Introduksjon til Generaliserte Lineære Modeller (GLM)

STK3100 - 23. august 2010

Sven Ove Samuelsen/Anders Rygh Swensen

Plan for første forelesning:

1. Introduksjon, Litteratur, Program
2. Eksempler
3. Uformell definisjon av GLM
4. Noen utvidelser av GLM
5. Plan for kurset

Introduksjon

Generaliserte lineære modeller (med utvidelser) omhandler sentrale klasser av mer kompliserte, men likevel standard modeller utover multippel regresjon / anova.

Spesielt skal vi se på hvordan binære data, telledata, kategoriske (multinomiske) data og levetidsdata kan analyseres innen rammen av regresjon.

Målet med emnet er både å lære å benytte disse modellene til konkrete analyser og kjenne den matematiske bagrunnen for analysene.

Emnet skal altså ha både et praktisk og et teoretisk perspektiv.

Lærebok (litteratur)

Som lærebok skal vi bruke ""Generalized Linear Models for Insurance Data" av Piet de Jong og Gillian Z. Heller."
Denne boka kan kjøpes i Akademika.

Boken har egen hjemmeside:

`http://www.actuary.mq.edu.au/research/books/GLMsforInsuranceData`

Her finnes blant annet de fleste data settene som benyttes.

Kurset vil som tidligere år vise eksempler fra mange fagfelt: medisin / biologi, samfunnvitenskap / økonomi, teknikk (?). Men vi får nå styrket eksemplene fra forsikring.

Mulig støttelitteratur er "Julian J. Faraway: Extending the linear model with R. Generalized linear, mixed effect and nonparametric regression models. Chapman & Hall/CRC 2006"
Denne kan lånes på biblioteket

Statistikk-program

Vi skal bruke programpakken R som kjører under de vanlige operativsystemer og som kan lastes ned gratis fra `http://mirrors.sunsite.dk/cran/`

I hovedsak skal vi benytte rutiner som er implementert i R. Det vil ikke bli behov for å programmere mye på egenhånd.

En kort introduksjon til R finnes på hjemmesiden til STK1120 for V08, se `http://www.math.uio.no/avdc/kurs/STK1120/R.html` og

`http://www.math.uio.no/avdc/kurs/STK1120/V08/introR.pdf`

En god innføring i R er boka til Peter Dalgaard: Introductory Statistics with R,2nd ed., 2008, Springer

Dataeksempel 1: Fødselsvekt og svangerskapslengde

	Gutter		Jenter	
	Varighet(uker)	Fødselsvekt (gram)	Varighet (uker)	Fødselsvekt (gram)
	40	2968	40	3317
	38	2795	36	2729
	40	3163	40	2935
	35	2925	38	2754
	36	2625	42	3210
	37	2847	39	2817
	41	3292	40	3126
	40	3473	37	2539
	37	2628	36	2412
	38	3176	38	2991
	40	3421	39	2875
	38	2975	40	3231
Gj.sn.	38.33	3024.00	38.75	2911.33

En er interessert i å studere veksthastigheten pr. uke i slutten av svangerskapet, og om denne er forskjelig for de to kjønn.

Dataeksempel 2: Dødelig giftdose for biller

Ca. 60 biller ble utsatt for hver av 8 ulike konsentrasjoner av CS_2, og antallet som døde ved hver av konsentrasjonene ble registrert.

Dose ($\log_{10} CS_2 mg\, l^{-1}$)	Antall biller	Antall døde
1.6907	59	6
1.7242	60	13
1.7552	62	18
1.7842	56	28
1.8113	63	52
1.8369	59	53
1.8610	62	61
1.8839	60	60

Ønsker å studere sammenhengen mellom dose og dødelighet.

Dataeksempel 3: Antall barn blant gravide

de Jong & Heller, side 15-16: Data over antall tidligere barn blant 141 gravide kvinner i ulike aldre.

Ikke uventet synes antall barn å øke med alder, se figur 1.11 i deJ&H

Dataeksempel 3b: Antall krav fra tredjepart

de Jong & Heller, side 17: Data over antall krav i 176 geografiske områder i New South Wales i en 12-måneders periode.

Forklaringsvariable:

- Statistisk kategori, 13 kategorier
- Antall uhell i området
- Antall drepte og skadede
- Befolkningsstørrelse

I begge eksempler: Antall, dvs. *tellevariable*, kanskje Poissonfordeling.

Typisk modell for Eks 1: Lineær regresjon

For $k = 1, ..., 12$ og $j = 1, 2$ (der $j = 1$ angir gutt og $j = 2$ jente)

$$Y_{jk} = \text{ fødselsvekt for baby nr. } k \text{ kjønn nr. } j$$

$$x_{jk} = \text{ svangerskapsvarighet for baby nr. } k \text{ kjønn nr. } j$$

antas

$$Y_{jk} = \alpha_j + \beta x_{jk} + \varepsilon_{jk}$$

der $\varepsilon_{jk} \sim \text{N}(0, \sigma^2)$, dvs. normalfordelte med forventning 0 og samme varians σ^2 og dessuten uavhengige.

