

## Modeling Risk

Assume for fixed  $n_{jk}$ ,  $d_{jk} \sim \text{Poisson}(n_{jk}\lambda_{jk})$ .

The pdf for  $d_{jk}$  is

$$p(d_{jk}) = \frac{e^{-n_{jk}\lambda_{jk}} (n_{jk}\lambda_{jk})^{d_{jk}}}{d_{jk}!}$$

so,

$$\log L = -n_{jk}\lambda_{jk} + d_{jk} \log(n_{jk}\lambda_{jk}) - \log(d_{jk}!)$$

Note that the range of the Poisson distribution is  $0, 1, 2, \dots, \infty$ , but  $d_{jk}$  is bounded by the number of subjects, clearly  $d_{jk}$  cannot be Poisson, however, for small  $n_{jk}\lambda_{jk}$ , the Poisson is a good approximation.

What if  $n_{jk}\lambda_{jk}$  is large?

Suppose hazard is constant (exponential survival), individual  $i$  has hazard  $\lambda$  and is followed for time  $T_i$ .

If death occurs at time  $t_i < T_i$ , then the likelihood is  $\lambda e^{-\lambda t_i}$ .

Otherwise, the probability of surviving to time  $T_i$  is  $e^{-\lambda T_i}$ .

So if  $\delta_i$  is indicator of death, at time  $t_i$  (possibly  $= T_i$ ), the likelihood is  $\lambda^{\delta_i} e^{-\lambda t_i}$  or  $\log L = -\lambda t_i + \delta_i \log \lambda$ .

The full likelihood is

$$\log L = -\lambda \sum t_i + \sum \delta_i \log \lambda$$

which matches the Poisson likelihood up to

$$d_{jk} \log n_{jk} - \log d_{jk}!$$

(which are irrelevant for the purposes of estimation/inference).

The MLE,  $\hat{\lambda} = \sum \delta_i / \sum t_i$ , is the same estimate we get using the Poisson assumption.

Therefore, the Poisson assumption yields the correct likelihood even if the rates  $n_{jk}\lambda_{jk}$  are high, provided that the hazard doesn't change much over the interval.

## Poisson Model as a GLM (1)

A generalized linear model (GLM) is defined by the following exponential family of distributions:

$$f(y_i) = \exp \left\{ \frac{y_i \theta_i - b(\theta_i)}{\phi} - c(y_i, \phi) \right\}$$

For a Poisson model with parameter  $\mu_i$ ,

$$\begin{aligned} f(y_i) &= \frac{\mu_i^{y_i} e^{-\mu_i}}{y_i!} \\ &= \exp(y_i \log \mu_i - \mu_i - \log y_i!) \end{aligned}$$

where  $\theta_i = \log \mu_i$ ,  $b(\theta_i) = e^{\theta_i} = \mu_i$ ,  $c(y_i, \phi) = \log y_i!$ , and  $\phi = 1$ .

$$E(Y_i) = b'(\theta_i) = \mu_i$$

$$\text{Var}(Y_i) = b''(\theta_i)/\phi = \mu_i$$

## Poisson Model as a GLM (2)

A natural link:

$$g(\mu_i) = \log \mu_i = \eta_i = x_i' \beta$$

In terms of the underlying Poisson model,  $\mu_i = \lambda_i n_i$ .

Suppose the following model is desirable:

$$\log \lambda_i = x_i' \beta$$

Then,

$$\log \mu_i = \log \lambda_i + \log n_i = x_i' \beta + \log n_i$$

The last term is known as *offset*, which is a part of the linear predictor with parameter fixed at  $i$  as follows:

$$x_i^* = (\log n_i, x_{1i}, \dots, x_{pi})'$$

and

$$\beta^* = (1, \beta_1, \dots, \beta_p)'$$

## Another Strategy for GLMs

Consider the following distribution:

$$f(y_i) = \exp \left\{ \frac{y_i \theta_i - b(\theta_i)}{a_i(\phi)} - c(y_i, \phi) \right\}$$

where  $a_i(\phi) = \phi/w_i$  and  $w_i$  prior weights.

For a Poisson model,

$$\begin{aligned} f(y_i) &= \exp \left\{ \frac{d_i \log \mu_i - \mu_i}{1} - \log d_i! \right\} \\ &= \exp \left\{ \frac{d_i \log(\lambda_i n_i) - \lambda_i n_i}{1} - \log d_i! \right\} \\ &= \exp \left\{ \frac{(d_i/n_i) \log \lambda_i - \lambda_i}{1/n_i} - (\log d_i! - d_i \log n_i) \right\} \end{aligned}$$

So that  $y_i = d_i/n_i$ ,  $w_i = n_i$ , and  $\log \lambda_i = x_i' \beta$ .

Thus, there is no need to include the offset term in the model.

## Two Models for Risk

- Additive model:  $\lambda_{jk} = \alpha_j + \beta_k$

Additive models may make more sense in certain situations and multiplicative models in others.

