Epidemiology Models

i. General: $E[\#events] = Rate \times PT$

ii. Specific way that rates are interrelated (form of 'rate model')

(a) (Additive, Rate Difference): $Rate = Rate_0 + \beta_1 X_1 + \beta_2 X_2 \ldots$

(b) (Multiplicative, Rate Ratio): $Rate = Rate_0 \times \exp\{\beta_1 X_1 + \beta_2 X_2 \ldots\}$

(or, equivalently, ...): $\log(Rate) = \log(Rate_0) + \beta_1 X_1 + \beta_2 X_2 \ldots$

Statistical Fitting of these Models

i. General: $E[\#events] = Rate \times PT$

ii. Specifically, how model is implemented in statistical packages:
In both instances, expand the $Rate \times PT$ product

(a) (Add.): $E[\#events] = \{Rate_0 + \beta_1 X_1 + \beta_2 X_2 \ldots\} \times PT$

$E[\#events] = Rate_0 \times \underline{PT} + \beta_1 \times \underline{X_1 \times PT} + \beta_2 \times \underline{X_2 \times PT} \ldots$

(specify 'no-intercept' ; in `R`, $\#events \sim -1 + ...,$)

(b) (Mult): $E[\#events] = Rate_0 \times \exp\{\beta_1 X_1 + \beta_2 X_2 \ldots\} \times PT$

$\log\{E[\#events]\} = \log(Rate_0) + \beta_1 \times \underline{X_1} + \beta_2 \times \underline{X_2} \cdots + \log(PT)$

(use '$log(PT)$ as 'offset' ; cf worked e.g.'s for `R` / SAS code)

<u>1 Rate</u> : no. of cases / {amount of experience (P-T)}

- ***Inference Model:*** Poisson distribution for numerator.
- **Déjà:** Exact (discrete distrn.) & Gaussian approximations
- **New:** Regression Approach:

"Usual" Linear model : (***not appropriate***)

$$\begin{aligned} E[\text{cases}] &= \text{rate} \times \text{Denominator} \\ &= \beta \times \text{Denominator} \end{aligned}$$

"No-intercept" model (Gaussian variation around line)

```
# Ayas et al data on PI's in Residents

## Incidence, Overall

cases=c(498); InternMonths=c(17003);

## regression through origin,
   Gaussian variation around mean [true line]

summary( lm(cases ~ -1+InternMonths) )

Residuals:
ALL 1 residuals are 0: no residual degrees of freedom!

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
InternMonths  0.02929         NA      NA       NA

Residual standard error: NaN on 0 degrees of freedom
Multiple R-Squared:  1,   Adjusted R-squared:   NaN
F-statistic:   NaN on 1 and 0 DF,  p-value: NA
```

Estimate, namely 0.02929 cases/InternMonth is sensible.

<u>But</u> : no df with which to calculate SE or CI

- **New:** Regression Approach:

"***Generalized***" Linear model : (Poisson variation)

$$\begin{aligned} E[\text{cases}] &= \text{rate} \times \text{Denominator} \\ &= \beta \times \text{Denominator} \end{aligned}$$

same "no-intercept" model, but Poisson variation

```
summary(glm(cases ~ -1 + InternMonths,
            family = poisson(link=identity)))

Deviance Residuals: [1]  0

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
InternMonths 0.029289   0.001312   22.32   <2e-16
              --rate--

(Dispersion parameter for poisson family taken to be 1)

    Null deviance:        Inf  on 1  df
Residual deviance: 3.1086e-15  on 0  df     AIC: 10.049

Check: SE[rate]     = sqrt[#cases) / Denominator
               = sqrt[498] / 17003 = 0.001312
```

- **Since rates > 0; safer to model (natural) log of the rate**

$$\begin{aligned} E[\text{cases}] &= \text{rate} && \times && \text{Denominator} \\ \log [\, E[\text{cases}] \,] &= \log\{\text{rate}\} && + && \log[\,\text{Denominator}\,] \\ \log [\, \mu \mid \,] &= \gamma && + && \log[\,\text{Denominator}\,] \\ \log [\, \mu \mid \,] &= \gamma_0 && + && 1 \times \log[\,\text{Denominator}\,] \\ &= \gamma_0 && + && \gamma_1 \times \log[\,\text{Denominator}\,] \end{aligned}$$

(no need to estimate γ_1 ; already know $\gamma_1 == 1$)

in this instance, log[Denominator] is an "***<u>offset</u>***"

To model log [μ | x] , use log "***link***" (default link for Poisson)

<u>Canonical links</u> (Binomial : logit; Poisson: log) ensure that whatever the value of the linear predictor, any fitted proportion will be between 0 and 1, and any rate (no. cases) between 0 and infinity.

