## Dagens program:

Mandag den 14 marts
**Eksempler på generaliserede lineære modeller**
Regressions- og faktormodeller, forskellige responsfordelinger

- Resume
- Poisson regression (brug af offset). Data nematod
- Eksempel på ikke-kanonisk link, bakterietælling, Data bakt
- Alternativer til logistisk link ved binær respons, Data Gnist
- Tofaktormodeller, vekselvirkning. Data uheld, Lombard
- Brug af SAS-procedurer
- Empiriske varianser fra normaltfordelte observationer, Data vask
- Gamma-regression, Data blod
- Evt Ordnet kategorisk respons, Data buscum, busmult
- Opsamling



## Generaliserede lineære modeller, Resume

I stedet for Center plus støj, bruger vi middelværdi - og fordelingsmodel for obs, karakteriseret ved **variansfunktion** $V(\mu)$

- I stedet for lineær model for middelværdier betragter vi en lineær model for en funktion $\eta = g(\mu)$ (linkfunktionen) af middelværdien $\mu$
- Fordelingsformen bestemmer en **variansfunktion** $V(\mu)$ og en **enhedsdeviansfunktion** $d(y;\mu)$
- Og så minimerer vi bare devianser i stedet for kvadratafvigelser
- og **opspalter** devianser ligesom Pythagoras ($\chi^2$ fordelte størrelser)
- og vi kan bruge **Error** leddet til modelkontrol

Sædvanligvis foretrækker vi Likelihood ratio konfidensintervaller (fra profilikelihood) frem for Wald intervaller, og deviansresidualer frem for Pearson.



## Indskud: Bestemmelse af p-værdier, tabelopslag

Man får ofte brug for den kumulerede $\chi^2$-fordeling.
I stedet for at slå op i en tabel kan man lave et lille program:

**I SAS**:

```
DATA regn;
x = 12.3;
df = 6 ; P = 1- PROBCHI(x,df);
OUTPUT;
PROC PRINT;
RUN;
```

Eller man kan gå på nettet:
feks: http://calculators.stat.ucla.edu/cdf
eller man kan bruge Excel



## Poisson regression (brug af offset)

side 182

DATA nematod angiver antal nematoder (ormeagtige parasitter) i forskellige prøvevolumener (10, 20 og 40 ml) af et medium.

Ved tilfældig spatiel fordeling af nematoderne i mediet vil man forvente at antallet er proportionalt med volumen.

**Grafisk vurdering:**
Scatterplot: Antal mod volumen.

**Statistisk model:**
$Y =$ antal, $Y_i \sim \mathrm{P}(\rho x_i)$ (Poisson), hvor $x_i$ angiver prøvevolumen.

## Analyse ved FIT-menu

**Direkte:**
ant $\longrightarrow$ Y
vol $\longrightarrow$ X
No Intercept
Method:
Response Dist: Poisson
Link: Identity



## Analyse ved FIT-menu

**Ved kanonisk link:** (log)

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

med $\beta = \ln(\rho)$.
Vi ønsker ikke at bestemme en koefficient til $\log(x_i)$. Derfor betragter vi modellen

$$\eta_i - \ln(x_i) = \beta$$

dvs en model med **offset** lig med $\ln(x_i)$ og blot et intercept led.
Proportionalitetsfaktoren $\rho = \exp(\beta)$



## Brug af offset ved Poisson regression

**Fit menu:**
ant $\longrightarrow$ Y
Intercept
Method:
Response Dist: Poisson
Link: Log
lvol $\longrightarrow$ Offset

Er undertiden nyttig, hvor man bruger kanonisk link til andre forklarende variable med multiplikativ effekt.



## Eksempler på uhensigtsmæssig adfærd:

Hvad går galt, hvis man bruger:
rate $\longrightarrow$ Y
Intercept
Response Dist: Poisson
????

Eller med ant $\longrightarrow$ Y
lvol $\longrightarrow$ X
Response Dist: Poisson
Link: log
????