Regresjonsparametre:

$$\beta = \text{ stigningskoeffisient}$$

$$\alpha_j = \text{ konstantledd for kjønn } j$$

Modellspesifikasjonen for Eks 1 kan alternativt skrives:

- Linearitet: $\mathrm{E}[Y_{jk}] = \mu_{jk} = \alpha_j + \beta x_{jk}$
- Konstant varians: $\mathrm{Var}[Y_{jk}] = \sigma^2$
- Uavhengige responser: Y_{jk}-ene uavhengige
- Normalitetsantagelse: $Y_{jk} \sim N(\mu_{jk}, \sigma^2)$

I STK3100 ser vi på utvidelser av lineære regresjonsmodeller til

- Linearitet etter transformasjon via "link-funksjon" $g()$: $g(\mu_{jk}) = \alpha_j + \beta x_{jk}$
- Andre fordelinger for responsene: Binomiske, Poisson, Gamma, ...
- Variansen avhenger av forventningen til responsene

I Eks. 2: Dødelighet for biller

er det rimelig å anta at Y_i = antall døde biller med dose x_i er binomisk fordelt

$$Y_i \sim \text{bin}(n_i, \pi_i)$$

der π_i = sannsynligheten for at en bille dør med dose x_i og n_i = antall biller som får dose x_i

En lineær modell for π_i tilpasset med vanlig minste kvadrater er problematisk fordi

- $0 \leq \pi_i \leq 1$ i motsetning til lineært utrykk $\alpha + \beta x_i$
- $\text{Var}(Y_i) = n_i \pi_i (1 - \pi_i)$ Ikke-konstant (heteroskedastisk) variansstruktur

Vanlig løsning for Eks. 2: Logistisk regresjon

Logistisk regresjonsmodell:

$$\pi_i = \frac{\exp(\alpha + \beta x_i)}{1 + \exp(\alpha + \beta x_i)}$$

Da blir $0 \leq \pi_i \leq 1$

Tilpasser så den logistiske regresjonsmodellen med Maximum Likelihood (ML).

- Tar hensyn til binomiske responser (og ikke-konstant varians)
- Effisiente estimater (tilnærmet med "mye" data)

Logistisk regresjon for Eks. 2: Andel døde biller

MLE: $\hat{\alpha} = -60.72, \hat{\beta} = 34.27$

Predikerte sannsynligheter: $\hat{\pi} = \frac{\exp(\hat{\alpha}+\hat{\beta}x)}{1+\exp(\hat{\alpha}+\hat{\beta}x)}$

Estimering logistisk regresjon

Storvik: "Numerical optimization of likelihoods: Additional literature for STK1120" gir en Newton-Rahpson rutine i R for å tilpasse logistisk regresjon til disse dataene.

Heldigvis er dette allerede implementert R. Bruk kommando

```
glm(cbind(Dode,Ant-Dode)~Dose,family=binomial)
```

- `glm` = Generalisert Lineær Modell
- `family=binomial` angir at vi har binære eller binomiske data
- Ved binomiske data angir `cbind(Dode,Ant-Dode)` antall suksesser og antall ikke-suksesser

I eks. 3: Y_i = Antall barn tidligere for mor nr. i

kan det være rimelig å anta at Y_i er Poissonfordelt med forventning μ_i der μ_i avhenger av x_i = mors alder.

Tilsvarende Eks 2:

- Forventningene $\mu_i > 0$
- Variansen til Y_i er lik μ_i, dvs. ikke-konstant varians

Vanlig løsning: Poisson-regresjon

$Y_i \sim \text{Po}(\mu_i)$ der $\mu_i = \exp(\alpha + \beta x_i)$

Dette er også en generalisert lineær modell og kan tilpasses ved glm-rutinen.

Må bare spesifisere at data er antatt Poissonfordelte ved `family=poisson`

Poisson-regresjon for eks. 3

MLE for (α, β) ble $(\hat{\alpha}, \hat{\beta}) = (-4.0895, 0.1129)$

Får dermed tilpassede forventninger $\hat{\mu}_i = \exp(\hat{\alpha} + \hat{\beta} x_i)$

Definisjon av GLM

Uavhengige responser: $Y_1, Y_2, \ldots, Y_n$
Vektorer av forklaringsvariable $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$
der $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ip})$ er p-dimensjonale.