"*Generalized*" Linear model : (Poisson variation, log link)

```
summary( glm(cases ~ 1, family=poisson,
             offset=log(InternMonths) ) )

Deviance Residuals: [1]  0

Coefficients:
            Estimate Std. Error  z value Pr(>|z|)
(Intercept) -3.53055    0.04481*  -78.79   <2e-16

             log[ rate ]
    i.e.,    log[498/17003]

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 1.0747e-13  on 0  df
Residual deviance: 1.0747e-13  on 0  df    AIC: 10.049

*Check: SE[log rate] = sqrt[ 1 / #cases ]

                    = sqrt[ 1 /  498   ] = 0.04481
```

Note that we are able to calculate an SE for the estimate of the rate (previous page] and for the estimate of the log rate, even though we have no df with which to estimate the residual variation around he line (the line goes through our one data point). We are able to do this because the variance of a Poisson random variable is equal to the mean of the random variable. So, since the fitted mean no. of cases is 498, the model is able to provide an estimate of how much variation there would be if the mean were indeed 498, ie SD = sqrt[498]. The SE 'borrowed from' the model' is called a "model-based" SE.

In the usual regression with Gaussian variation, (i) the variance is estimated separately, using the mean of the squared residuals and (ii) the variance about the (true) line of means is assumed to be the same at all values of x. The Poisson model better reflects the variability of counts: the variation is higher when the expected (or average) count is higher (but the ***cv*** is smaller, the larger the count ie cv = σ/μ = sqrt[μ]/μ = 1/sqrt[μ].

Comparison of 2 Rates

• Rate difference / Ratio•

Example

Extended Periods (coded 'X' = 1)

35 percutaneous injuries in 26667 opportunities

vs.

Non-Extended Periods (coded 'X' = 0)

46 percutaneous injuries in 60763 opportunities

Regression framework for rate difference (RD)

observed rate when X = 0 (reference category)

$$b_0 = \text{rate}[0] = 46 / 60763 = \underline{0.0007570}$$

observed rate difference, rd

$$\text{rd} = 35/26667 - 46 / 60763 = \underline{0.0005554}$$

In general (single, binary X)

$$\begin{aligned} \text{RATE} \mid X &= \text{RATE}_0 + \text{RD} \times X \\ &= B_0 + B_1 \times X \end{aligned}$$

So...

$$\begin{aligned} E[\ \#\text{CASES} \mid X\ ] &= \text{RATE}_x \times \text{PT} \\ &= (B_0 + B_1 \times X) \times \text{PT} \\ &= B_0 \times \text{PT} + B_1 \times \underline{X \times \text{PT}} \\ &= B_0 \times Z_0 + B_1 \times Z_1 \end{aligned}$$

This is a regression with 2 terms (Z_0 & Z_1 ***), and no intercept***

Rate difference (rd), and CI for RD, via regression framework

```
cases=c(35,46); PT=c(26667,60763);
e=c(1,0); ePT = e*PT;

ds <- data.frame(cases=cases, opps=PT, extended=e,
extended.opps = ePT);

ds

  cases  opps extended extended.opps

1    35 26667        1         26667
2    46 60763        0             0

attach(ds)

# regression to obtain rate difference

summary(glm(cases ~ -1 + opps + extended.opps,
            family=poisson(link=identity)))

Deviance Residuals:
[1]  0  0

Coefficients:
               Estimate Std. Error z value Pr(>|z|)
opps          0.0007570  0.0001116   6.782 1.18e-11 ***
extended.opps 0.0005554  0.0002483   2.237   0.0253 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance:        Inf  on 2  df
Residual deviance: 1.3323e-15  on 0  df   AIC: 15.068
```