## Eksempel på brug af ikke-kanonisk link

Bestemmelse af bakteriekoncentration i et medium, p. 171

Man kan kun registrere vækst/ ikke vækst

Man bruger et *dilution assay*, hvor man bestemmer vækst/ ikke vækst i en række forskellige fortyndinger.
Ofte flere prøver ved hver fortynding.

Data bakt:

| Dilution | | Number of plates | | |
|---|---|---|---|---|
| Index | Concentration | total | non-fertile | fraction |
| 1 | 1 | 5 | 1 | 0.2 |
| 2 | 0.5 | 5 | 3 | 0.6 |
| 3 | 0.25 | 5 | 2 | 0.4 |
| 4 | 0.125 | 5 | 5 | 1.0 |



## Bestemmelse af bakteriekoncentration

**Model**:
Antag bakterier tilfældigt spredt i mediet. (Poissonproces).

Grundlæggende ide:
Betragt en prøve med concentration $x$ [ml sample/ml suspension].

$$p(x) = \text{P[ej vaekst]} = \text{P[0 bacterier]} = \exp(-\text{mx}) \ .$$

Det følger da, at

$$\log(p(x)) = -mx$$

er proportional med concentrationen $x$ of mediet i prøven



## Bestemmelse af bakteriekoncentration

**Grafisk vurdering**:
Tegn scatterplot af log andel mod concentration

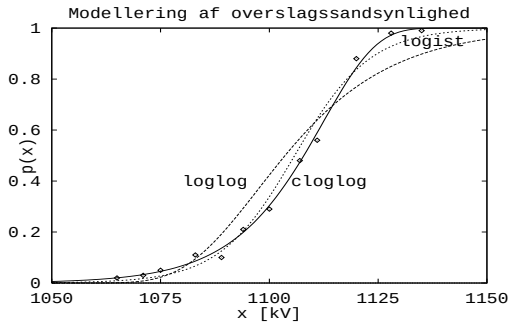
**Modelfit**:
ejvaekst → Y
conc → X
Ingen Intercept
**Method**:
Response = Poisson
Link = log

Estimatet kaldes **Most Probable Number**.
Metoden klassisk, estimationen udledt af Fisher



## Alternativer til logistisk link ved binær respons

Eks. 4.14.1 side 168 ff. Linkfunktioner side 202



## Fortolkning af dosis-respons relation

Ved dosis-respons forsøg beskriver dosis-respons funktionen fordelingen af **tolerancer** i populationen.

Tidligere brugte man ofte en **probit**- model (normalfordeling).
Nok mest fordi den var tabelleret. Ikke væsensforskellig fra logitmodellen.



## Tofaktormodeller, vekselvirkning

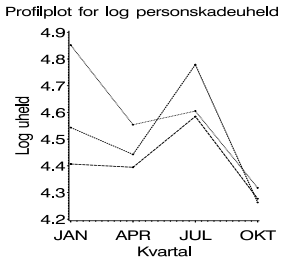
Eksempel 4.15.1 side 191 ff.
DATA uheld
Kvartalsvise uheldstal
**Model:**
$Y_{ij} \sim \mathrm{P}(\mu_{ij})$ (Poisson)

**Hypotese**:
proportionalitet, dvs $\mu_{ij} = \rho_i \theta_j$
For kanonisk link $\eta_{ij} = \alpha_i + \beta_j$

Vurderes grafisk ved **profilplot**



## Profilplot




Tilnærmelsesvis parallelitet - Vi kan vurdere paralleliteten ved et test.



## Tofaktormodeller, vekselvirkning

**Modelfit**:
uheld $\rightarrow$ Y
aar $\rightarrow$ X
kvt $\rightarrow$ X
Intercept
Method:
Response = Poisson
Link = log

dvs. model uden vekselvirkning.

