Forelesning 10 STK3100

3. november 2008

S. O. Samuelsen

Plan for forelesning:

1. Multinomisk fordeling
2. Multinomisk regresjon - ikke-ordnede kategorier
3. Multinomisk regresjon - ordnede kategorier

Multinomisk fordeling:

- n uavhengige forsøk
- Observerer ett av J ulike kjennetegn i hvert forsøk
- Sannsynligheten for kjennetegn j er lik π_j i hvert forsøk der altså $\pi_1 + \pi_2 + \cdots + \pi_J = 1$

Med Y_j = antall forsøk med kjennetegn j sier vi at $Y = (Y_1, Y_2, \ldots, Y_J)$ er multinomisk fordelt

$$Y \sim \text{Multinom}(n, \pi_1, \pi_2, \ldots, \pi_J)$$

Punktsannsynlighetene for Y gis da ved

$$f(y|n) = \frac{n!}{y_1!y_2!\ldots y_J!}\pi_1^{y_1}\pi_2^{y_2}\ldots\pi_J^{y_J}$$

for $y_j \in \{0, 1, \ldots, n\}$ og $\sum_{j=1}^J y_j = n$

Momenter i multinomisk fordeling

Spesielt er $Y_j \sim \text{Bin}(n, \pi_j)$ marginalt. Derfor får vi

$$\text{E}[Y_j] = n\pi_j \qquad \text{og} \qquad \text{Var}[Y_j] = n\pi_j(1-\pi_j)$$

Dessuten er, for $k \neq j$,

$$\text{Cov}[Y_j, Y_k] = -n\pi_j\pi_k$$

Dette følger ved å se på Y_{ij} = indikator for kjennetegn j i forsøk i. Da blir

$$\begin{aligned}\text{Cov}[Y_{ij}, Y_{ik}] &= \text{P}(Y_{ij}=1, Y_{ik}=1) - \text{P}(Y_{ij}=1)\text{P}(Y_{ik}=1)\\ &= 0 - \pi_j\pi_k\end{aligned}$$

og

$$\text{Cov}[Y_j, Y_k] = \text{Cov}[\sum_{i=1}^n Y_{ij}, \sum_{i=1}^n Y_{ik}] = \sum_{i=1}^n \text{Cov}[Y_{ij}, Y_{ik}] = -n\pi_j\pi_k$$

Sammenheng med Poissonfordeling

Anta at $Y_j \sim \text{Po}(\mu_j), j = 1, \cdots, J$ er uavhengige. Da er

$$(Y_1, \ldots, Y_J | \sum_{j=1}^J Y_j = n) \sim \text{Multinom}(n, \pi_1, \ldots, \pi_J)$$

der sannsynlighetene $\pi_j = \mu_j / \sum_{k=1}^J \mu_k$.

Dette følger av at $\sum_{j=1}^J Y_j \sim \text{Po}(\sum_{k=1}^J \mu_k)$. Dermed

$$\begin{aligned}\text{P}(Y_1 = y_1, \ldots, Y_J = y_J | \textstyle\sum_{j=1}^J Y_j = n) &= \frac{\text{P}(Y_1=y_1,\ldots,Y_J=y_J)}{\text{P}(\sum_{j=1}^J Y_j = n)}\\ &= \frac{\Pi_{j=1}^J \left[\frac{\mu_j^{y_j}}{y_j!}\exp(-\mu_j)\right]}{\frac{(\sum_{j=1}^J \mu_j)^n}{n!}\exp(-\sum_{j=1}^J \mu_j)}\\ &= \frac{n!}{y_1!y_2!\ldots y_J!}\frac{\mu_1^{y_1}\mu_2^{y_2}\ldots\mu_n^{y_n}}{(\sum_{j=1}^J \mu_j)^n}\\ &= \frac{n!}{y_1!y_2!\ldots y_J!}\pi_1^{y_1}\pi_2^{y_2}\ldots\pi_J^{y_J}\end{aligned}$$

"Multinomisk regresjon"

Eks. (Faraway, "Extending the linear model with R"):
Undersøkelse om hvem som stemte på Demokrater, Republikanerne eller Uavhengige ved valg i USA i 1996 etter utdannelse og inntekt.

```
> table(sPID)
sPID
   Democrat Independent  Republican
        380         239         325
> table(nes96$educ)
    MS HSdrop     HS   Coll  CCdeg  BAdeg  MAdeg
    13     52    248    187     90    227    127
> table(nes96$income)
  $3Kminus    $3K-$5K    $5K-$7K    $7K-$9K   $9K-$10K  $10K-$11K
        19         12         17         19         18         13
 $11K-$12K  $12K-$13K  $13K-$14K  $14K-$15K  $15K-$17K  $17K-$20K
        11         17         10         15         23         35
 $20K-$22K  $22K-$25K  $25K-$30K  $30K-$35K  $35K-$40K  $40K-$45K
        26         39         68         70         62         48
 $45K-$50K  $50K-$60K  $60K-$75K  $75K-$90K $90K-$105K  $105Kplus
        51        100        103         53         47         68
```


Eks. Faraway, forts.

