Forelesning 7 STK3100

6. oktober 2008

S. O. Samuelsen

Plan for forelesning:

1. Parameterfortolkning logistisk regresjon
2. Parameterfortolkning andre linkfunksjoner
3. Goodness-of-fit: Hosmer-Lemeshow-test

GLM Binomiske / binære responser

$Y_i \sim \text{Bin}(n_i, \pi_i)$ der linkfunskjonen $g(\pi_i) = \eta_i = \beta' x_i$ er invers av kontinuerlig kumulativ fordelingsfunksjon på $\mathbb{R}$.

Følgende linkfunksjoner er implementert i R:

- Logistisk regresjon: $g(\pi_i) = \log(\pi_i/(1-\pi_i))$ ekvivalent med $g^{-1}(\eta_i) = \frac{\exp(\eta_i)}{1+\exp(\eta_i)}$
- Probit-analyse: $g(\pi_i) = \Phi^{-1}(\pi_i)$ der $\Phi()$ er kumulativ for N(0,1)
- "Cauchit-analyse" $g(p_i) = \tan(\pi(p_i - 0.5))$
- clog-log-link $g(\pi_i) = \log(-\log(1-\pi_i))$ ekvivalent med $\pi_i = 1 - \exp(-\exp(\eta_i))$
- log-link $g(\pi_i) = \log(\pi_i)$ (som ikke er invers av kumulativ over $\mathbb{R}$)

Parameterfortolkning logistisk regresjon

Vi definerer odds for begivenhet ved: $\frac{\pi}{1-\pi} = \text{Odds}$

For logistisk regresjon blir oddsen, med $\eta = \beta' x$,

$$\text{Odds} = \frac{\frac{\exp(\eta)}{1+\exp(\eta)}}{1 - \frac{\exp(\eta)}{1+\exp(\eta)}} = \frac{\frac{\exp(\eta)}{1+\exp(\eta)}}{\frac{1}{1+\exp(\eta)}} = \exp(\eta)$$

La to kovariatvektor være $x = (x_1, \ldots, x_p)$ og $x' = (x'_1, \ldots, x'_p)$ der

$$x'_j = x_j + 1$$
$$x'_k = x_k \text{ for } k \neq j$$

slik at $x' - x = (0, \ldots, 0, 1, 0, \ldots, 0)$,
kovariatene aviker bare for komponent j.

Parameterfortolkning logistisk regresjon: Odds-ratio

Da blir forholdet mellom oddsene med kovariater x og x', kalt *odds-ratioen*, (med $\pi' = e^{\eta'}/(1+e^{\eta'})$ og $\eta' = \beta' x'$)

$$\begin{aligned}\text{OR}_j &= \frac{\frac{\pi'}{1-\pi'}}{\frac{\pi}{1-\pi}} = \frac{\text{Odds}'}{\text{Odds}} = \exp(\eta' - \eta) = \exp(\beta'(x' - x)) \\ &= \exp(\beta_j)\end{aligned}$$

eller omvendt

$$\beta_j = \log(\text{OR}_j),$$

dvs. regresjonsparametrene fortolkes som log-odds-ratioer (eller log-odds-forhold).

Legg merke til at

$$\text{OR}_j = \frac{\pi'}{\pi}\frac{1-\pi}{1-\pi'}$$

Odds-ratio ≈ Relativ Risk når sannsynlighetene er små

En "relativ risk" er definert som forholdet mellom to sannsynligheter, f.eks.

$$\text{RR} = \frac{\pi'}{\pi}$$

Spesielt når både π og π' er små blir $1-\pi \approx 1$ og $1-\pi' \approx 1$.
Dermed får vi

$$\text{OR} = \frac{\pi'}{\pi}\frac{1-\pi}{1-\pi'} \approx \frac{\pi'}{\pi} = \text{RR}$$

Faktisk: Tilnærmelsen er rimelig god selv når π og π' er så store som 0.2.