$\alpha_j = \lambda_{j1}$ , baseline rate for stratum  $j$

With  $\beta_1 = 0$ , the effect of exposure at level  $k$  is to add a constant amount  $\beta_k$  to the stratum-specific rate  $\lambda_{j1}$

- Multiplicative model:  $\log \lambda_{jk} = \alpha_j + \beta_k$  or  $\lambda_{jk} = \theta_j \psi_k$

Multiplicative models are more commonly used and are easier to fit and interpret.

$\theta_j = e^{\alpha_j}$  represents the baseline rate for stratum  $j$ .

$\psi_k = e^{\beta_k}$  represents the relative risk (risk ratio) of disease for exposure at level  $k$  relative to baseline at level 1 ( $\psi_1 = 1$ ).

- We will not be discussing additive models here.

## Multiplicative Models

Note that, under multiplicative models,

$$\frac{\text{CMF}_2}{\text{CMF}_1} = \frac{\sum_j w_j \lambda_{j2}}{\sum_j w_j \lambda_{j1}} = \frac{\sum_j w_j \theta_j \psi_2}{\sum_j w_j \theta_j \psi_1} = \frac{\psi_2}{\psi_1} = \psi_2$$

$$\frac{\text{SMR}_2}{\text{SMR}_1} = \frac{\sum_j n_{j2} \lambda_{j2} / \sum_j n_{j2} \lambda_j^*}{\sum_j n_{j1} \lambda_{j1} / \sum_j n_{j1} \lambda_j^*} = \psi_2 \frac{\sum_j n_{j2} \theta_j / \sum_j n_{j2} \lambda_j^*}{\sum_j n_{j1} \theta_j / \sum_j n_{j1} \lambda_j^*}$$

Thus, the ratio of CMFs is unbiased, but more variable.

The ratio of SMRs is unbiased if either  $\theta_j / \lambda_j^*$  or  $n_{j2} / n_{j1}$  is constant.

This means either that age-specific rates are both proportional to the external standard rates or that age distributions are identical.

## Estimation (1)

The MLE equates margins to expected margins:

$$d_{jk} = n_{jk} \hat{\lambda}_{jk} = n_{jk} \hat{\theta}_j \hat{\psi}_k.$$

So that

$$D_j = \sum_k d_{jk} = \sum_k n_{jk} \hat{\theta}_j \hat{\psi}_k$$

and

$$O_k = \sum_j d_{jk} = \sum_j n_{jk} \hat{\theta}_j \hat{\psi}_k.$$

Therefore, the MLEs satisfy

$$\hat{\theta}_j = \frac{D_j}{\sum_k n_{jk} \hat{\psi}_k}$$

and

$$\hat{\psi}_k = \frac{O_k}{\sum_j n_{jk} \hat{\theta}_j}.$$

## Estimation (2)

Solutions can be obtained by iteration:

First fix  $\psi_1 = 1$ ,

$$\hat{\psi}_k^{(0)} = 1, \quad \hat{\theta}_j^{(1)} = \frac{D_j}{N_j}$$

$$\hat{\psi}_k^{(1)} = \frac{O_k}{\sum_j n_{jk} \hat{\theta}_j^{(1)}} = \frac{O_k}{\sum_j n_{jk} D_j / N_j}$$

= SMR using aggregate sample as standard

$$\hat{\theta}_j^{(i+1)} = \frac{D_j}{\sum_k n_{jk} \hat{\psi}_k^{(i)}}$$

and

$$\hat{\psi}_k^{(i+1)} = \frac{O_k}{\sum_j n_{jk} \hat{\theta}_j^{(i+1)}}$$

Continue until convergence achieved.

Note:  $\hat{\psi}_k$  can be interpreted as an SMR using the estimated rates  $\hat{\theta}_j$  as the standard rates.

## General Multiplicative Model for Grouped Data

Consider

$$\log \lambda_{jk} = \alpha_j + x_{jk}\beta$$

where  $\alpha_j$  denote nuisance parameters representing strata effect and  $x_{jk}$  the exposure of interest.

Then

$$\frac{\lambda_{jk}}{\lambda_{jk'}} = \exp\{(x_{jk} - x_{jk'})\beta\}$$

depends only on exposure.

Under a Poisson model,

$$\log E[d_{jk}] = \log n_{jk} + \alpha_j + x_{jk}\beta.$$

Note that  $\log n_{jk}$  is a *known* constant – does not need to be estimated, i.e., it is NOT a parameter, nor does it have a parameter associated with it.

It is the offset term.

## Software Implementation

- SAS

```
PROC GENMOD;  
  Model deaths=age exposure  
    /dist=poisson offset=log_year;
```

where `log_year` is a variable whose value for cell  $(j, k)$  is  $\log n_{jk}$ .