- Respons: Partivalg
  - 3 kategorier: Demokrater, Republikanerne eller Uavhengige
  - Multinomiske data
  - Ikke opplagt ordning mellom kategoriene
- Forklaringsvariable: Utdannelse og inntekt
  - → Regresjon

Regresjon for multinomiske responser med ikke-ordnede kategorier:
Nominell multinomisk logit modell

Nominell multinomisk logit modell:

Med π_{ji} = sannsynlighet for responskategori j for individ med kovariat x_i:

$$\pi_{1i} = \frac{1}{1+\sum_{j=2}^{J}\exp(\beta_j' x_i)}$$
$$\pi_{ji} = \frac{\exp(\beta_j' x_i)}{1+\sum_{k=2}^{J}\exp(\beta_k' x_i)} \text{ for } j > 1$$

Merk:

- $\sum_{j=1}^{J}\pi_{ji} = 1$
- $\frac{\pi_{ji}}{\pi_{1i}} = \exp(\beta_j' x_i)$
- Ulike $\beta_j = (\beta_{j1}, \ldots, \beta_{jp})'$ for hvert nivå $j = 2, 3, \ldots, J$

Likelihood: Med responsvektor $Y^i = (Y_1^i, \ldots, Y_J^i)$

$$L = \prod_{i=1}^{n} \frac{n_i!}{Y_1^i! Y_2^i! \ldots Y_J^i!} \pi_{1i}^{Y_1^i} \pi_{2i}^{Y_2^i} \ldots \pi_{Ji}^{Y_J^i}$$

Nominell multinomisk logit modell: R-implementasjon

Optimering av likelihooden for nominelle multinomiske data kan gjøres i R med rutinen `multinom` som finnes i standard-R-biblioteket `nnet` (kort for "nevralt nett").

```
> library(nnet)
> mmod <- multinom(sPID ~ neduc + nincome, nes96)
> mmod
> multinom(sPID~neduc+nincome)
# weights:  12 (6 variable)
initial  value 1037.090001
iter  10 value 992.514868
final  value 992.514853
converged

Coefficients:
            (Intercept)        neduc    nincome
Independent   -1.161305 -0.003621094 0.01614688
Republican    -1.052716  0.028144037 0.01709310

Residual Deviance: 1985.030
AIC: 1997.030
```

Eks: Utdannelse spiller ingen rolle gitt inntekt

```
> multinom(sPID~1)$dev
# weights:  6 (2 variable)
initial  value 1037.090001
final  value 1020.636052
converged
[1] 2041.272
> multinom(sPID~neduc)$dev
[1] 2030.478
> multinom(sPID~nincome)$dev
[1] 1985.424
> multinom(sPID~nincome+neduc)$dev
[1] 1985.030
```

- Dev.endring 2041.27-2030.48 = 10.79 med bare utdannelse
- Dev.endring 2041.27-1985.42 = 55.75 med bare inntekt
- Dev.endring 1985.42-1985.03 = 0.39 med utdannelse i tillegg til inntekt

Eks: Standardfeil fra `summary.multinom`

```
mod<-multinom(sPID~nincome+neduc)
# weights:  12 (6 variable)
initial  value 1037.090001
iter  10 value 992.514868
final  value 992.514853
converged
> summary(mod,cor=F)
Call:
multinom(formula = sPID ~ nincome + neduc)

Coefficients:
            (Intercept)    nincome        neduc
Independent   -1.161305 0.01614688 -0.003621094
Republican    -1.052716 0.01709310  0.028144037

Std. Errors:
            (Intercept)     nincome      neduc
Independent   0.2561357 0.003081403 0.05712273
Republican    0.2380262 0.002860900 0.05257820

Residual Deviance: 1985.030
AIC: 1997.030
```


Eks: Må selv lage tabell med

estimater, standardfeil, t-verdier og p-verdier:

```
koeftab<-function(nnetmod){
koef<-summary(nnetmod)$coefficients
se<-summary(nnetmod)$standard.errors
koefb<-c(koef[1,],koef[2,])
seb<-c(se[1,],se[2,])
tval<-koefb/seb
pval<-2*pnorm(-abs(koefb/seb))
tab<-cbind(koefb,seb,tval,pval)
tab
}
> round(koeftab(mod),4)
              koefb    seb    tval   pval
(Intercept) -1.1613 0.2561 -4.5339 0.0000
nincome      0.0161 0.0031  5.2401 0.0000
neduc       -0.0036 0.0571 -0.0634 0.9495
(Intercept) -1.0527 0.2380 -4.4227 0.0000
nincome      0.0171 0.0029  5.9747 0.0000
neduc        0.0281 0.0526  0.5353 0.5925
```