Tilnærmelsen OR ≈ RR for små sannsynligheter

	Relativ risk				Odds-ratio			
π	0.01	0.05	0.10	0.20	0.01	0.05	0.10	0.20
$\pi' = 0.01$	1	0.2	0.1	0.05	1.00	0.19	0.09	0.04
$\pi' = 0.05$	5	1.0	0.5	0.25	5.21	1.00	0.47	0.21
$\pi' = 0.10$	10	2.0	1.0	0.50	11.00	2.11	1.00	0.44
$\pi' = 0.20$	20	4.0	2.0	1.00	24.75	4.75	2.25	1.00

Derimot når sannsynlighetene er nær 0.5

F.eks $\pi = 0.4$ og $\pi = 0.6$ får vi odds-ratio

$$\text{OR} = \frac{\pi'}{\pi}\frac{1-\pi}{1-\pi'} = \frac{0.6}{0.4}\frac{0.6}{0.4} = 2.25 = 1.5^2 = \text{RR}^2$$

eller litt mer generelt $\pi' = 0.5 + \delta$ og $\pi = 0.5 - \delta$.

Da blir $1-\pi' = 0.5 - \delta = \pi$ og $1-\pi = 0.5 + \delta = \pi'$ slik at

$$\text{OR} = \frac{\pi'}{\pi}\frac{1-\pi}{1-\pi'} = \left(\frac{\pi'}{\pi}\right)^2 = \text{RR}^2$$

dvs. ikke tilnærmelse mellom størrelsene
og OR avviker vesentlig mer fra 1 enn RR

På den annen side: Hvis sannsynlighetene er nær 1

F.eks. hvis $\pi = 0.98$ og $\pi' = 0.99$ blir

$$\text{RR} = \frac{0.99}{0.98} \approx 1.01,$$

mens

$$\text{OR} = \frac{0.99}{0.98}\frac{0.02}{0.01} \approx 2.02,$$

så OR har en symmetriegenskap som "fanger" opp at
$\text{P}(Y=0) = 2\text{P}(Y'=0)$

(I slike tilfeller vil en antagelig redefinere responsen
$Y_{\text{ny}} = 1 - Y$.)

Uttrykket odds: Spill

I ett pengespill satser man en innsats I og får deretter utbetalt $U = G_0 + I$ hvis man vinner. Hvis man taper får man ikke innsatsen tilbake. Gevinsten etter å ha spilt er derfor

$$G = \begin{cases} -I & \text{hvis en taper spillet} \\ G_0 & \text{hvis en vinner spillet} \end{cases}$$

Vi antar at sannsynlighet for å vinne er p. Hvis spillet er rettferdig er

$$0 = \mathrm{E}[G] = G_0 p - I(1-p),$$

dvs.

$$G_0 = I\frac{1-p}{p} = I\frac{q}{1-q} = I \text{ Odds}$$

der $q = 1 - p$ og Odds $= \frac{q}{1-q}$.

Uttrykket odds, forts.

Oddsen i spill som "Langoddsen" tar imidlertid utgangspunkt i *utbetalt* beløp $U = G_0 + I$ og med rettferdig spill skulle

$$\text{Odds} = \frac{U}{I} = \frac{G_0 + I}{I} = \frac{1-p}{p} + 1 = \frac{1}{p}$$

For et spill som Langoddsen (fotballtipping) kan man tippe på hjemmeseier (Odds H), Uavgjort (Odds U) og borteseier (med Odds B). Med et rettferdig spill skulle da

$$\begin{aligned} 1 &= \text{P(Hjemmeseier)} + \text{P(Uavgjort)} + \text{P(Borteseier)} \\ &= \frac{1}{\text{Odds H}} + \frac{1}{\text{Odds U}} + \frac{1}{\text{Odds B}} \end{aligned}$$

Det kommer imidlerid på en provisjon, slik at typisk

$$\frac{1}{\text{Odds H}} + \frac{1}{\text{Odds U}} + \frac{1}{\text{Odds B}} \approx 1.25$$

Eksempel: Fra Langodds

1/H	1/U	1/B	1/H+1/U+1/B
0.870	0.204	0.175	1.249
0.513	0.270	0.465	1.248
0.588	0.303	0.357	1.248
0.417	0.317	0.513	1.247
0.800	0.233	0.217	1.250
0.417	0.317	0.513	1.247
0.476	0.274	0.500	1.250
0.571	0.357	0.317	1.246
0.645	0.357	0.247	1.249
0.400	0.417	0.435	1.251
0.400	0.417	0.435	1.251
0.588	0.244	0.417	1.249
0.488	0.274	0.488	1.250
0.488	0.400	0.364	1.251
0.625	0.225	0.400	1.250
0.282	0.303	0.667	1.251
0.513	0.392	0.345	1.250
0.556	0.377	0.317	1.250
0.278	0.345	0.625	1.248
0.333	0.357	0.556	1.246
0.278	0.345	0.625	1.248

