) viser sig at være 3.4, og på nær en anden værdi på 4.0 er alle de andre større end 5. Dette gør det forsvarligt at fortsætte testningen på sædvanlig måde.

Næste skridt må være at prøve at fjerne én af de tre tofaktorvekselvirkninger. Det viser sig, at kun én af dem kan fjernes:

```
FITLOGLINEAR antal=1+anc*holdn+holdn*funk
```

giver output

```
Convergence after 9 iterations.
36 observations, 18 parameters estimated.
-2log(Likelihood) =            30.1451
Likelihood ratio test against full model
P[ ChiSquare(18) > -2log(Likelihood) ] =  0.036060
Pearsons's chi-square test for goodness of fit
P[ ChiSquare(18) >      27.0635 ] =    0.077812
```

som i det mindste for Pearson–testets vedkommende giver klar accept. Det mere korrekte successive test (altså mod modellen uden tre–faktor vekselvirkning i stedet for den fulde model), udføres med kommandoen

```
TESTMODELCHANGE
```

som giver output

```
6 parameters removed
-2log(Q) =             8.9925
P[ ChiSquare(6) > -2log(Q) ] =  0.174002
```

Vi konkluderer, at modellen `ANC*HOLDN+HOLDN*FUNK` er fuldtud acceptabel. Herfra viser det sig at man ikke kan komme videre, så det bliver vores slutmodel.

Spørgsmålet er så, hvordan denne model kan fortolkes. Opskrevet i almindeligt formelsprog betyder modellen, at middelværdien af antallet i celle (anc,holdn,funk) har formen

$$\alpha_{\text{anc,holdn}} \times \beta_{\text{holdn,funk}}$$

Det betyder åbenbart, at vi for fastholdt niveau af faktoren `HOLDN` har en almindelig tosidet multiplikativ model i de to andre faktorer. Sagt med reference til polynomialfordelingsfortolkningen, hvor totalsummen 5157 opfattes som given: For ethvert givet niveau af `HOLDN` har vi (i den betingede fordeling, givet niveauet af `HOLDN`) uafhængighed mellem `ANC` og `FUNK`. Dette indebærer for eksempel, at hvis vi stillede os den opgave at forudsige folks anciennitet ud fra svarene på de to holdningsspørgsmål, så ville det være tilstrækkeligt at kende svaret på det generelle tilfredshedsspørgsmål. Den yderligere oplysning om graden af tilfredshed med funktion ville ikke være til nogen hjælp.

Man formulerer denne hypotese kort ved at sige at `ANC` og `FUNK` er betinget uafhængige, givet `HOLDN`. Hypoteser om forskellige former for uafhængighed og betinget uafhængighed er de vigtigste, og stort set de eneste der er til at forstå, i analysen af flerdimensionale antalstabeller.

Bemærk at hypotesen som vi fik godkendt ovenfor også kan fortolkes mere kontant på følgende måde: Hvis vi, for et hvilket som helst fast niveau af faktoren `HOLDN`, danner den tosidede antalstabel svarende til de to andre faktorer, så holder hypotesen om uafhængighed. Det test vi har udført kan altså opfattes som et simultant test for uafhængighed i tre forskellige 4×3–tabeller, nemlig de "lag" i den oprindelige tresidede tabel, der svarer til de tre niveauer af `HOLDN`. Hvis vi helt konkret udførte disse tre tests ved

```
FOCUSONLEVEL HOLDN 1
FITLOGLINEAR ANTAL=1+ANC+FUNK
INCLUDEALL
FOCUSONLEVEL HOLDN 2
...
```

ville vi netop få tre kvotientteststørrelser, hvis sum er 30.1451. Noget tilsvarende gælder for Pearson teststørrelserne. Bemærk overensstemmelsen med χ^2–fordelingens foldningsegenskab (frihedsgradsregnskabet passer, idet 6+6+6=18). Det simultane test for de tre hypoteser foregår altså simpelthen ved en vurdering af de tre tilsvarende teststørrelser, som består i at danne deres sum og vurdere den i sin approksimative fordeling under antagelse af at alle tre hypoteser er opfyldt.