Sammenheng vanlig og nominell multinomisk logit

Vi har at $Y_j^i | Y_1^i + Y_j^i = m_i \sim \text{Bin}(m_i, \pi_{ji}/(\pi_{1i} + \pi_{ji}))$ og

$$\frac{\pi_{ji}}{\pi_{1i}+\pi_{ji}} = \frac{\frac{\exp(\beta_j' x_i)}{1+\sum_{k=2}^J \exp(\beta_k' x_i)}}{\frac{1}{1+\sum_{j=2}^J \exp(\beta_j' x_i)} + \frac{\exp(\beta_j' x_i)}{1+\sum_{k=2}^J \exp(\beta_k' x_i)}} = \frac{\exp(\beta_j' x_i)}{1+\exp(\beta_j' x_i)},$$

dvs. vanlig logistisk regresjons-modell for $Y_j^i | Y_1^i + Y_j^i = m_i$.

- $\exp(\beta_{jl}) = \text{OR}_{jl}$ = oddsratio mellom nivå j og 1 for kovariat l
- Parametrene kan også estimeres ved separate logistiske regresjoner.

Eks: Separate logistiske regresjoner

```
> Rep<-1*(sPID=="Republican")
> Ind<-1*(sPID=="Independent")

> summary(glm(Ind~nincome+neduc,family=binomial,subset=Rep!=1))
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.106370   0.250014  -4.425 9.63e-06 ***
nincome      0.015536   0.003063   5.073 3.92e-07 ***
neduc       -0.009809   0.056135  -0.175    0.861

    Null deviance: 825.71  on 618  degrees of freedom
Residual deviance: 794.50  on 616  degrees of freedom
AIC: 800.5

> summary(glm(Rep~nincome+neduc,family=binomial,subset=Ind!=1))
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.052653   0.241909  -4.351 1.35e-05 ***
nincome      0.017754   0.002966   5.985 2.17e-09 ***
neduc        0.021639   0.054081   0.400    0.689

    Null deviance: 973.04  on 704  degrees of freedom
Residual deviance: 924.26  on 702  degrees of freedom
AIC: 930.26
```


Eks: Sammendrag separate logistiske regresjoner

- Estimater med multinomisk og separate logistisk regresjon er tilsvarende, men ikke identiske
- Standardfeil er også tilsvarende, men ikke identiske
- Multinomisk regresjon er en full likelihood analyse
- Kan være like oversiktlig å gjøre de separate logistisk regresjonene

Ordnede (ordinale) kategoriske responser

Eks: Fødselsvekt ble kategorisert i et eksempel til mindre (større) enn 2800 g. Kan også benytte flere terskelverdier, f.eks. 3500 g.

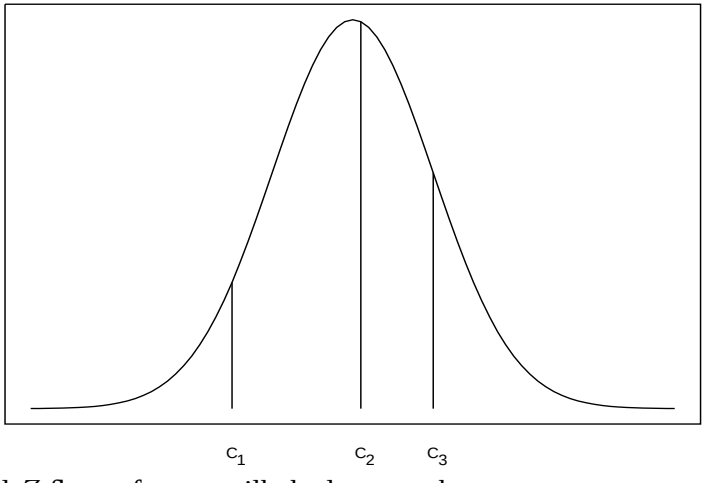
Eks: Muligens en skala fra Demokrater via Uavhengige til Republikanere

Generelt: en underliggende (typisk latent) skala Z og et sett av terskelverdier $C_1 < C_2 < \cdots < C_{J-1}$ slik at individet kategoriseres til

$$\begin{aligned} &\text{nivå } 1 \text{ dersom } Z \leq C_1 && \Leftrightarrow Y = 1 \\ &\text{nivå } j \text{ dersom } C_{j-1} < Z \leq C_j && \Leftrightarrow Y = j \\ &\text{nivå } J \text{ dersom } C_{J-1} < Z && \Leftrightarrow Y = J \end{aligned}$$

Merk: Y defineres på annen måte enn for nominelle data.