Log-link for binære data

Anta $\pi = \exp(\beta' x) = \exp(\eta)$ (eller $\log(\pi) = \beta' x = \eta$)
og la som før $x' - x = (0, \ldots, 0, 1, 0, \ldots, 0)$, kovariatene aviker bare for komponent j. Da får vi

$$\exp(\beta_j) = \frac{\exp(\eta')}{\exp(\eta} = \text{RR}_j$$

dvs. log-link gir at regresjonsparametrene er en log-relativ risk

$$\beta_j = \log(\text{RR}_j)$$

Desverre kan $\pi = \exp(\eta)$ være større enn 1, og det er derfor praktiske problemer med å benytte denne linkfunksjonen.

Parameter-fortolkning med clog-log-link

$\pi = 1 - \exp(-\exp(\beta' x))$ eller $\eta = \beta' x = \log(-\log(1-\pi))$

Når sannsynlighetene π og π' er små får vi faktisk også

$$\exp(\beta_j) \approx \frac{\pi'}{\pi} = \text{RR}_j$$

Dette fordi $1 - \exp(-\exp(\eta)) \approx 1 - (1 - \exp(\eta)) = \exp(\eta)$ når $\exp(\eta)$ er liten.

Denne linken har en nær tilknytning til proporsjonale hazard-modeller som benyttes mye for levetids-data. Vi skal derfor kalle $\exp(\beta_j) = \text{HR}_j$, "hazard-ratio" med denne linken.

Sammenligning HR med OR og RR

Vi antar at cloglog-linken er korrekt. Nokså generelt has enten

$$1 < \text{RR}_j < \text{HR}_j < \text{OR}_j$$

eller

$$\text{OR}_j < \text{HR}_j < \text{RR}_j < 1$$

dvs. OR avviker mest fra 1 og RR minst.

F.eks. la $\eta = \alpha + \beta x = \log(-\log(1-\pi))$ med binær kovariat slik at $x = 0$ og $x' = 1$. Sett $\pi(0) = \pi = 1 - \exp(-\exp(\alpha))$. Da får vi $\pi(1) = \pi' = 1 - (1 - \pi(0))^{HR}$ samt

$$\text{RR} = \frac{\pi(1)}{\pi(0)} = \frac{1 - (1-\pi(0))^{\text{HR}}}{\pi(0)}$$

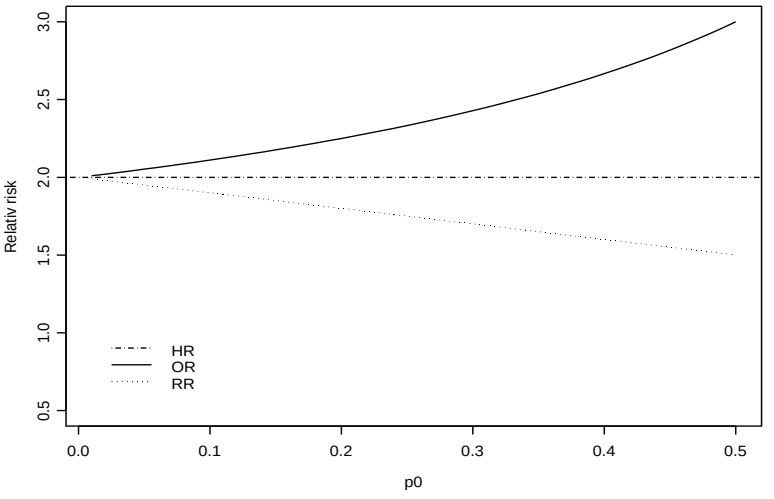
og

$$\text{OR} = \frac{\pi(1)(1-\pi(0))}{\pi(0)(1-\pi(1))} = \text{RR}(1-\pi(0))^{1-\text{HR}}$$

dvs. begge uttrykt ved $\pi(0)$ og HR.