13.3. Logistisk regression.

En klassisk forsøgstype inden for biomedicinsk forskning, som i høj grad var motiverende for udviklingen af de logit–lineære modeller, er det såkaldte dosis–respons forsøg. Dette går i sin simpleste form ud på følgende. Et antal forsøgsdyr (eller mennesker eller planter eller ...), nummereret $i = 1, \ldots n$, udsættes for doser $x_1, \ldots, x_n$ af et eller andet præparat. Doserne antages tilfældigt tilordnet dyrene, således at det for eksempel ikke netop er de største som får høje doser eller lignende. Man observerer så for hvert forsøgsdyr en respons y_i, som er udtryk for om præparatet har haft en bestemt effekt.

Hvis responserne er reelle tal (for eksempel faldet i blodtryk i et forsøg med et formodet blodtrykssænkende præparat) ligger det lige for at opstille en simpel lineær regressionsmodel. Bortset fra at både doserne x_i og responserne y_i muligvis skal transformeres (ofte skal doserne f.eks. angives på logaritmisk skala) går en simpel og naturlig model ud på at antage y_i'erne normalfordelte med middelværdier $\alpha + \beta x_i$ og samme varians σ^2.

I nogle tilfælde (f.eks. hvis det drejer sig om en allergisk reaktion, som enten kan være til stede eller ikke til stede, men er svær at kvantificere) må man lade sig nøje med binære responser af formen $y_i \in \{0, 1\}$, hvor 1 betyder “reaktion” og 0 betyder “ingen reaktion”. Følgende ræsonnement fører til en klasse af modeller, som naturligt kan bruges i denne situation:

Antag at responsen “i virkeligheden” er kontinuert, vi kan bare ikke observere den. Det vil sige, vi har en underliggende naturlig model som er en simpel lineær regressionsmodel, med responser y_i^* som er normalfordelte med middelværdier $\alpha + \beta x_i$ og fælles varians σ^2. De binære responser, som vi observerer, er dannet ud fra de ikke–observerbare y_i^*'er ved “dikotomisering”, nemlig ved

$$y_i = \begin{cases} 1 \text{ for } y_i^* > y_0^* \\ 0 \text{ for } y_i^* \leq y_0^* \end{cases}$$

hvor y_0^* er en ukendt tærskelværdi. Et dyr reagerer hvis og kun hvis dets (ikke–observerbare) kontinuerte respons overstiger denne tærskelværdi.

Selv om idéen om en sådan skjult kontinuert respons i nogle tilfælde kan virke lidt spekulativ, er det vel klart at denne konstruktion giver en fornuftig model til beskrivelse af, hvordan sådanne binære responser kan opføre sig. Fordelingen af y_i'erne er åbenbart givet ved at de er uafhængige, og fordelingen af den i'te bestemt ved

$$P(y_i = 1) = P(y_i^* > y_0^*) = P\left(\frac{y_i^* - (\alpha + \beta x_i)}{\sigma} > \frac{y_0^* - (\alpha + \beta x_i)}{\sigma}\right)$$

$$= 1 - \Phi\left(\frac{y_0^* - (\alpha + \beta x_i)}{\sigma}\right) = \Phi\left(\frac{\alpha - y_0^*}{\sigma} + \frac{\beta}{\sigma} x_i\right)$$

Modellen har fire ukendte parametre (α, β, σ^2 og den ukendte tærskelværdi y_0^*). Men fordelingen af de binære responser afhænger åbenbart kun af disse gennem de to størrelser $\frac{\alpha - y_0^*}{\sigma}$ og $\frac{\beta}{\sigma}$. Det hænger sammen med, at enhver omskalering af y_i^*'erne kan ophæves af en ændring af tærskelværdien y_0^* og hældningen β på en sådan måde at sandsynlighederne $P(y_i = 1)$ ikke ændrer sig. Hvis vi i stedet indfører betegnelserne $\alpha' = \frac{\alpha - y_0^*}{\sigma}$ og $\beta' = \frac{\beta}{\sigma}$ får vi

$$P(y_i = 1) = \Phi(\alpha' + \beta' x_i).$$

Denne model kaldes den *probit–lineære model* (jvf. probit diagrammet; betegnelsen probit bruges løst om den normerede normalfordelings fraktiler, og mere præcist kaldes funktionen Φ^{-1} undertiden for probit funktionen).