*Check: SE[rate difference] , as in Rothman,

$= \text{sqrt}[\ \text{Var}[\text{rate}_1] + \text{Var}[\text{rate}_0\ ]\ ] = \underline{0.0002483}$

Regression framework for rate ratio (RR) [2 Rates]

(same, e.g., ignoring, for now, the self-paired structure)

observed rate when X = 0 (reference category)

$$b_0 = \text{rate}[0] = 46 / 60763 = \underline{0.0007570}$$

observed rate ratio, rr

$$rr = (35/26667) / (46/60763) = \underline{1.73}$$

In general

$$\text{RATE} \mid X = \text{RATE}_0 \times \begin{matrix} 1 & \text{if } X=0 \\ RR & \text{if } X=1 \end{matrix}$$

$$= \text{RATE}_0 \times \exp[\ \log[RR] \times X\ ]$$

So...

$$\log[\ \text{RATE} \mid X] = \log[\text{RATE}_0] + \log[RR] \times X$$

$$= B_0 + B_1 \times X$$

So...

$$E[\ \#\text{CASES} \mid X\ ] = \text{RATE}_x \times PT$$

$$\log[\ E[\ \#\text{CASES} \mid X\ ]\ ] = \log[\text{RATE}_x] + \log[PT]$$

$$= B_0 + B_1 \times X + \log[PT]$$

$$= B_0 + B_1 \times X + 1 \times \underset{\text{"offset"}}{\underline{\log[PT]}}$$

(an "offset" is a term whose coefficient is KNOWN to be 1)

Rate ratio (rr), and CI for RR, via regression framework

```
# to obtain rate ratio  [ log.opps = log(opps) ]

summary(glm(cases ~ extended,
    family=poisson(link=log*),offset=log.opps))

Deviance Residuals:
[1]  0  0

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -7.1861     0.1474 -48.739   <2e-16 ***
extended      0.5503     0.2243   2.453   0.0142 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance:  5.8023e+00  on 1  df
Residual deviance: -3.7748e-15  on 0  df  AIC: 15.068
```

*Check:

-7.1861 = log[rate_0] = log[46/60763]

0.5503 = log[rateRatio]

= log[(35/26667) / (46/60763)]

So, rateratio = exp[log rateRatio] = exp[0.5503] = 1.73

SE[log of rate ratio] , as in Rothman,

= sqrt[1/ 35 + 1/46] = 0.2243

95% CI for log[RateRatio]: 0.5503 +/- 1.96 × 0.2243

95% CI for RateRatio: exp[0.5503 +/- 1.96 × 0.2243]