Sammenligning HR, RR og OR

Fikserer HR=2. Plotter HR, RR og OR for $0 \leq \pi(0) \leq 0.5$:



Eksempel: Studie av dødelighet med Wilm's tumor

```
> glm(d~unfav+factor(stg),family=binomial(link=logit))$coef
 (Intercept)        unfav factor(stg)2 factor(stg)3 factor(stg)4
  -3.2415851    1.9927784    0.6957588    1.0305140    1.7935930
> glm(d~unfav+factor(stg),family=binomial(link=cloglog))$coef
 (Intercept)        unfav factor(stg)2 factor(stg)3 factor(stg)4
  -3.2240445    1.7404373    0.6591325    0.9664677    1.6147868
> glm(d~unfav+factor(stg),family=binomial(link=log))$coef
Error: no valid set of coefficients has been found: please supply starti
>
> b1<-glm(d~unfav+factor(stg),family=binomial(link=logit))$coef
> b2<-glm(d~unfav+factor(stg),family=binomial(link=cloglog))$coef
> startverdi<-b2-(b1-b2)
> startverdi
 (Intercept)        unfav factor(stg)2 factor(stg)3 factor(stg)4
  -3.2065038    1.4880962    0.6225062    0.9024214    1.4359807
>
> glm(d~unfav+factor(stg),family=binomial(link=log),start=startverdi)
Coefficients:
 (Intercept)     unfav  factor(stg)2  factor(stg)3  factor(stg)4
     -3.1925    1.4723        0.6331        0.9107        1.3802

Degrees of Freedom: 3914 Total (i.e. Null);  3910 Residual
```

Eksempel: Sammendrag

Vi ser her

- Parmeter-estimater $\hat{\beta}_{\text{logit}}$ med logistisk regresjon større enn med cloglog ($\hat{\beta}_{\text{cloglog}}$)
- Initielt ingen konvergens med log-link
- Foreslår startverdi $\beta_0 = \hat{\beta}_{\text{cloglog}} - (\hat{\beta}_{\text{logit}} - \hat{\beta}_{\text{cloglog}})$ (gjetter at cloglog-estimater ligger midt mellom logit- og log-estimater)
- Med startverdien får vi konvergens også for log-link
- Har $0 < \hat{\beta}_{\text{log}} < \hat{\beta}_{\text{cloglog}} < \hat{\beta}_{\text{logit}}$
- Dermed $1 < \text{RR}_j < \text{HR}_j < \text{OR}_j$

Spesialtilfelle: En binær kovariat

slik at $x' = 1$ og $x = 0$. Vi betegner individene med $x = 0$ for gruppe 1 og dem med $x' = 1$ for gruppe 2. Vi lar også at Y være en indikator for sykdom. Dataene kan da aggregeres til en 2x2 tabell

Tabell 1

	Gruppe 1	Gruppe 0	Totalt
Syke	A	B	A+B
Friske	C	D	C+D
Totalt	A+C	B+D	n=A+B+C+D

Odds-ratio i 2x2-tabell

La $\pi(0)$ og $\pi(1)$ være sannsynligheten for sykdom i gruppe 0 og gruppe 1 (hhv). Da blir

$$A \sim \text{Bin}(A + C, \pi(1))$$
$$B \sim \text{Bin}(B + D, \pi(0))$$

og vi estimerer $\hat{\pi}(1) = \frac{A}{A+C}$ og $\hat{\pi}(0) = \frac{B}{B+D}$.
Odds-ratio defineres tilsvarende logistisk regresjon ved

$$\text{OR} = \frac{\frac{\pi(1)}{1-\pi(1)}}{\frac{\pi(0)}{1-\pi(0)}} = \frac{\pi(1)}{\pi(0)}\frac{1-\pi(0)}{1-\pi(1)}$$

og estimeres ved

$$\widehat{\text{OR}} = \frac{\hat{\pi}(1)}{\hat{\pi}(0)}\frac{1-\hat{\pi}(0)}{1-\hat{\pi}(1)} = \frac{\frac{A}{A+C}}{\frac{B}{B+D}}\frac{\frac{D}{B+D}}{\frac{C}{A+C}} = \frac{AD}{BC}$$