I forbindelse med generaliserede lineære modeller kan vi karakterisere denne model som et specialtilfælde (nemlig det, hvor den lineære struktur er den samme som i en simpel lineær regression) af den klasse af generaliserede lineære modeller man får ved som middelværdifunktion at tage funktionen Φ, og som klassen af responsfordelinger alle mulige fordelinger på $\{0, 1\}$. For stokastiske variable med værdier i $\{0, 1\}$ gælder jo $P(y_i = 1) = \mathrm{E}(y_i)$.

Valget af middelværdifunktion i dette tilfælde er lidt arbitrært. Det eneste man nødvendigvis må kræve af middelværdifunktionen i dette tilfælde er at den skal være voksende, og afbilde hele aksen (= det naturlige domæne for alle mulige linearkombinationer af givne baggrundsvariable) på enhedsintervallet (= det naturlige domæne for sandsynligheder, eller middelværdier af binære variable). De logit–lineære modeller, som vi skal beskæftige os med i det følgende, adskiller sig kun fra de probit–lineære modeller ved at middelværdifunktionen $\mu = \Phi(\eta)$ er erstattet med funktionen

$$\mu = \frac{\exp(\eta)}{1 + \exp(\eta)}.$$

De to funktioner ligner hinanden meget (på nær en omskalering, som er ligegyldig for fortolkningen), så i praksis er det nærmest ligegyldigt hvilken af dem man bruger. Men der er nogle fordele ved at vælge den sidste, blandt andet at log–likelihood'en er nemmere at beregne og maksimere.

Den tilhørende linkfunktion, som kaldes *logit funktionen*, ses at være givet ved

$$\eta = \log \frac{\mu}{1 - \mu} = \mathrm{logit}(\mu)\,.$$

Denne funktion har visse pæne algebraiske egenskaber, som gør det lettere at fortolke logit–lineære modeller end probit–lineære modeller. For en 0–1–variabel Y med middelværdi μ har vi jo $\mu = P(Y = 1)$ og dermed

$$\mathrm{logit}(\mu) = \log \frac{P(Y = 1)}{P(Y = 0)}.$$

Ideen med de logit–lineære modeller er således, at de er lineære på den skala, hvor success–sandsynligheder repræsenteres ved det logaritmiske forhold mellem sandsynlighederne for success og fiasko ("log–odds"). Dette valg af "logit–skalaen" som den skala, hvor additivitet og linearitet specificeres, er på mange måder naturligt, blandt andet fordi de logit–lineære modeller har visse pæne relationer til de log–lineære modeller (som vi dog ikke vil gå i detaljer med).

Stokastiske variable som kun antager værdierne 0 og 1 kaldes undertiden for *Bernoulli variable*, og den tilhørende klasse af fordelinger kaldes Bernoulli fordelingerne. En Bernoulli fordeling er åbenbart karakteriseret ved én parameter

$$p = P(Y = 1) = \mathrm{E}Y.$$

En Bernoulli fordeling kan også opfattes som en binomialfordeling med antalsparameter 1. Omvendt er binomialfordelingen pr. definition fordelingen af en sum af uafhængige, identisk fordelte Bernoulli variable.