Goodness of fit teststørrelse er et test for multiplikativ Poissonfordeling, dvs uden vekselvirkning

Modellen uden vekselvirkning giver simple fortolkninger:
Forskel på kvartaler afhænger ikke af året.

## Logit link ikke altid naturligt for binært respons

Eksempel 4.15.2 p. 196 ff., Datasæt Lombard
Antal personer med ringe viden om cancer:

| | læsning | |
|---|---|---|
| | Nej | Ja |
| Avislæsn. | | |
| Nej | 393/477 | 83/150 |
| Ja | 156/231 | 177/378 |



## Ringe viden om cancer (fortsat)

Naturligt at tænke på et serielt system:

Ringe viden: [ej udbytte af avis] $\cap$ [ej udbytte af læsn]
Viden: [udbytte af avis] $\cup$ [udbytte af læsn]

Multiplikativ model for ssh for ringe viden

$$\ln(p_{i,j}) = \kappa + \alpha_i + \beta_j$$

dvs binomial model uden vekselvirkning for Link=log og respons = ukendsk.
ukendsk $\rightarrow$ Y  AVIS $\rightarrow$ X  LAES $\rightarrow$ X
Response = Binomial  Pers $\rightarrow$ Binomial  Link = log

Bemærk: Modellen har ikke symmetri mellem ukendsk og kendsk

Model uden vekselvirkning for Link=logit ville svare til proportionalitet af **odds**, og det er noget andet.



## Brug af SAS-procedurer, PROC GENMOD og PROC LOGISTIC:

Man kan også skrive SAS-programmer med SAS-procedurer.
PROC GENMOD er dedikeret til generaliserede lineære modeller
PROC LOGISTIC er dedikeret til "logistiske regressionsproblemer" (også med kategoriske forklarende variable)

Information om procedurerne fås i SAS-hjælp:
Help $\rightarrow$ SAS System Help $\rightarrow$ SAS/Stat Software (i venstre side) $\rightarrow$ procedurenavn (i højre side).

Udførligere dokumentation (Online documentation) http://v8doc.sas.com/sashtml/

Procedurer bygget op af sætninger. Hver sætning afsluttes med semikolon.



## PROC GENMOD og PROC LOGISTIC:

Procedurerne skrives i **Program**-vinduet, og submittes ved klik på submit.

Strukturen i MODEL statement minder om FIT-menuen i Insight;
Options skrives i sætningen efter en /

Procedurerne giver mulighed for lidt flere variationer i output, end SAS/Insight, fx konfidensgrænser for lineæ prædiktor, fittede værdier mv.

Giver printet output, og mulighed for output til datasæt. Og muligheder for lidt flere responsfordelinger, f.eks. Negativ Binomial og Multinomial.

## Empiriske varianser fra normaltfordelte observationer

Side 183 ff.
$X_1, X_2, \ldots, X_n$ uafhængige med $X_i \sim \mathsf{N}(\mu, \sigma^2)$. Betragt den empiriske varians,

$$S^2 = \frac{\sum_{i=1}^n (X_i - \overline{X}_.)^2}{n-1}$$

med $\overline{X} = \sum_{i=1}^n X_i/n$.

$$S^2 \sim \sigma^2 \chi^2(f)/f$$

med $f = n - 1$.

En gammafordeling: $S^2 \sim \mathsf{G}\big(f/2, \sigma^2/(f/2)\big)$ med

$$\mathsf{E}[\mathsf{S}^2] = \sigma^2 \ ,$$

variansfunktionen $V_G(\sigma^2) = (\sigma^2)^2$ og præcisionsparameter $w = f/2$.
Kanonisk link er den **reciprokke**, $\eta = 1/\sigma^2$.

Lidt forvirrende at symbolet $\sigma^2$ nu optræder som **middelværdi** (i fordelingen af $S^2$).



