

Statistical Practice in Epidemiology

Poisson and Logistic Regression

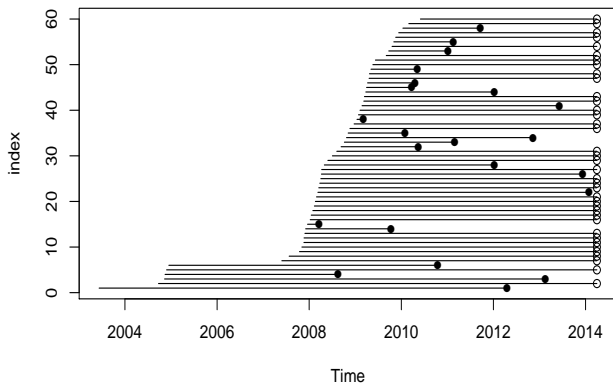
Janne Pitkaniemi (EL)

Points to be covered

- ▶ Incidence rates, rate ratios and rate differences from *follow-up studies* can be computed by fitting *Poisson regression models*.
- ▶ Odds ratios can be computed from binary data by fitting *Logistic regression models*.
- ▶ Odds-ratios can be estimated from case-control studies.
- ▶ Both models are special instances of *Generalized linear models*.
- ▶ There are various ways to do these tasks in R.

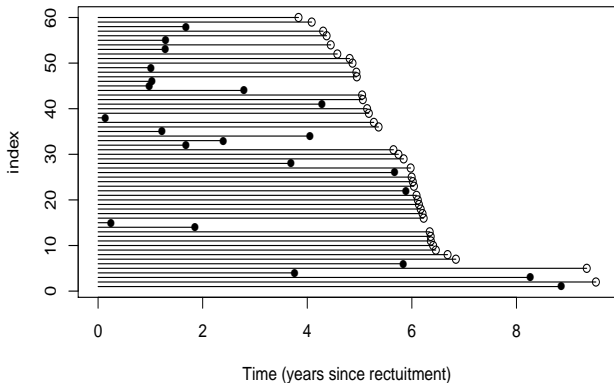
The Estonian Biobank cohort: survival among the elderly

Follow-up of 60 random individuals aged 75-103 at recruitment, until death (●) or censoring (○) in April 2014 (linkage with the Estonian Causes of Death Registry).



The Estonian Biobank cohort: survival among the elderly

Follow-up time for 60 random individuals aged 75-103 at recruitment (time-scale: time in study).



Events, dates and risk time

- ▶ Mortality as the outcome:

d: indicator for **status** at exit:

1: death observed

0: censored alive

- ▶ Dates:

doe = date of **E**ntry to follow-up,

dox = date of e**X**it, end of follow-up.

- ▶ Follow-up time (years) computed as:

$$y = (\text{dox} - \text{doe}) / 365.25$$

Crude overall rate computed in two ways

Total no. cases, person-years & rate (/1000 y):

```
> D <- sum( d ); Y <- sum(y) ; R <- D/(Y/1000)
> round( c(D=D, Y=Y, R=R), 2)
      D      Y      R
884.00 11678.24  75.70
```

Poisson regression model with only intercept ("1").

```
> m1 <- glm( d ~ 1, family=poisson, offset=log(y))
> coef(m1)
( Intercept )
-2.581025

> exp( coef(m1) )*1000
( Intercept )
75.69636
```

Why do we get the same results?

Constant hazard — Poisson model

Let $Y \sim \exp(\lambda)$, then $f(y; \lambda) = \lambda e^{-\lambda y} I(y > 0)$

Constant rate: $\lambda(y) = \frac{f(y; \lambda)}{S(y; \lambda)} = \lambda$

Observed data $\{(y_i, \delta_i); i = 1, \dots, n\}$.

The likelihood $L(\lambda) = \prod_{i=1}^n \lambda^{\delta_i} e^{-\lambda y_i}$ and

$$\log(L) = \sum_{i=1}^n [\delta_i \log(\lambda) - \lambda y_i]$$

Solving the *score equations*: $\frac{\partial \log L(\lambda)}{\partial \lambda} = \sum \left[\frac{\delta_i}{\lambda} - y_i \right]$
 $= \frac{D}{\lambda} - Y = 0$ and $D - \lambda Y = 0$