Vi kan nu give en

FORELØBIG DEFINITION. *En logit–lineær model, også kaldet en logistisk regressionsmodel, er en generaliseret lineær model med linkfunktion* logit *og fordelingstype Bernoulli.*

Vi vil dog straks foretage en mindre udvidelse, som går ud på at tillade generelle binomialfordelte responser, også med antalsparametre som er større end 1. For at få middelværdistrukturen til at passe er vi så nødt til at transformere de binomialfordelte antal til relative hyppigheder ved division med antalsparametrene. Sådanne modeller fremkommer naturligt ved summation af responser i logit–lineære modeller for egentlige binære variable i situationer, hvor observationerne kan opdeles i grupper med samme værdier af alle baggrundsvariable. Hvis for eksempel et dosis–repons forsøg er udført ved at 100 forsøgsdyr hver har fået én ud af 5 mulige doser, idet dyrene på forhånd er (tilfældigt) fordelt i 5 lige store grupper, så ligger det lige for at reducere de 100 binære variable til 5 binomialfordelte variable med antalsparameter 20, dannet som "antal succes'er" inden for hver af de fem grupper. For de fem tilsvarende relative hyppigheder har vi så en logit–lineær model med samme middelværdistruktur som den oprindelige model, blot med en mere generel definition af klassen af mulige responsfordelinger. Bemærk, at hvis der ikke er tale om ægte binære data, så spiller antalsparametrene i binomialfordelingerne naturligvis en rolle for estimationen, og skal derfor indgå i specifikationen af modellen.

Vi vil derfor udvide definitionen til følgende:

DEFINITION. *En logit–lineær model, også kaldet en logistisk regressionsmodel, er en generaliseret lineær model med linkfunktion logit og fordelingstype givet ved, at responserne er relative hyppigheder svarende til* *binomialfordelte variable med kendte antalsparametre; altså $y_i = y_i'/m_i$ hvor y_i' er binomialfordelt med antalsparameter m_i og sandsynlighedsparameter $\mu_i = \mathrm{E}y_i$.*

EKSEMPEL 13.4.

De mest almindelige binomialfordelingsmodeller kan uden videre fortolkes som logit–lineære modeller. Betragt f.eks. tabellen

	Tv.auk.	Ikke tv.auk.
–30	139	985
30–40	240	1587
40–50	256	1982
50–	105	1074

som klassificerer 6368 låntagere i BRF Kredit efter alder og tvangsauktion/ikke–tvangsauktion (opgave 5.5.1, side 38). En naturlig model går ud på at betragte antal personer i hver af de fire aldersgrupper som givne, og fortolke antal tvangsauktioner i hver gruppe som binomialfordelt. Hvis data findes på en fil `BRF.TXT` som indeholder de fire linier

```
139  985
240 1587
256 1982
105 1074
```

kan vi i ISU indlæse data i form af en variate `M` med antalsparametrene og en variate `Y` med de relative hyppigheder på følgende måde:

```
VARIATE y1 y2 4
OPENINFILE brf.txt
READ y1 y2
COMPUTE m=y1+y2
COMPUTE y=y1/m
```

For at få udført testet for om de fire sandsynlighedsparametre er ens kan vi prøve at fitte homogenitetsmodellen:

```
FITLOGITLINEAR y=1/m
```

Bemærk syntaksen: Antalsparametrene specificeres ved at modelspecifikationen på højre side "divideres" med den tilsvarende variate. Uden denne specifikation antages det at alle antalsparametre er 1, svarende til at vi har egentlige binære responser.

Denne kommando giver følgende output (som indeholder det kvotienttest der bliver spurgt om i opgave 5.5.1):

```
Convergence after 5 iterations.
4 observations, 1 parameters estimated.
-2log(Likelihood) =             13.7245
Likelihood ratio test against full model
P[ ChiSquare(3) > -2log(Likelihood) ] =  0.003305
```

Herefter kan vi, for illustrationens skyld, prøve at fitte den fulde logit–lineære model og udskrive parameterestimaterne:

```
FACTOR aldgr 4 4
COMPUTE aldgr=#
FITLOGITLINEAR y=aldgr/m
LISTPARAMETERS
```

Output fra `FITLOGITLINEAR` ser i dette tilfælde sådan ud:

```
Convergence after 5 iterations.
4 observations, 4 parameters estimated.
-2log(Likelihood) =              -0.0000
```

Ikke særligt informativt, men det er jo også den fulde model vi har fittet. Output fra `LISTPARAMETERS` ser sådan ud:

	Estimate	Std.dev.	U	P
ALDGR[1]	-1.9582	0.09061	-21.612	0.000000
ALDGR[2]	-1.8890	0.06926	-27.274	0.000000
ALDGR[3]	-2.0467	0.06641	-30.817	0.000000
ALDGR[4]	-2.325	0.1022	-22.741	0.000000

Bemærk at (f.eks.)

$$-1.9582 = \log \frac{139}{985} = \operatorname{logit} \frac{139}{139 + 985}$$

De angivne estimater er altså logit–transformerede relative hyppigheder. Bortset fra logit–tranformationen svarer parametriseringen helt til hvad vi kender fra ensidet variansanalyse, når modellen specificeres uden konstantled. De tests der udføres er tests for $\operatorname{logit}(p_i) = 0$, svarende til $p_i = \frac{1}{2}$ — altså helt irrelevante. Vi kunne have tilføjet et konstantled for at få udført de (lidt mere relevante) parvise sammenligninger af p_1, p_2 og p_3 med p_4. De ville i så fald blive udført ved hjælp af et test, der approksimativt er det samme som det der er beskrevet på side 15–17. U–størrelserne udregnes altid som Estimate/Std.dev., og de angivne P–værdier er udregnet ved approksimation af U'ernes fordeling med en normeret normalfordeling. Men præcis hvordan parameterestimaternes standardafvigelser er estimerede vil vi ikke gå i detaljer med her.

Man kunne her (lidt kunstigt, måske) forsøge at fitte en model med logit–lineær effekt af alderen, ved i stedet for faktoren `ALDGR` at bruge en variate af længde 4 med (approksimative) intervalmidtpunkter som værdier:

```
VARIATE alder 4
COMPUTE alder=[25 35 45 55]
FITLOGITLINEAR y=1+alder/m
```

Output fra den sidste kommando

```
Convergence after 5 iterations.
4 observations, 2 parameters estimated.
-2log(Likelihood) =               4.5359
```

```
Likelihood ratio test against full model
P[ ChiSquare(2) > -2log(Likelihood) ] =  0.103522
```

viser at denne model godkendes ved testet mod den fulde model. Vi prøver med

```
LISTPARAMETERS
```

som giver

```
                  Estimate      Std.dev.          U          P
CONSTANT            -1.548        0.1616     -9.581   0.000000
   ALDER          -0.01201      0.003964     -3.030   0.002442
```

Konklusionen er altså, at sandsynligheden for tvangsauktion aftager med alderen efter formlen

$$P(\text{tv.auk.}) = \frac{\exp(-1.548 - 0.01201 \times \text{alder})}{1 + \exp(-1.548 - 0.01201 \times \text{alder})}.$$

I denne formel skal “alder” ganske vist fortolkes som “alderen afrundet til nærmeste intervalmidtpunkt”. Men det, at vi får godkendt modellen, giver et vist håb om at vi (i det oprindelige datasæt) ville kunne få godkendt en tilsvarende model med de rigtige aldre.

Ifølge princippet for successiv testning bør vi også prøve at foretage testet for homogenitet imod denne model:

```
FITLOGITLINEAR y=1/m
TESTMODELCHANGE
```

Output fra den sidste kommando bliver her

```
1 parameters removed
-2log(Q) =              9.1885
P[ ChiSquare(1) > -2log(Q) ] =  0.002435
```

Altså, som venteligt, en endnu skarpere afvisning af homogenitetshypotesen end vi fik ved test mod den fulde model.

EKSEMPEL 13.5.