## Empiriske varianser fra normaltfordelte observationer

**EKSEMPEL**
Datasæt vask indeholder resultater fra et vaskeforsøg med fire forskellige enzymer, ENZYM = A,B,C og D, der blev afprøvet på tre forskellige positioner, POS = 1,2,3 i en vaskemaskine.

Den variable VAR angiver den empiriske varians for hver kombination af de to faktorer.

Der indgik forskellige antal vaske for de forskellige faktorkombinationer som indikeret ved den variable DF, der angiver frihedsgraderne knyttet til SAK.

Vi vil vurdere om der kan antages at være en multiplikativ effekt af de to faktorer



## Empiriske varianser fra normaltfordelte observationer

**Fit menu i SAS Insight**:
Den variable WGT i datasættet er netop DF/2.

VAR $\longrightarrow$ Y
POS $\longrightarrow$ X
ENZYM $\longrightarrow$ X
WGT $\longrightarrow$ Weight
Distribution = Gamma
Link = log
Scale = Constant (konstanten skal være 1)

Vægten $f/2$ bliver tilgodeset i alle beregnede størrelser, devianser, residualer mv.



## Empiriske varianser ved PROC GENMOD

I PROC GENMOD skal man huske at deklarere variable af typen Nominal

```
PROC GENMOD DATA=vask;
CLASS pos enzym ;
MODEL var = pos enzym / TYPE1 TYPE3 DIST=gamma NOSCALE LINK=log
OBSTATS ;
WEIGHT WGT ;
RUN;
QUIT;
```

Bemærk at man skal skrive NOSCALE, ellers begynder den at estimere skalaparameter.

Også her indgår vægten i alle relevante størrelser

## To-parametret Gamma fordeling som responsfordeling

Gammafordelingen er nævnt i oversigten side 123:
$Y \sim \mathsf{G}(\alpha, \beta/\alpha)$

har tæthedsfunktion

$$g_Y(y;\alpha,\beta) = \frac{1}{\Gamma(\alpha)\beta/\alpha}\left(\frac{y}{\beta/\alpha}\right)^{\alpha-1}\exp(-y\alpha/\beta) \qquad \text{for } 0 < y$$

med E[Y] $= \beta$ og V[Y] $= \beta^2/\alpha$, dvs variansfunktion $V(\mu) = \mu^2$ og **præcisionsparameter** $\lambda = \alpha$.

I SAS-notation , $\phi = 1/\alpha$ og SCALE $= \alpha$.



## Eksempel på Gamma regression

Datasæt bloddata er resultater af et forsøg, hvor man bestemte den gennemsnitlige størkningstid for blod (i sek) ved tilsaetning af prothromboplastin til forskellige koncentrationer af prothrombofri plasma fra to forskellige plasmabatches.

**Grafisk vurdering:**
Datasæt blodplot bruges til den indledende grafiske analyse.
Scatterplot af tid mod koncentration, ser hyperbolsk ud
Prøv med tid mod reciprok koncentration
og reciprok tid mod log-koncentration.

Vi vælger en lineær model for reciprok tid af log-koncentration
dvs Link = reciprok, $\eta(\mu) = 1/\mu$ (kanonisk for Gamma) og $\eta_i = \alpha + \beta x_i$ hvor $x_i$ angiver log koncentration.



## Gamma regression i Insight

**Fit menu:**
tid $\longrightarrow$ Y
logkonc $\longrightarrow$ X
tilsaet $\longrightarrow$ X
Method:
Response Dist: Gamma
Link: Power (Potensen vælges til -1)
Scale: Maximum likelihood
Exact Distribution

Vi ser, at den estimerer scale (parameteren $\alpha$). Men den **skalerer ikke devianser mv**, så vi kan ikke stole på goodness of fit teststørrelsen. (Den skal multipliceres med værdien af Scale).



