Poisson and Logistic Regression

Janne Pitkäniemi

Finnish Cancer Registry

Tampere university

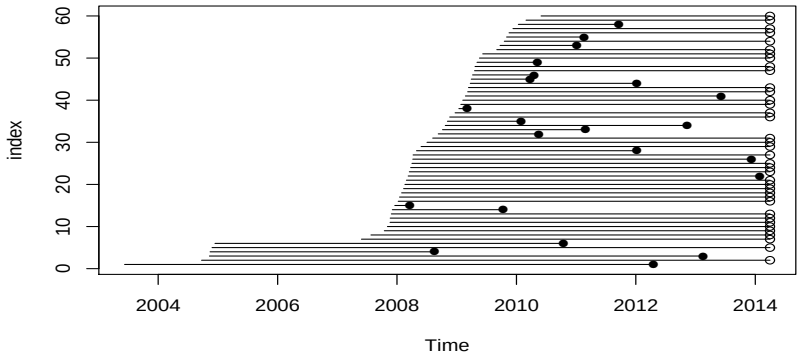
Statistical Practice in Epidemiology (2023, Tartu)

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*.
- Both models are special instances of *Generalized linear models*.
- There are various ways to do these tasks in R.

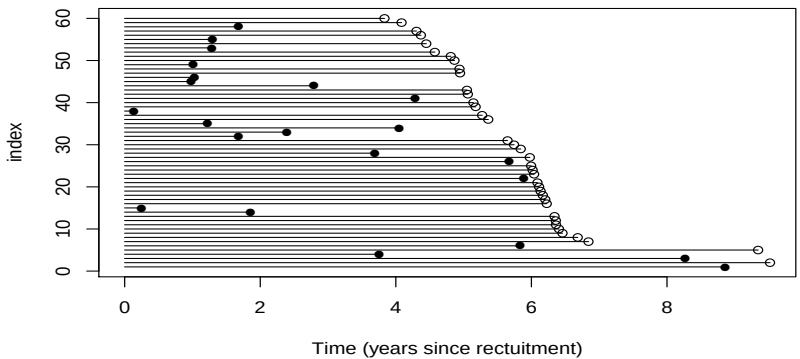
The Estonian Biobank cohort: survival among the elderly

Follow-up of 60 random individuals aged 75-103 at recruitment, until death (•) or censoring (o) in April 2014 (linkage with the Estonian Causes of Death Registry). (time-scale: calendar time).



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 = (dox - doe)/365.25
```

Crude overall rate computed by hand and model

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
```

R-implementation of the rate estimation with Poisson regression:
A model with poisreg−family (Epi package)

```
> glm(cbind(D, Y) ~1, family=poisreg)

 Coefficients :
( Intercept )
      −2.581
```

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 model: $\lambda(y) = \frac{f(y;\lambda)}{S(y;\lambda)} = \lambda$ and observed data $\{(y_i, \delta_i); i = 1, ..., n\}$.

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 $D - \lambda Y = 0$

$\rightarrow$ **maximum likelihood estimator** (MLE) of λ:

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

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 sex,age,treatment group,...
- 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)
- This is all taken care of by family=poisreg – recommend to use

$$log(\frac{\mu}{y}) = \beta_0 + \beta_1 x_1$$
$$log(\mu) = 1 \times log(y) + \beta_0 + \beta_1 x_1$$

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.

Tabulating 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 estimation with Poisson regression

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

```
> m2 <- glm( cbind(d,y/1000) ~ contrast,family = poisreg(link="log") )
> 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( cbind(d,y/1000) ~ factor(contrast)-1,family = poisreg)
> 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
```

Rate difference estimation with Poisson regression

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

```
> contrast<-c(0,1)
> m5 <-glm(cbind(d,y/1000) ~contrast,
          family=poisreg(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
```

Binary data: Treatment success Y/N

85 diabetes-patients with foot-wounds:

- Dalterapin (Dal)
- Placebo (Pl)

Treatment/Placebo given to diabetes patients, the design is prospective and outcome is measured better(Y)/worse(N). 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

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

Binary data: Crosstabulation analysis of 2x2 table

```
> library(Epi)
> dlt <- rbind( c(29,14), c(20,22) )
> colnames( dlt ) <- c("Better","Worse")
> rownames( dlt ) <- c("Dal","Pl")
> 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
Pl      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
       Asymptotic P-value: 0.0665
```

Binary regression – estimation of odds ratio

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

```
> library(Epi)
> library(xtable)
> dlt <- data.frame(rbind( c(29,14),c(20,22) ))
> colnames( dlt ) <- c("Better","Worse")
> dlt$trt <- c(1,0)
> b2<-glm(cbind(Better,Worse)~trt,
+           family=binomial(link="logit"),
+           data=dlt)
> xtable(round( ci.exp( b2 ), digits=6 ))
```

	exp(Est.)	2.5%	97.5%
(Intercept)	0.91	0.50	1.67
trt	2.28	0.95	5.49

- 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% $\frac{p}{1-p} \not\approx p$)

Binary regression - Estimation of risk ratio (Relative risk)

```
> library(Epi)
> library(xtable)
> dlt <- data.frame(rbind( c(29,14),c(20,22) ))
> colnames( dlt ) <- c("Better","Worse")
> dlt$trt <- c(1,0)
> b2<-glm(cbind(Better,Worse)~trt,
+           family=binomial(link="log"),
+           data=dlt)
> xtable(round( ci.exp( b2 ), digits=6 ))
```

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

Diabetics with Dalterapin treatment are 1.4 times likely to get better than those treated with placebo

Binary regression - Estimation of risk difference

```
> library(Epi)
> library(xtable)
> dlt <- data.frame(rbind( c(29,14),c(20,22) ))
> colnames( dlt ) <- c("Better","Worse")
> dlt$trt <- c(1,0)
> b2<-glm(cbind(Better,Worse)~trt,
+           family=binomial(link="identity"),
+           data=dlt)
> xtable(round( ci.exp( b2,Exp=F ), digits=6 ))
```

	Estimate	2.5%	97.5%
(Intercept)	0.48	0.33	0.63
trt	0.20	-0.01	0.40

Twenty percent more of the Diabetics with Dalterapin treatment are getting better compared to Diabetics treated with placebo

Conclusion: What did we learn?

- Rates, their ratio and difference can be analysed by Poisson regression
- In Poisson models the response can be either:
  - case indicator `d` with `offset = log(y)`, or
  - case and person-years `cbind(d,y)` with poisreg-family (Epi-package)
- Both may be fitted on either grouped data, or individual records.
- Binary outcome can be modeled with binary regression.