I SAS’s eksempelsamling findes bl.a. følgende eksempel: 40 personer er blevet spurgt om de vil abonnere på en bestemt avis. For hver person foreligger køn, alder og svaret, kodet som 1 for ja, 0 for nej. Data foreligger på filen AVIS.TXT, som (udskrevet i fire spalter) ser sådan ud:

```
Female  35   1    Female  48   1      Male  50   0    Female  39   1
  Male  44   1    Female  56   0    Female  45   1      Male  34   1
  Male  45   0      Male  46   0    Female  47   1    Female  52   0
Female  47   0    Female  59   0    Female  30   0    Female  46   1
Female  51   1    Female  46   0    Female  39   1      Male  58   0
Female  47   1      Male  59   0    Female  51   1    Female  50   0
  Male  54   0      Male  38   0    Female  45   1    Female  32   1
  Male  47   0    Female  39   1    Female  43   0    Female  52   0
Female  35   1      Male  49   0      Male  39   0    Female  35   1
Female  34   1      Male  42   0      Male  31   1    Female  51   1
```

Vi indlæser data således:

```
FACTOR koen 40 2
VARIATE alder svar 40
OPENINFILE avis.txt
READ koen=,Male,Female alder svar
```

Som grundmodel tager vi den model, der beskriver den logit–transformerede sandsynlighed for at svare ja som en lineær funktion af alderen, hvor både hældning og afskæring i første omgang kan afhænge af respondentens køn:

```
FITLOGITLINEAR svar=1+koen+alder+koen*alder
LISTPARAMETERS
```

Output fra den første kommando vil vi ikke gengive her, men blot bemærke at det *ikke* indeholder testet for modellen mod den fulde model, da dette test viser sig at være meningsløst og ubrugeligt i tilfælde af rent binære data. Output fra LISTPARAMETERS ser sådan ud:

	Estimate	Std.dev.	U	P
CONSTANT	6.30	3.250	1.938	0.052573
KOEN[1]	5.21	7.800	0.667	0.504461
KOEN[2]	set to zero			
ALDER	-0.1246	0.06928	-1.798	0.072197
KOEN[1]*ALDER	-0.184	0.1907	-0.963	0.335758
KOEN[2]*ALDER	set to zero			

På grund af de sædvanlige parametriseringskonventioner gælder det her, at parameteren "KOEN[1]*ALDER", som er estimeret til -0.184, i virkeligheden skal fortolkes som differensen mellem de to hældninger. Testet for om denne parameter kan sættes lig med 0 kan defor fortolkes som et test for parallellitet af de to "regressionslinier". P–værdien 0.34 tyder på at dette kan godkendes. Men et mere pålideligt test er kvotienttestet, som vi kan foretage ved

```
FITLOGITLINEAR svar=1+koen+alder
TESTMODELCHANGE
```

Output fra den sidste kommando giver

```
1 parameters removed
-2log(Q) =              1.2066
P[ ChiSquare(1) > -2log(Q) ] =  0.271998
```

som bekræfter at hypotesen om parallelitet kan godkendes. Hvis vi herefter udskriver parameterestimaterne i den reducerede model ved

```
LISTPARAMETERS
```

får vi

	Estimate	Std.dev.	U	P
CONSTANT	8.18	3.096	2.643	0.008212
KOEN[1]	-2.422	0.9559	-2.534	0.011270
KOEN[2]	set to zero			
ALDER	-0.1649	0.06519	-2.530	0.011416

Her er både forskellen mellem afskæringerne og den fældes hældning signifikant forskellige fra 0 (P = ca. 0.01 i begge tilfælde), så vi vil ikke forsøge at reducere yderligere. Konklusionen bliver altså, at sandsynligheden for at en person siger ja til at abonnere på den pågældende avis kan estimeres ved