Fra Oblig I, punkt m

has at variansen til log-odds $\log(\frac{\hat{\pi}(1)}{1-\hat{\pi}(1)})$ estimeres med

$$\hat{V}_1 = \frac{1}{A} + \frac{1}{C}$$

og tilsvarende variansen til $\log(\frac{\hat{\pi}(0)}{1-\hat{\pi}(0)})$ estimert ved

$$\hat{V}_0 = \frac{1}{B} + \frac{1}{D}$$

og siden

$$\log(\widehat{\text{OR}}) = \log(\frac{\hat{\pi}(1)}{1-\hat{\pi}(1)}) - \log(\frac{\hat{\pi}(0)}{1-\hat{\pi}(0)})$$

der de to log-oddsene er uavhengige blir variansestimatet for $\log(\widehat{\text{OR}})$:

$$se^2 = \hat{V}_1 + \hat{V}_0 = \frac{1}{A} + \frac{1}{B} + \frac{1}{C} + \frac{1}{D}$$

95% Konfidensintervall (KI) for OR:

Fra $\log(\widehat{\text{OR}}) \sim \text{N}(\log(\text{OR}), se^2)$ blir

$$\log(\widehat{\text{OR}}) \pm 1.96se$$

et tilnærmet 95% KI for $\log(\text{OR})$. Dette betyr at

$$\text{P}(\log(\widehat{\text{OR}}) - 1.96se < \log(\text{OR}) < \log(\widehat{\text{OR}}) + 1.96se) \approx 0.95,$$

men denne ulikheten er ekvivalent med

$$\widehat{\text{OR}} \exp(-1.96se) < \text{OR} < \widehat{\text{OR}} \exp(+1.96se)$$

og har samme sannsynlighet. Derfor får vi at

$$\widehat{\text{OR}} \exp(\pm 1.96se)$$

blir et tilnærmet 95% KI for OR.

Eksempel: Dødelighet blant barn med Wilm's tumor

```
> table(d,unfav)
   unfav
d      0    1
  0 3206  265
  1  270  174

> OR<-(174*3206)/(265*270)
> OR
[1] 7.796562

> sesq<-1/3206+1/265+1/270+1/174
> OR*exp(c(-1,1)*1.96*sesq^0.5)
[1] 6.206798 9.793516
```

Vi finner altså at estimert OR med 95% KI blir

$$7.80(6.21, 9.79)$$

Fortolkning av parametre med probitanalyse

Noen ganger har vi kontinuerlige responser, $Y_{i0} \sim \text{N}(\beta' x_i, \sigma^2)$ (f.eks. normalfordelt), men velger å studere

$$Y_i = \begin{cases} 1 & \text{hvis } Y_{i0} < \gamma = \text{ terskelverdi} \\ 0 & \text{hvis ikke} \end{cases}$$

Eks. $Y_{i0} =$ fødselsvekt

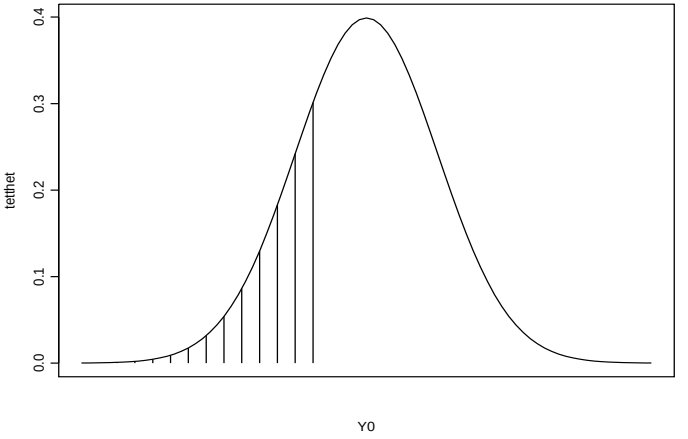
$$Y_i = \begin{cases} 1 & \text{hvis } Y_{i0} < 2500 \text{ gram} \\ 0 & \text{hvis ikke} \end{cases}$$

Eks. Psykometriske målinger, $Y_{i0} =$ score på depresjonsskala

$$Y_i = \begin{cases} 1 & \text{hvis } Y_{i0} > \text{ terskelverdi} \\ 0 & \text{hvis ikke} \end{cases}$$

Underliggende skala

$$Y_i = \begin{cases} 1 & \text{hvis } Y_{i0} < \gamma = \text{ terskelverdi} \\ 0 & \text{hvis ikke} \end{cases}$$



Probit, forts.

Hvorfor binære respons?