→ **maximum likelihood estimator** (MLE) of λ :

$$\hat{\lambda} = \frac{D}{Y} = \frac{\text{number of cases}}{\text{total person-time}} = \text{empirical rate!}$$

offset term — Poisson model

- ▶ Previous model without offset: Intercept $6.784 = \log(884)$
- ▶ We should use an offset if we suspect that the underlying **population sizes (person-years) differ** for each of the observed counts – For example varying person-years by treatment group, sex, age,...
- ▶ We need a term in the model that "scales" the likelihood, but does not depend on model parameters (include a **term with reg. coef. fixed to 1**) – offset term is $\log(y)$

$$\begin{aligned}\log\left(\frac{\mu}{y}\right) &= \beta_0 + \beta_1 x_1 \\ \log(\mu) &= 1 \times \log(y) + \beta_0 + \beta_1 x_1\end{aligned}$$

Comparing rates: The Thorotrast Study

- ▶ Cohort of seriously ill patients in Denmark on whom angiography of brain was performed.
- ▶ Exposure: contrast medium used in angiography,
 1. `thor` = thorotrast (with ^{232}Th), used 1935-50
 2. `ctrl` = other medium (?), used 1946-63
- ▶ Outcome of interest: death

`doe` = date of **E**ntry to follow-up,

`dox` = date of e**X**it, end of follow-up.

- ▶ `data(thoro)` in the `Epi` package.

Comparing rates: thorotrast vs. control

Tabulating cases, person-years & rates by group

```
> stat.table( contrast ,  
+             list ( N = count(),  
+                   D = sum(d),  
+                   Y = sum(y),  
+                   rate = ratio(d,y,1000) ) )
```

contrast	N	D	Y	rate
ctrl	1236	797.00	30517.56	26.12
thor	807	748.00	19243.85	38.87

Rate ratio, $RR = 38.89/26.12 = 1.49$,

Std. error of log-RR, $SE = \sqrt{1/748 + 1/797} = 0.051$,

Error factor, $EF = \exp(1.96 \times 0.051) = 1.105$,

95% confidence interval for RR:

$(1.49/1.105, 1.49 \times 1.105) = (1.35, 1.64)$.

Rate ratio estimation with Poisson regression

- ▶ Include contrast as the explanatory variable (factor).
- ▶ Insert person years in units that you want rates in

```
> m2 <- glm( d ~ contrast, offset=log(y/1000),  
+           family = poisson )  
> round( summary(m2)$coef, 4)[, 1:2]
```

	Estimate	Std. Error
(Intercept)	3.2626	0.0354
contrast thor	0.3977	0.0509

- ▶ Rate ratio and CI?
Call function `ci.exp()` in `Epi`

```
> round( ci.exp( m2 ), 3 )
```

	exp(Est.)	2.5%	97.5%
(Intercept)	26.116	24.364	27.994
contrast thor	1.488	1.347	1.644

Rates in groups with Poisson regression

- ▶ Include contrast as the explanatory variable (factor).
- ▶ Remove the intercept (-1)
- ▶ Insert person-years in units that you want rates in

```
> m3 <- glm( d ~ contrast - 1,
              offset=log(y/1000),
              family = poisson )
> round( summary(m3)$coef, 4)[, 1:2]
```

	Estimate	Std. Error
contrast ctrl	3.2626	0.0354
contrast thor	3.6602	0.0366

```
> round( ci.exp( m3 ), 3 )
```

	exp(Est.)	2.5%	97.5%
contrast ctrl	26.116	24.364	27.994
contrast thor	38.870	36.181	41.757

Rates in groups with Poisson regression

- You can have it all in one go:

```
> CM <- rbind( c(1,0), c(0,1), c(-1,1) )  
> rownames(CM) <- c("Ctrl","Thoro","Th vs.Ct")  
> colnames(CM) <- names( coef(m3) )  
> CM
```

	contrast ctrl	contrast thor
Ctrl	1	0
Thoro	0	1
Th vs. Ct	-1	1

```
> round( ci.exp( m3, ctr.mat=CM ),3 )
```

	exp(Est.)	2.5%	97.5%
Ctrl	26.116	24.364	27.994
Thoro	38.870	36.181	41.757
Th vs. Ct	1.488	1.347	1.644

Rate ratio estimation with Poisson regression

- ▶ Response may also be specified as individual *rates*:
d/y

weights= instead of offset= are needed.

```
> m4<-glm( d/(y/1000)~contrast, weights=y/1000,  
+          family=poisson)  
> round( ci.exp(m4), 3 )
```

	exp(Est.)	2.5%	97.5%
(Intercept)	26.116	24.365	27.994
contrast thor	1.488	1.347	1.644

Rate difference estimation with Poisson regression

- The approach with d/y enables additive rate models too:

```
> m5 <-glm(d/(y/1000) ~contrast,weights=y/1000,  
+          family=poisson(link="identity") )  
> round( ci.exp(m5,Exp=F), 3 )
```