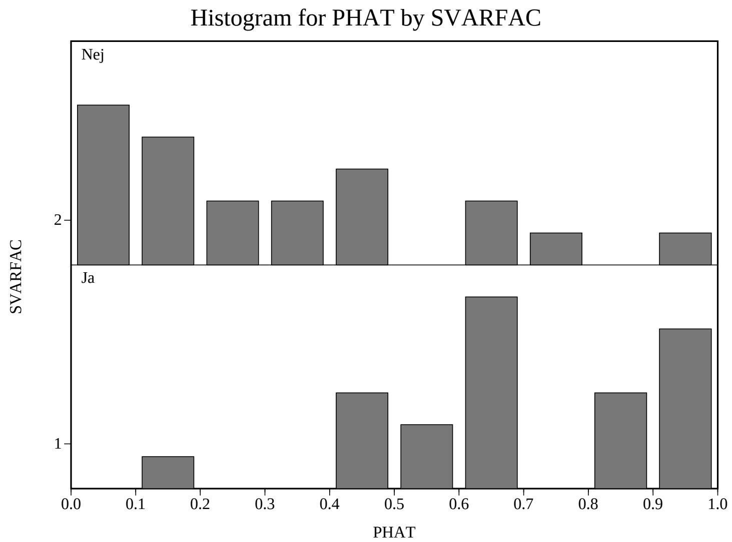
$$P(\text{Ja}) = \frac{\exp(8.18 - 2.422 \times 1_{\text{Mand}} - 0.1649 \times \text{alder})}{1 + \exp(8.18 - 2.422 \times 1_{\text{Mand}} - 0.1649 \times \text{alder})}$$

hvor leddet $2.422 \times 1_{\text{Mand}}$ naturligvis skal forstås sådan at 2.422 skal subtraheres for mænd, ikke for kvinder.

Mulighederne for at foretage modelkontrol i logistiske regressionsmodeller for rent binære data er begrænsede. Men man kan få et indtryk af, hvor godt modellen er i stand til at skille Ja–svarerne fra Nej–svarerne ved at tegne to "parallelle" histogrammer, der viser de estimerede Ja–sandsynligheders fordelinger for henholdsvis Ja–svarere og Nej–svarere:

```
FACTOR svarfac 40 2
COMPUTE svarfac=2-svar
SAVEFITTED phat
HISTOGRAM phat=0.0,10,1 svarfac=,Ja,Nej
```

Den sidste kommando producerer følgende tegning:

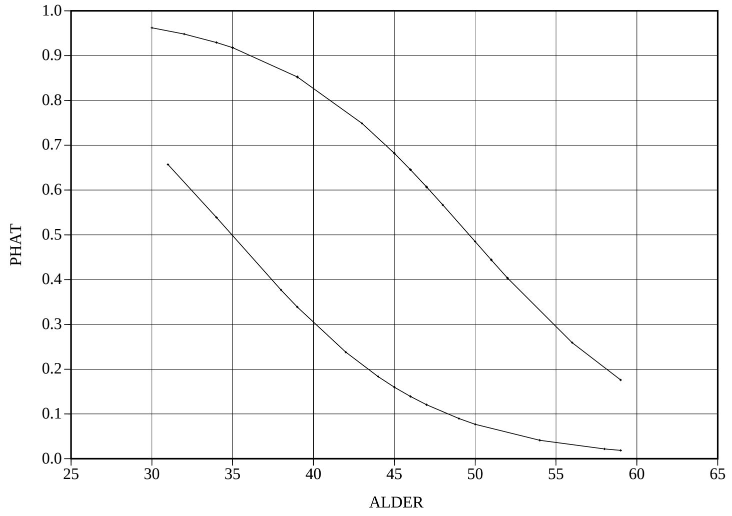


`HISTOGRAM phat=0.0,10,1 svarfac=,Ja,Nej`

En simpel illustration af hvordan sandsynligheden for "Ja" afhænger af køn og alder får vi ved

```
SORT koen alder +
PLOT alder=25,-8,65 phat=0.0,-10,1 koen=,1,4 koen=L
```

Her er den øverste kurve (som er rød på originalen) for kvinder, den nederste (blå på originalen) for mænd.




```
PLOT alder=25,-8,65 phat=0.0,-10,1 koen=,1,4 koen=L
```

Dette (meget lille, og muligvis konstruerede) eksempel illustrerer en vigtig anvendelse af logistisk regression indenfor markedsanalyse. Hvis man ønsker at sælge en vare ved personlig (for eksempel telefonisk) henvendelse, er det naturligvis vigtigt at vide hvem det betaler sig at henvende sig til. Konklusionen i dette tilfælde er, at kvinder under 40 går til biddet i over 80% af tilfældene (hvilket forekommer en anelse urealistisk), medens det er næsten umuligt at få ældre mænd til at købe denne avis.