## Gamma regression ved PROC GENMOD

```
PROC GENMOD data = bloddata;
CLASS tilsaet;
MODEL tid = logkonc tilsaet /DIST = GAMMA
TYPE1 TYPE3 OBSTATS ;
RUN; QUIT;
```

Bemærk at vi ikke siger noget om SCALE. For Gammafordelingen vælger den pr default maximum-likelihood estimation af scale og skalering af relevante størrelser.

Vi får udskrevet både deviansen og den skalerede devians.

Skaleringen influerer ikke på estimationen af middelværdiparametrene. Skaleringen betyder bare en fælles faktor på alle deviansbidrag.

(Vægtning tillægger derimod forskellig vægt til de enkelte bidrag).

## Ordnet kategorisk respons

Tilfredshed hos buspassagerer ved forskellige forsinkelser, side 175 ff.
5 svarkategorier, ssh $p_1, p_2, p_3, p_4, p_5$
Men $p_1 + p_2 + p_3 + p_4 + p_5 = 1$, så egentlig kun fire sandsynligheder.

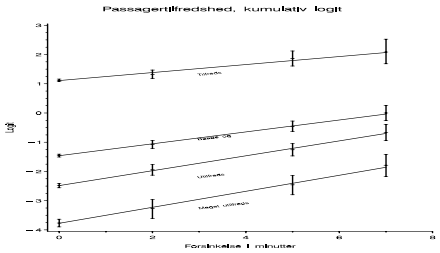
Mange odds-muligheder.
Vi vælger **kumulative odds** svarende til kumulering af svarkategorier (fra neden), $\Pi_1 = p_1$, $\Pi_2 = P_1 + p_2$ etc.

**Analyse ved Fit-menu i Insight:**
Datasæt buscum indeholder de kumulerede svar i de fire variable mutil, utilcum, bogcum og tilfcum, samt den variable ialt.
Man kan lave logistisk regression for hver af de enkelte kumulative kategorier
Eksempel:

ukendsk → Y  fors → X
Response = Binomial  ialt → Binomial  Link = logit



## Tilfredshed hos buspassagerer, kumulativ logit




Individuelle hældninger.
Markeringerne angiver de logittransformerede konfidensintervaller fra de observerede andele



## Ordnet kategorisk respons i PROC GENMOD og PROC LOGISTIC

Datasæt busmult indeholder data fra tilfredshedsundersøgelsen, hvor responserne er samlet i den variable resp, der antager værdierne '0mutil', '1util', '2bog', etc. Antallet respondenter med den pågældende respons er angivet i den variable ant.

Datasættet kan behandles såvel af PROC GENMOD som af PROC LOGISTIC. Begge procedurer benytter kumulativ logit som standardoption, og begge procedurer tilpasser en model med **proportionale odds**, dvs parallelle logit'er.



## Ordnet kategorisk respons i PROC GENMOD

I PROC GENMOD er det nødvendigt at specificere at fordelingen er en multinomialfordeling i MODEL sætningen som vist nedenfor.

```
PROC GENMOD data=busmult;
MODEL resp = fors /DIST=mult ;
FREQ ant;
OUTPUT OUT = busud pred = prob XBETA = eta;
RUN; QUIT;
```

## Ordnet kategorisk respons i PROC LOGISTIC

**PROC LOGISTIC** er indrettet mod kategorisk respons, så i dette tilfælde med flere responskategorier ved den godt, at det drejer sig om en multinomialfordeling.

```
PROC LOGISTIC data=busmult;
FREQ ant;       MODEL resp = fors / LACKFIT CLODDS=PL ;
RUN; QUIT;
```

Bemærk, at proceduren udfører et test for "parallellitet" (proportional odds). Nøgleordet LACKFIT bevirker desuden at der foretages et goodness of fit test



## Opsamling:

- diverse GENLM for kombinationer af kanonisk link og responsfordeling
- brug af offset værdi
- Vi kan tilmed klare ordnet kategorisk respons og empiriske varianser
- Kanonisk link ofte "naturligt" - men ikke altid
- For Gammafordeling kan vi også estimere "formparameter" ved ML