	Estimate	2.5%	97.5%
(Intercept)	26.116	24.303	27.929
contrast thor	12.753	9.430	16.077

Rates difference

- ▶ As before you can have it all:

```
> m6 <- glm( d/(y/1000) ~ contrast -1,  
+ family = poisson(link="identity"),  
+ weights = y/1000)  
> round(ci.exp(m6, ctr.mat=CM, Exp=F ), 3)
```

	Estimate	2.5%	97.5%
Ctrl	26.116	24.303	27.929
Thoro	38.870	36.084	41.655
Th vs. Ct	12.753	9.430	16.077

```
> round( ci.exp( m3, ctr.mat=CM), 3 )
```

	exp(Est.)	2.5%	97.5%
Ctrl	26.116	24.364	27.994
Thoro	38.870	36.181	41.757
Th vs. Ct	1.488	1.347	1.644

Binary data: Treatment success Y/N

85 diabetes-patients with foot-wounds:

- ▶ Dalterapin (Dal)
- ▶ Placebo (PI)

Treatment/Placebo given to diabetes patients, the design is prospective because outcome is measured better/worse. Is the probability of outcome more than 15% – yes, then use the risk difference or risk ratio (RR)

	Treatment group	
	Dalterapin	Placebo
Better	29	20
Worse	14	22
Total	43	42

$$\hat{p}_{\text{Dal}} = \frac{29}{43} = 67\% \quad \hat{p}_{\text{PI}} = \frac{20}{42} = 47\%$$

The difference between the probabilities is the fraction of the patients that benefit from the treatment: $p_{\text{Dal}} - p_{\text{PI}}$

```
> library(Epi)
> dlt <- rbind( c(29,14), c(20,22) )
> colnames( dlt ) <- c("Better", "Worse")
> rownames( dlt ) <- c("Dal", "PI")
> kable(twoby2( dlt ), "latex")
```

2 by 2 table analysis :

/ ... /

	Better	Worse	P(Better)	95% conf. interval
Dal	29	14	0.6744	0.5226 0.7967
PI	20	22	0.4762	0.3316 0.6249

	95% conf. interval
Relative Risk: 1.4163	0.9694 2.0692
Sample Odds Ratio: 2.2786	0.9456 5.4907
Conditional MLE Odds Ratio: 2.2560	0.8675 6.0405
Probability difference : 0.1982	-0.0110 0.3850

Exact P-value: 0.0808

Logistic regression for binary data

For grouped binary data, the response is a two-column matrix with columns (successes, failures).

```
trt <- factor(c("Dal", "Pl"))  
trt <- relevel( trt , 2 )  
b1 <- glm( dlt ~ trt, family=binomial )  
round( ci.exp( b1 ), 4 )
```

	exp(Est.)	2.5%	97.5%
(Intercept)	0.9091	0.4962	1.6657
trtDal	2.2786	0.9456	5.4907

- ▶ The default parameters in logistic regression are **odds** (the intercept: $20/22 = 0.9090$) and the **odds-ratio** ($((29/14)/(20/22) = 2.28)$).
- ▶ This is not what you want, because odds ratio is biased estimate of the risk ratio. (recall if $p > 10\%$ $\frac{p}{1-p} \not\approx p$)

Logistic regression for binary data - Risk ratio (Relative risk)

```
> library(Epi)
> library(xtable)
> dlt <- rbind( c(29,14), c(20,22) )
> diab<-expand.grid(dlt)
> colnames(diab)[1]<-"d"
> diab$out <- c("Better","Better","Worse","Worse")
> diab$trt <- as.factor(c("Dal","Pl","Dal","Pl"))
> diab$totals<-rep(rowSums(dlt),2)
> diab$trt<-relevel( diab$trt, 2 )
> print(xtable(diab,digits=c(0,0,0,0,0)),include.rownames = F)
```

d	out	trt	totals
29	Better	Dal	43
20	Better	Pl	42
14	Worse	Dal	43
22	Worse	Pl	42

Logistic regression for binary data - risk ratio

```
> library(Epi)
> library(xtable)
> b2 <- glm(d/totals~trt,
+           weights=totals,
+           family=binomial(link="log"),
+           data=diab[c(1,2),])
> xtable(round( ci.exp( b2 ), digits=6 ))
```

	exp(Est.)	2.5%	97.5%
(Intercept)	0.48	0.35	0.65
trtDal	1.42	0.97	2.07

Diabetics with Daltelapin treatment have 1.4 times the risk of getting better than those treated with placebo

Logistic regression in case-control studies

- ▶ Model for disease occurrence in the target population:

$$\ln \left[\frac{p}{1-p} \right] = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