- Tradisjon for tabellanalyse
- Direkte score Y_{i0} kan være svært skjevfordelt
- Direkte score er kanskje ikke registert, bare noe vi forestiller oss ("latent" variabel)

Vi finner sammenhengen mellom

- $Y_{i0} \sim \mathrm{N}(\beta' x_i, \sigma^2)$
- $Y_i = \mathrm{I}(Y_{i0} \leq \gamma)$

ved

$$\pi_i = \mathrm{P}(Y_i = 1) = \mathrm{P}(Y_{i0} \leq \gamma) = \Phi(\frac{\gamma}{\sigma} - (\frac{\beta}{\sigma})' x_i)$$

Sammenheng paramtre i probit og underliggende skala

Forventning for $\mathrm{E}[Y_{i0}] = \beta' x_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}$ svarer altså til probitmodell

$$\Phi^{-1}(\pi_i) = \alpha_0 + \alpha_1 x_{i1} + \cdots + \alpha_p x_{ip}$$

der

- $\alpha_0 = \frac{\gamma - \beta_0}{\sigma}$
- $\alpha_j = \frac{-\beta_j}{\sigma}$ for $j = 1, \ldots, p$

Merk: Standardavviket σ for den underliggende skalaen er ikke mulig å identifisere.

Eksempel: Fødselsvekt og svangerskapsvarighet

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

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -1447.24     784.26  -1.845   0.0791 .
svlengde      120.89      20.46   5.908 7.28e-06 ***
sex          -163.04      72.81  -2.239   0.0361 *
---
Residual standard error: 177.1 on 21 degrees of freedom
Multiple R-Squared:  0.64,      Adjusted R-squared: 0.6057
F-statistic: 18.67 on 2 and 21 DF,  p-value: 2.194e-05
```

Vi får altså estimert $\hat{\sigma} = 177.1$.

Eksempel: Fødselsvekt og svangerskapsvarighet, forts.

```
> lavvekt<-1*(vekt<2800)
> table(lavvekt)
 0  1
17  7
>
> glm(lavvekt~svlengde+sex,family=binomial(link=probit))$coef
(Intercept)    svlengde         sex
 24.1550285  -0.6801164   0.7522067


> lm(vekt~svlengde+sex)$coef/177.1
(Intercept)    svlengde         sex
 -8.1718986   0.6826331  -0.9206059
```

Definerer $Y_i = 1$ hvis fødselsvekten er mindre enn 2800 gram.

Får probit-estimater $\hat{\alpha}_j \approx -\frac{\hat{\beta}_j}{\hat{\sigma}}$ fra lineær regresjon.

Goodness of fit-tester for binomiske data

Hvis $Y_i \sim \text{Bin}(n_i, \pi_i)$ og (a) $n_i\pi_i > 5$ og (b) $n_i(1-\pi_i) > 5$ for $i = 1, \ldots, N$ er tilnærmet

Residual devians	$\Delta = 2(\tilde{l} - \hat{l}) \sim \chi^2_{N-p}$
Pearson kjikvadrat	$X^2 = \sum_{i=1}^{n} \frac{(Y_i - \hat{\pi}_i)^2}{n_i\hat{\pi}_i(1-\hat{\pi}_i)} \sim \chi^2_{N-p}$

der $\tilde{l}$ er log-likelihood i mettet modell, $\hat{l}$ log-likelihood for den tilpassede modellen med p parametre og $\hat{\pi}_i$ estimerte sannsynligheter in denne modellen.
Hvis D og X^2 er vesentlig større enn $N - p$ tyder det på at modellen passer dårlig.

Ofte er imidlertid Y_i-ene binære og betingelsen (a) og (b) er da ikke oppfylt.

To strategier for goodness-of fit med binære data

- Med kategoriske kovariater: Aggreger til binomiske data
- Hosmer-Lemeshow test

Aggregering består i å

- Tell opp antall individer etter alle nivåer av de kategoriske variablene
- Tell opp antall $Y_i = 1$ etter alle nivåer av de kategoriske variablene
- Gjør `glm`-tilpasning på aggregerte data
- Modellen er OK hvis D og X^2 små i forhold til $\chi^2_{\tilde{N}-p}$ der $\tilde{N}$ er antall komb. av nivåer over de kategoriske variablene
- Dette krever at forventet antall suksesser og fiaskoer i hver gruppe > 5

Eks. Aggregering: Wilm's tumor