Underliggende skala og terskelverdier



Her vil Z flyttes fram og tilbake langs x-aksen ettersom kovariater varierer.

Modellering av ordinale multinomiske data

Siden det er mer struktur i ordnede kategoriske data kan man modellere med færre parametre.
Vi antar, at gitt kovariat x_i, er den latente variablen $Z_i = \beta' x_i + Z_{0i}$ der $\mathrm{P}(Z_{0i} \leq z) = F(z)$ for en passende kumulativ fordelingsfunksjon $F(z)$.
Da fås

$$\mathrm{P}(Y_i \leq j) = \mathrm{P}(Z_i \leq C_j) = \mathrm{P}(Z_{0i} \leq C_j - \beta' x_i) = F(C_j - \beta' x_i)$$

- Merk at det bare er en parametervektor β
- Dessuten er C_j-ene ukjente, dvs. parametre
- Vi får ulike modeller med ulike valg av (invers) "link"-funksjon $F(z)$

Proporsjonal oddsmodell: Logit-link

Anta at $F(z) = \exp(z)/(1+\exp(z))$. Da blir

$$\mathrm{P}(Y_i \leq j) = \exp(C_j - \beta' x_i)/(1+\exp(C_j - \beta' x_i))$$

og med $\pi_{ji} = \mathrm{P}(Y_i = j)$ (som ved nominelle responser) og $\gamma_{ji} = \mathrm{P}(Y_i \leq j) = \pi_{1i} + \cdots + \pi_{ji}$ fås

$$\frac{\gamma_{ji}}{1-\gamma_{ji}} = \exp(C_j - \beta' x_i)$$

og dermed blir odds-ratioen

$$\frac{\frac{\gamma_{ji'}}{1-\gamma_{ji'}}}{\frac{\gamma_{ji}}{1-\gamma_{ji}}} = \exp(\beta'(x_{i'} - x_i))$$

uavhengig av nivået C_j.

Eks: "Democrats" < "Independent" < "Republican"

Proporsjonale oddsmodeller tilasses i `R` med rutinen `polr` som finnes i `MASS`-biblioteket.
Spesielt skal responsen - den ordnede faktoren - være spesifisert som en faktor.

```
> library(MASS)
> summary(polr(as.factor(sPID)~nincome+neduc))

Coefficients:
            Value  Std. Error   t value
nincome 0.01261746 0.002126369 5.9338071
neduc   0.02565506 0.040922490 0.6269184

Intercepts:
                          Value  Std. Error t value
Democrat|Independent   0.3025 0.1867     1.6204
Independent|Republican 1.3853 0.1921     7.2108

Residual Deviance: 1994.970
AIC: 2002.970
```


Eks: Sammendrag

- Estimatene for `nincome` og `neduc` har lignende koeffisienter som ved nominell logit modell
- Estimatene for `nincome` og `neduc` har også tilsvarende grad av signifikans
- Residual Devians var 1985.03 for nominell logit og altså mindre enn 1994.97 proporsjonale odds

Utvidelser: Probit og Cloglog

Spesifikasjonen

$$\mathrm{P}(Y_i \leq j) = F(C_j - \beta' x_i)$$

tillater andre (inverse) linkfunksjoner

- Probit: $F(z) = \phi(z) = \int_{-\infty}^{z} \frac{\exp(-z^2/2)}{\sqrt{2\pi}} dz$
- Clog-log: $F(z) = 1 - \exp(-\exp(z))$
- Cauchit: $F(z) = \int_{-\infty}^{z} \frac{dz}{\pi(1+z^2)} = \frac{\arctan(z)}{\pi} + \frac{1}{2}$

Eks: Probit og Cloglog

```
> summary(polr(as.factor(sPID)~nincome+neduc,method="probit"))
Coefficients: Value  Std. Error   t value
nincome 0.007887085 0.001312480 6.0092976
neduc   0.014567640 0.025430970 0.5728307
Intercepts:            Value  Std. Error t value
Democrat|Independent   0.1812 0.1154     1.5704
Independent|Republican 0.8505 0.1173     7.2519

Residual Deviance: 1994.564
AIC: 2002.564

> summary(polr(as.factor(sPID)~nincome+neduc,method="cloglog"))
Coefficients: Value  Std. Error   t value
nincome 0.009374659 0.001431614 6.5483141
neduc   0.007236995 0.028933467 0.2501254
Intercepts:             Value  Std. Error t value
Democrat|Independent   0.5669 0.1303     4.3503
Independent|Republican 1.3610 0.1366     9.9613

Residual Deviance: 1989.348
AIC: 1997.348
```