- ▶ Sampling fractions: $P(\text{inclusion in study} \mid \text{control}) = s_{\text{ctr}}$
 $P(\text{inclusion in study} \mid \text{case}) = s_{\text{case}}$
- ▶ Model for observed case-control data:

$$\begin{aligned} \ln[\text{odds (case } \mid \text{ incl.)}] &= \ln \left[\frac{p}{1-p} \right] + \ln \left[\frac{s_{\text{cas}}}{s_{\text{ctr}}} \right] \\ &= \left(\ln \left[\frac{s_{\text{cas}}}{s_{\text{ctr}}} \right] + \beta_0 \right) + \beta_1 x_1 + \beta_2 x_2 \end{aligned}$$

Logistic regression in case-control studies

Analysis of $P(\text{case}|\text{inclusion})$ — i.e. binary observations:

$$Y = \begin{cases} 1 & \sim \text{case} \\ 0 & \sim \text{control} \end{cases}$$

$$\ln[\text{odds (case | incl.)}] = \left(\ln \left[\frac{s_{\text{cas}}}{s_{\text{ctr}}} \right] + \beta_0 \right) + \beta_1 x_1 + \beta_2 x_2$$

- ▶ Effect of covariates is estimated correctly.
- ▶ Intercept is meaningless
depends on s_{cas} and s_{ctr} that are often unknown.

Case-control study: Food-poisoning outbreak

- ▶ An outbreak of acute gastrointestinal illness (AGI) occurred in a psychiatric hospital in Dublin in 1996.
- ▶ Out of all 423 patients and staff members, 65 were affected during 27 to 31 August, 1996.
- ▶ 65 cases and 62 randomly selected control subjects were interviewed.
- ▶ Exposure of interest: chocolate mousse cake.
- ▶ 47 cases and 5 controls reported having eaten the cake.

Ref: <http://www.eurosurveillance.org/ViewArticle.aspx?ArticleId=188> – here original numbers somewhat modified.

Outbreak: crude summary of data

- ▶ Target population information
 - ▶ $N = 423$, size of the whole study population
 - ▶ $D = 65$, no. of cases of AGI
 - ▶ $B = 358$, no. of non-cases
- ▶ Case-control data
 - ▶ $C = 62$, no. of controls, random sample from 358 non-cases
 - ▶ $f = 62/358 = 0.173$, sampling fraction of non-cases
 - ▶ $D1 = 47$ cases exposed to chocolate mousse
 - ▶ $D0 = 18$ unexposed cases
 - ▶ $C1 = 5$ controls exposed to chocolate mousse
 - ▶ $C0 = 57$ unexposed controls

Outbreak: results of analysis

Overall incidence proportion (IP) of AGI in the population

```
> D <- 65; N <- 423; IP <- D/N  
> round(IP, 3)
```

```
[1] 0.154
```

Analysis of case-control data

```
> D1 <- 47; D0 <- D - D1;  
> C <- 62 ; C1 <- 5; C0 <- C - C1
```

Case-control ratios by exposure (not as useful as the following!)

```
> round( c( D1/C1, D0/C0 ), 2)
```

```
[1] 9.40 0.32
```

Exposure odds in cases and controls

```
> round( c( D1/D0, C1/C0 ), 2)
```

```
[1] 2.61 0.09
```

Outbreak: results of analysis

Estimation of the incidence odds ratio (IOR) = exposure odds ratio

```
> IOR <- (D1/D0)/(C1/C0)
> SE.logIOR <- sqrt(1/D1 + 1/D0 + 1/C1 + 1/C0 )
> CI.IOR <- IOR * exp( c(-1,1)*1.96*SE.logIOR )
> round( c(IOR, SE.logIOR, CI.IOR ), 2)
```

```
[1] 29.77  0.54 10.28 86.21
```

Same with glm model

```
> count<-c(D1,D0,C1,C0)
> cc<-c(1,1,0,0)
> exposed<-c(1,0,1,0)
> mousse<-data.frame(cbind(cc,exposed,count))
> ci.exp(glm(cc~exposed,weights=count,family="binomial",data=mousse))
```

	exp(Est.)	2.5%	97.5%
(Intercept)	0.3157895	0.1858913	0.5364586
exposed	29.7666667	10.2778305	86.2102603

Conclusion: What did we learn?

- ▶ Poisson regression models.
- ▶ In Poisson models the response can be either:
 - ▶ case indicator d with `offset = log(y)`, or
 - ▶ rate d/y with `weights = y`.
- ▶ Both may be fitted on either grouped data, or individual records.
- ▶ Binary data can be modeled with odds.
- ▶ Case-control studies:
Odds-ratios can be computed by logistic regression models, but **Intercept** from model is **meaningless**.