```
> unfavaggr<-c(rep(0,4),rep(1,4))
> stgaggr<-rep(1:4,2)
> naggr<-numeric(0)
> for (i in 1:8)
  naggr[i]<-sum(unfav==unfavaggr[i]&stg==stgaggr[i])
> daggr<-numeric(0)
> for (i in 1:8)
  daggr[i]<-sum(d[unfav==unfavaggr[i]&stg==stgaggr[i]])

> glmfit<-glm(cbind(daggr,naggr-daggr)~unfavaggr+factor(stgaggr),
                                                  family=binomial)
> glmfit
(Intercept) unfavaggr factor(stgaggr)2 factor(stgaggr)3 factor(stgaggr)4
   -3.2416     1.9928           0.6958           1.0305           1.7936

Degrees of Freedom: 7 Total (i.e. Null);  3 Residual
Null Deviance:      413.4
Residual Deviance: 3.33         AIC: 56.85

> X2<-sum(residuals(glmfit,type="pearson")^2)
> X2
[1] 3.259168
```


Eks. Aggregering: Wilm's tumor

Siden residual devians $D = 3.33 \approx X^2 = 3.26 =$ Pearson kjikvadrat er lite sammenlignet med residualt antall frihetsgrader $df = 3$ virker modellen OK.

Men er forventet antall suksesser og "fiaskoer" > 5? Ja, beregner disse:

```
> round(naggr*glmfit$fit,2)
    1     2     3     4     5     6     7     8
53.81 63.55 75.95 76.70 25.19 43.45 61.05 44.30
> round(naggr*(1-glmfit$fit),2)
      1       2       3       4       5       6       7       8
1376.19  810.45  693.05  326.30   87.81   75.55   75.95   25.70
```

Hosmer-Lemeshow test

Hvis mange kategoriske variable eller skala-kovariater vil ikke aggregering hjelpe. Kan istedet bruke Hosmer-Lemeshow test:

- Gjør `glm`-tilpasning
- Ordner individene etter tilpassede sannsynligheter $\hat{\pi}_{(1)} \leq \hat{\pi}_{(2)} \leq \cdots \leq \hat{\pi}_{(n)}$
- Lager 10 like store grupper etter ordningen
- Beregner $\bar{\pi}_{gr}$ = gj.sn. av $\hat{\pi}_{(i)}$ i gruppe $gr = 1, 2, \ldots, 10$
- Beregner antall observasjoner n_{gr} og antall suksesser Y_{gr} i gruppe gr
- Beregner Hosmer-Lemeshow $X^2_{hl} = \sum_{gr=1}^{10} \frac{(Y_{gr} - n_{gr}\bar{\pi}_{gr})^2}{n_{gr}\bar{\pi}_{gr}(1-\bar{\pi}_{gr})}$
- Hvis modellen er OK has tilnærmet $X^2_{hl} \sim \chi^2_8$, dvs. $df = 10 - 2 = 8$

Eks. X^2_{hl}: Wilm's tumor

```
> glmfit<-glm(d~unfav+factor(stg)+yr.regis+age,family=binomial)
> kuttoff<-sort(glmfit$fit)[c(round(length(d)*(1:10)/10))]
> gr<-rep(1,length(d))
> for (i in 1:9) gr<-gr+(glmfit$fit>kuttoff[i])
> table(gr)
  1   2   3   4   5   6   7   8   9  10
392 392 391 392 392 390 391 392 392 391
> ngr<-as.numeric(table(gr))
> ngr
 [1] 392 392 391 392 392 390 391 392 392 391
> dgr<-numeric(0)
> for (i in 1:10) dgr[i]<-sum(d[gr==i])
> dgr
 [1]  10  14  16  26  20  28  36  48  79 167
> for (i in 1:10) pigr[i]<-mean(glmfit$fit[gr==i])
> round(pigr,3)
 [1] 0.024 0.032 0.040 0.049 0.061 0.076 0.095 0.128 0.202 0.427
> X2HL<-sum((dgr-ngr*pigr)^2/(ngr*pigr*(1-pigr)))
> X2HL
[1] 3.482061
> 1-pchisq(X2HL,8)
[1] 0.9005774
```