Faculty of Life Sciences

Logistisk regression

Claus Ekstrøm
E-mail: ekstrom@life.ku.dk

Program

- Odds og odds-ratios igen
- Logistisk regression
- Estimation og inferens
- Modelkontrol

Eksempel: pesticider og møl

20 møl udsættes for forskellige doser af et pesticid, og det registreres, hvor mange der dør.

	Dosis					
	1	2	4	8	16	32
Døde	1	4	9	13	18	20

I alt 120 møl.

Odds og odds-ratios (igen)

Har hidtil snakket om differencen mellem successandsynligheder for to binomialfordelinger. Det svarer altså til en ændring i procentpoint.
Procentpoint kan dække over forholdsmæssige store og små ændringer.

Et andet mål kunne være forholdet.
Odds for en hændelse A er

$$\text{odds for } A = \frac{P(A)}{1 - P(A)}.$$

Politik: Dansk politik 15 april 2011 - 17:00

Dansk valg 2011 - Folketingsvalg

Næste Statsminister

Valg	Odds
Lars Løkke Rasmussen	1.90
Helle Thorning-Schmidt	2.00
Villy Søvndal	9.00
Lene Espersen	9.00
Pia Kjærsgaard	10.00

Odds-ratio

Odds-ratio er forholdet mellem to odds, fx. odds fra to grupper:

$$\theta = \frac{\frac{p_1}{1-p_1}}{\frac{p_2}{1-p_2}} = \frac{p_1 \cdot (1-p_2)}{p_2 \cdot (1-p_1)}$$

Her svarer en værdi på 1 til ingen forskel mellem de to odds. Et test for $H_0 : \theta = 1$ svarer altså et test for at odds i de to grupper er ens!
Den centrale grænseværdisætning giver, at $\log \hat{\theta}$ opfører sig "pænt", og at estimatet har en standard error på

$$\mathrm{SE}(\log \hat{\theta}) = \sqrt{\frac{1}{y_{11}} + \frac{1}{y_{12}} + \frac{1}{y_{21}} + \frac{1}{y_{22}}},$$

Kan altså bruge dette til at lave konfidensinterval for $\log \theta$.
Hvorfor: fortolkning anderledes: forhold vs forskel.
Parametrisering i mere komplicerede modeller.

Den logistiske regressionsmodel

Den logistiske regressionsmodel modellerer en binær responsvariabel som funktion af en eller flere kategoriske og/eller kontinuerte forklarende variable.
Den logistiske regressionsmodel er givet ved

$$Y_i \sim \mathrm{bin}(1, p_i), \quad i = 1, \dots, n,$$

hvor Y_i angiver om observation i var en succes ($Y_i = 1$) eller en fiasko ($Y_i = 0$).

Vi kan lade successandsynligheden for hver enkelt observation, p_i, afhænge af forskellige forklarende variable: kontinuerte og/eller kategoriske.

Den logistiske regressionsmodel II

Den logistiske regressionsmodel er givet ved

$$Y_i \sim \mathrm{bin}(1, p_i), \quad i = 1, \dots, n, \tag{1}$$

hvor Y_i angiver om observation i var en succes ($Y_i = 1$) eller en fiasko ($Y_i = 0$).

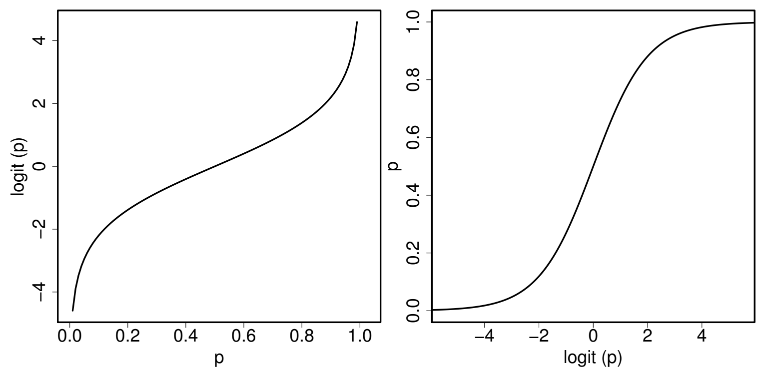
Log odds for hændelsen $Y = 1$ kaldes også logit af p_i og defineres som

$$\mathrm{logit}\ (p_i) = \log\left(\frac{p_i}{1-p_i}\right). \tag{2}$$

Vi kan modellere de individuelle successandsynligheder som

$$\mathrm{logit}\ (p_i) = \alpha + \beta_1 x_{i1} + \cdots + \beta_d x_{id}, \quad i = 1, \dots, n. \tag{3}$$

logit



$$p_i = \frac{\exp(\alpha + \beta_1 x_{i1} + \cdots + \beta_d x_{id})}{1 + \exp(\alpha + \beta_1 x_{i1} + \cdots + \beta_d x_{id})}. \tag{4}$$

Inferens i den logaritmiske regressionsmodel

Har allerede grundideerne i det, som vi skal bruge

Vi kan bruge tankegangen fra de tidligere uger til estimation (maximum likelihood), test af hypoteser (likelihood-ratio tests / Wald tests), konfidens- og prædiktionsintervaller og modelkontrol.