----

* no need to specify, as Log is default link for Poisson

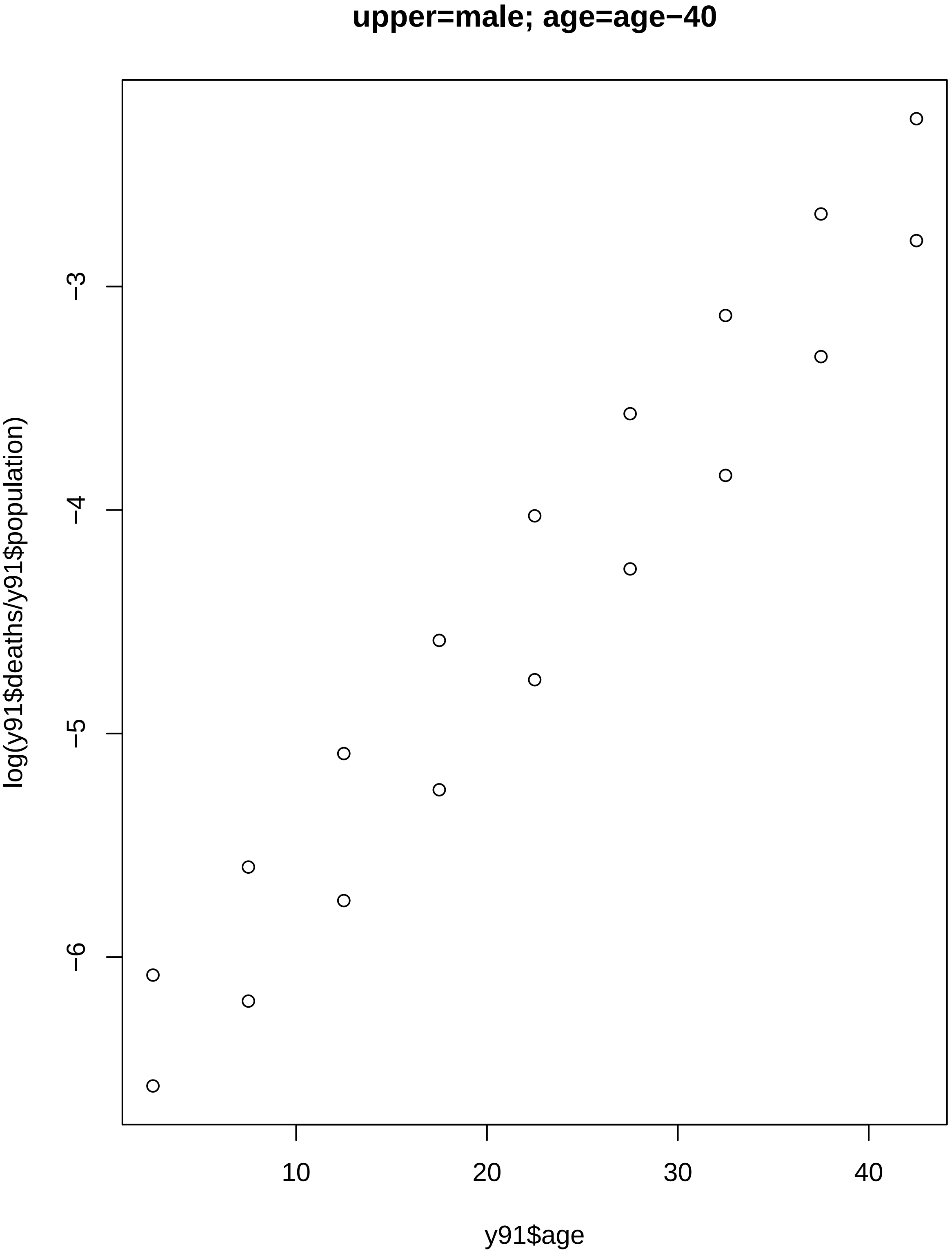
Rates & Rate ratios: multiple regression [Many Rates]

Example: age-specific death rates, male/female 1991, Québec.

```
ds=read.table("quedata.txt",header=T)

ds7191=ds[(ds$age > 40) & (ds$age < 85 ) & (
(ds$year==1971) | (ds$year==1991)),] ;attach(ds7191)
# age and sex specific death rates '91

y91=ds[(ds$age > 40) & (ds$age < 85 ) & (ds$year==1991) ,]

y91$age=y91$age - 40
(y91m$deaths/y91m$population) / (y91f$deaths/y91f$population)
1.64 1.82 1.93 1.95 2.08 2.00 2.04 1.89 1.72
mean 1.90
plot(y91$age,  log( y91$deaths / y91$population ) ) next page
summary( glm(deaths ~ age + male, family=poisson,
               offset=log(population),data=y91) )
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -6.9030313  0.0172228 -400.81   <2e-16 ***
age          0.0956970  0.0004896  195.47   <2e-16 ***
male         0.6506765  0.0106740   60.96   <2e-16 ***
rr = exp[0.0956970] = 1.10   rr= exp[0.6506765] = 1.91
    Null deviance: 47005.299  on 17  df
Residual deviance:    63.095  on 15  df  AIC: 234.73
summary( glm(deaths ~ age + male + age*male,
  family=poisson, offset=log(population),data=y91)
)
Coefficients:
              Estimate Std. Error  z value Pr(>|z|)
(Intercept) -6.9308474  0.0256471 -270.239   <2e-16 ***
age          0.0965921  0.0007816  123.579   <2e-16 ***
male         0.6952540  0.0321127   21.650   <2e-16 ***
age:male    -0.0014773  0.0010029   -1.473    0.141
Residual deviance:    60.923  on 14  df AIC: 234.56
```