En GLM = Generalisert Lineær Modell er definert ved

- $Y_1, Y_2, \ldots, Y_n$ kommer fra samme eksponensiell klasse (Eksponensielle klasser defineres senere, nok å vite at normalfordelinger, binomiske, Poisson-, gammafordelinger etc. utgjør eksp. klasser)
- Lineære komponenter (prediktorer)
  $\eta_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}$
- Linkfunksjon $g()$: Med $\mu_i = \mathrm{E}[Y_i]$ kobles forventningen til lineær komponent ved at $g(\mu_i) = \eta_i$

Lineær regresjon er en GLM

- Responser (Y_i-er) fra normalfordelinger
- Lineær komponent $\eta_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}$
- $\mathrm{E}[Y_i] = \mu_i = \eta_i$, dvs. linkfunksjonen $g(\mu_i) = \mu_i$ er identitetsfunksjonen

Spesielt gjør R-kommandoene `lm` for lineær regresjon og `glm` essensielt det samme bare med litt forskjellig utskrift.

Lineær regresjon er spesielt default-spesifikasjonen av for `glm`

Eks. 1: Fødselsvekter

```
> lm(vekt~sex+svlengde)

Call:
lm(formula = vekt ~ sex + svlengde)

Coefficients:
(Intercept)          sex     svlengde
    -1447.2       -163.0        120.9

> glm(vekt~sex+svlengde)

Call:  glm(formula = vekt ~ sex + svlengde)

Coefficients:
(Intercept)          sex     svlengde
    -1447.2       -163.0        120.9

Degrees of Freedom: 23 Total (i.e. Null);  21 Residual
Null Deviance:      1830000
Residual Deviance: 658800       AIC: 321.4
```

Logistisk regresjon er en GLM

- Responser (Y_i-er) fra binomiske fordelinger bin(n_i, π_i)
- Lineær komponent $\eta_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}$
- $\mathrm{E}[Y_i]/n_i = \pi_i = \frac{\exp(\eta_i)}{1+\exp(\eta_i)}$.
  Dermed fås linkfunksjon $g(\pi_i) = \log(\frac{\pi_i}{1-\pi_i})$

Kaller $g(\pi) = \log(\frac{\pi}{1-\pi}) = \mathrm{logit}(\pi)$ for logit-funksjonen.

```
> glm(cbind(Dode,Ant-Dode)~Dose,family=binomial)

Call:  glm(formula = cbind(Dode, Ant - Dode) ~ Dose, family = binomial)
Coefficients:
(Intercept)         Dose
     -60.72        34.27

Degrees of Freedom: 7 Total (i.e. Null);  6 Residual
Null Deviance:      284.2
Residual Deviance: 11.23        AIC: 41.43
```

Poisson-regresjon er en GLM

- Responser $Y_i \sim \text{Po}(\mu_i)$
- Lineær komponent $\eta_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}$
- $\text{E}[Y_i] = \mu_i = \exp(\eta_i)$, dvs. linkfunksjonen $g(\mu_i) = \log(\mu_i)$ er (den naturlige) logaritmefunksjonen

```
> glm(children~age,family=poisson)

Call:  glm(formula = children ~ age, family = poisson)

Coefficients:
(Intercept)          age
    -4.0895       0.1129

Degrees of Freedom: 140 Total (i.e. Null);  139 Residual
Null Deviance:      194.4
Residual Deviance: 165  AIC: 290
```

Noen utvidelser

Andre GLM-er:

- Telledata med negativ binomisk fordeling: Overspredning
- Kontinuerlige, ikke-normale responser: Gammafordeling, Invers gaussisk fordeling

Utvidelser av GLM:

- Multinomiske responser (ordinale og nominelle)
- Levetidsdata
- Analyse av avhengige data
- Generaliserte additive modeller (GAM)

Vi skal gå inn på multinomiske responser og levetidsdata.

Oversikt boka til de Jong & Heller

- Kap. 1: Introduksjon, Dataeksempler, Gjennomgåes ikke detaljert
- Kap. 2: Diverse fordelinger (med noen unntak kjent fra før)
- Kap. 3: Eksponensielle klasser, ML-estimering
- Kap. 4: Lineær modellering (stort sett kjent fra STK1110/STK1120)
- Kap. 5: Generaliserte lineære modeller
- Kap. 6: Telledata (Poissonregresjon, overspredning)
- Kap. 7: Katergoriske responser (binomiske data, multinomiske data)
- Kap. 8: Kontinuerlige responser
- Kap. 9: Korrelerte data

Plan for kurset

Vi følger boka til de Jong & Heller, men ikke til punkt og prikke, og heller ikke helt kronologisk. Dessuten må boka må fylles ut på en del punkter.

Omtrentlig forelesningsplan:

- Introduksjon, idag!
- Kap. 4. Lineaere modeller, stort sett repetisjon fra STK1110/STK2120, mandag 30. august
- Kap. 3: Eksponensielle klasser, 6. september
- Kap. 5: GLM-rammen og ML-teori 13. september.
- Kap. 7: Binomiske og Binære data
- Kap. 6: Telledata

Plan for kurset, forts.

- Kap. 7: Multinomiske data
- Kap. 8: Kort om kontinuerlige responser
- Levetidsanalyse (ikke i boka)
- Utvidelser: Korrelerte data og GAM