Ideerne er *helt* de samme — vi skal ikke komme ind på de konkrete metoder.

Eksempel: pesticider og møl

```
> dose <- c(1, 2, 4, 8, 16, 32)
> moths <- matrix(c(1,  4,  9, 13, 18, 20,
+                  19, 16, 11,  7,  2,  0), ncol=2)
> logreg <- glm(moths ~dose, family=binomial)
> summary(logreg)

Call:
glm(formula = moths ~ dose, family = binomial)

Deviance Residuals:
      1        2        3        4        5        6
-1.5729  -0.0954   1.1798   0.3631  -0.7789   0.1426

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.92771    0.40195  -4.796 1.62e-06 ***
dose         0.29723    0.06254   4.752 2.01e-06 ***
```


Inferens for logistisk regressionsmodeller

Konfidensintervaller for parametre i logistiske regressionsmodeller udregnes "som sædvanligt" (brug normalfordelingsfraktiler):

$$\text{estimat} \pm \text{fraktil} \cdot \text{SE}(\text{estimat}).$$

Test af hypoteser for en enkelt parameter kan foretages ved et Wald test

$$Z_{\text{obs}} = \frac{\text{estimat} - \text{sande værdi}}{\text{SE}(\text{estimat})}$$

og Z_{obs} er approksimativt normeret normalfordelt, $N(0,1)$.

Sammensatte hypoteser kan altid testes ved hjælp af et likelihood ratio test

$$\text{LR} = 2 \cdot (\log(L_{\text{full}}) - \log(L_0)), \tag{5}$$

Eksempel: pesticider og møl

```
> summary(logreg)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.92771    0.40195  -4.796 1.62e-06 ***
dose         0.29723    0.06254   4.752 2.01e-06 ***

> logreg2 <- glm(moths ~ 1, family=binomial)
> anova(logreg2, logreg, test="Chisq")
Analysis of Deviance Table

Model 1: moths ~ 1
Model 2: moths ~ dose
  Resid. Df Resid. Dev Df Deviance P(>|Chi|)
1         5     71.138
2         4      4.634  1   66.504 3.493e-16 ***
```

Modelkontrol

Som altid bør man checke sine modelantagelser

Pearson residualer defineres som

$$r_i = \frac{y_i - \hat{y}_i}{\sqrt{\hat{y}_i \cdot (1 - \hat{y}_i)}}, \tag{6}$$

så de svarer til de almindelige residualer delt med deres estimerede spredning af y_i.

Lav et standardiseret residualplot som sædvanligt. Kan se lidt "spøjst" ud.

Grafisk modelkontrol

```
> phi <- summary(model)$dispersion
> hi <- hatvalues(model)
> rstd <- residuals(model,
+         type="pearson")/sqrt(phi*(1-hi))
> plot(fitted(model), rstd,
+      xlab="Predicted", ylab="Std. Pearson")
```


Modelkontrol: Pearsons chi-square test

Alternativ til grafisk modelkontrol: sammenlign observerede og fittede andele.

Gruppér kontinuerte variable for at ende med grupperede data.

Pearsons chi-square goodness-of-fit test

$$X^2 = \sum_{j=1}^{J} \frac{(\text{obs}_j - \text{forv}_j)^2}{\text{forv}_j} = \sum_{j=1}^{J} \frac{(\text{obs}_j - n_j \hat{p}_j)^2}{n_j \hat{p}_j}, \tag{7}$$

Summér over alle mulige grupper J, for responsen og de forklarende variable

- n_j antal observationer i gruppe j.
- $\hat{p}_j$ estimeret relativ frekvens for gruppe j.

Teststørrelsen følger en χ^2-fordeling med $J/2 - r$ frihedsgrader (r parametre i modellen, og $J/2$ er det mulige antal grupper).

Modelkontrol: Pearson test

```
> obs <- 20
> expect  <- obs * fitted(logreg)
> expect2 <- obs * (1-fitted(logreg))
> sum((deaths-expect)**2/expect) +
+ sum((alive-expect2)**2/expect2)
[1] 4.247966
> 1-pchisq(4.247966, df=6-2)
[1] 0.3734862
```

Dagens hovedpunkter

- Logistisk regression
- Inferens for logistisk regression
- Modelkontrol