Male and female death rates are 'close to proportional'

[see 'multiplicative' model of Clayton & Hills, Table 22.5 Ch 22]


upper=male; age=age−40
log(y91$deaths/y91$population)
−3
−4
−5
−6
10
20
30
40
y91$age

Male circumcision for HIV prevention in men in Rakai, Uganda: a randomised trial: The Lancet Feb 25 2007

Ronald H Gray, Godfrey Kigozi, David Serwadda, Frederick Makumbi, Stephen Watya, Fred Nalugoda, Noah Kiwanuka, Lawrence H Moulton, Mohammad A Chaudhary, Michael Z Chen, Nelson K Sewankambo, Fred Wabwire-Mangen, Melanie C Bacon, Carolyn F M Williams, Pius Opendi, Steven J Reynolds, Oliver Laeyendecker, Thomas C Quinn, Maria J Wawer

Summary

Background Ecological and observational studies suggest that male circumcision reduces the risk of HIV acquisition in men. Our aim was to investigate the effect of male circumcision on HIV incidence in men.

Methods: 4996 uncircumcised, HIV-negative men aged 15–49 years who agreed to HIV testing and counselling were enrolled in this randomised trial in rural Rakai district, Uganda. Men were randomly assigned to receive immediate circumcision (n=2474) or circumcision delayed for 24 months (2522). HIV testing, physical examination, and interviews were repeated at 6, 12, and 24 month follow-up visits. The primary outcome was HIV incidence. Analyses were done on a modified intention-to-treat basis. This trial is registered with ClinicalTrials.gov, with the number NCT00425984.

Findings: Baseline characteristics of the men in the intervention and control groups were much the same at enrolment. Retention rates were much the same in the two groups, with 90–92% of participants retained at all time points. In the modified intention-to-treat analysis, HIV incidence over 24 months was 0·66 cases per 100 person-years in the intervention group and 1·33 cases per 100 person-years in the control group (estimated efficacy of intervention 51%, 95% CI 16–72; p=0·006). The as-treated efficacy was 55% (95% CI 22–75; p=0·002); efficacy from the Kaplan-Meier time-to-HIV-detection as-treated analysis was 60% (30–77; p=0·003). HIV incidence was lower in the intervention group than it was in the control group in all sociodemographic, behavioural, and sexually transmitted disease symptom subgroups. Moderate or severe adverse events occurred in 84 (3·6%) circumcisions; all resolved with treatment. Behaviours were much the same in both groups during follow-up.

Interpretation Male circumcision reduced HIV incidence in men without behavioural disinhibition. Circumcision can be recommended for HIV prevention in men.

From Table 3 of article...

```
> ds=read.table("UgandaTrial.txt",header=T); ds
  t.from t.to intervention participants events     py  rr*
1      0    6            1         2263     14 1172.1 .76
2      0    6            0         2319     19 1206.7
3      6   12            1         2235      5 1190.7 .35
4      6   12            0         2229     14 1176.3
5     12   24            1          964      3  989.7 .25
6     12   24            0          980     12 1008.7
>
> fit= glm(events ~ intervention, family=poisson,
           offset=log(py),data=ds) ;
> summary(fit)

Deviance Residuals:
      1        2        3        4        5        6
 2.0379   0.7256  -1.0784  -0.4140  -1.5346  -0.3849

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)   -4.3224     0.1491 -28.996  < 2e-16 ***
intervention  -0.7040     0.2601  -2.706  0.00681 **

    Null deviance: 16.3106  on 5  df
Residual deviance:  8.5173  on 4  df   AIC: 37.095

> 0.01*round(100*exp(fit$coefficients))

 (Intercept) intervention
       0.01         0.49

observation:            1     2     3     4     5     6

0.1*round(10*fit$fit) 7.7  16.0   7.8  15.6   6.5  13.4
 ds$events           14    19     5    14     3    12

             * Proportional hazards?


Cumulative no. of participants     2387     2430
Cumulative incident events           22       45
Cumulative person-years          3352.4   3391.8  i.r.r.
Cumulative incidence per 100 p-y ??   0.66     1.33  0.49
                                                (0.28 to 0.84)
```