- Splus

```
glm(deaths~age+exposure+offset(log.year),  
  family=poisson, ...)
```

## Goodness-of-Fit

Note that the deviance statistics is

$$G^2 = 2 \left\{ \sum_j \sum_k d_{jk} \log \left( \frac{d_{jk}}{\hat{d}_{jk}} \right) + (\hat{d}_{jk} - d_{jk}) \right\}.$$

Since we always have  $\sum_{jk} \hat{d}_{jk} = \sum_{jk} d_{jk}$ , this is the usual

$$G^2 = 2 \sum_{\text{All cells}} \text{Obs} \times \log \left( \frac{\text{Obs}}{\text{Exp}} \right).$$

The statistic

$$\chi^2 = \sum_j \sum_k \frac{(d_{jk} - \hat{d}_{jk})^2}{\hat{d}_{jk}}$$

has a  $\chi^2$  distribution with degrees of freedom equal to the number of non-empty cells minus the number of independent parameters estimated.

## Residuals

One may inspect the residuals

$$r_{jk} = \frac{d_{jk} - \hat{d}_{jk}}{\sqrt{\hat{d}_{jk}}}$$

or more appropriately the *adjusted* residuals

$$\tilde{r}_{jk} = \frac{d_{jk} - \hat{d}_{jk}}{\sqrt{\hat{d}_{jk}(1 - h_{jk})}}$$

where

$$h_{jk} = \hat{d}_{jk} \text{Var}(\hat{\alpha}_j + x_{jk} \hat{\beta})$$

a diagonal element of the *hat* or projection matrix.

These residuals may be regarded roughly as equivalent normal deviates when assessing the fit for any particular cell.

The adjusted residuals account for the number of fitted parameters, and in fact the sum of  $h_{jk}$  equals  $J + p$ , the number of estimated parameters.

## Additive vs Multiplicative Models (1)

Consider

$$\lambda_{jk}^\rho = \alpha_j + x_{jk}\beta$$

or alternatively

$$\frac{\lambda_{jk}^\rho - 1}{\rho} = \alpha_j + x_{jk}\beta.$$

If  $\rho = 1$ , the power model is additive.

As  $\rho \rightarrow 0$ ,  $\frac{\lambda^\rho - 1}{\rho} \rightarrow \log \lambda$ , so it is multiplicative.

Under this model,  $E[d_{jk}] = \lambda_{jk} n_{jk} = (\alpha_j + x_{jk}\beta)^{1/\rho} n_{jk}$  which no longer has form  $g(y) = \alpha_j + x_{jk}\beta + \text{offset}$ .

Fitting this model requires that we redefine our response to be  $d_{jk}/n_{jk}$ , which is no longer Poisson.

Hence, a new family of distributions is required for fitting purposes.

We first note that the likelihood function can be rewritten as

$$\log L = n_{jk} \left( -\lambda_{jk} + \frac{d_{jk}}{n_{jk}} \log(\lambda_{jk}) + C(d_{jk}, n_{jk}) \right).$$

The *quasi* family is a general family of distributions in which one approximates the distribution as a function of the mean  $\mu = E[y]$  and variance  $V(\mu, \phi)$  where  $\phi =$  dispersion only.

In this case,  $E[y_{jk}] = E[d_{jk}/n_{jk}] = \lambda_{jk}$  and  $\text{Var}(y_{jk}) = \lambda_{jk}/n_{jk}$ .

## Additive vs Multiplicative Models (2)

The possible forms for the variance function (at least in Splus) do not allow the use of a variable ( $n_{jk}$ ), so we need to be a bit tricky.

In this case, the parenthesized expression above looks like the likelihood for a poisson with mean  $\lambda_{jk}$ , so we take the mean to be  $\mu = \lambda_{jk}$  and the variance to be  $\mu$  as well.

The Splus function `glm` accepts a `weight` argument which weights the contribution each term makes to the likelihood.

We choose the weight variable to be  $n_{jk}$ .

In Splus we model the variable `rate = deaths/time`,  
`time = person years exposure in each cell`.

```
glm(rate ~ age + exposure, family=  
    quasi(link=power(p),variance=mu), weight=time,...)
```

To choose the best value of  $\rho$ , we may plot  $G^2$  versus  $\rho$  for a given model.

In SAS

```
PROC GENMOD ;  
Model . . . / dist = poisson link = power(p) ... ;
```

where `p` is the desired value of  $\rho$ .

## Modeling using External Rates

Suppose that  $\theta_j = \theta \lambda_j^*$  under which a ratio of SMRs becomes unbiased.

Then  $\log \lambda_{jk} = \alpha_j^* + \gamma + x_{jk}\beta$  where  $\alpha_j^* = \log \lambda_j^*$  is known and  $\gamma = \log \theta$ , so that

$$\begin{aligned}\log E[d_{jk}] &= \log(\lambda_{jk} n_{jk}) \\ &= \alpha_j^* + \log n_{jk} + \gamma + x_{jk}\beta \\ &= \log(n_{jk} \lambda_j^*) + \gamma + x_{jk}\beta\end{aligned}$$

with the offset  $\log(n_{jk} \lambda_j^*)$ .

The intercept term  $\gamma$  measures departures of rates in baseline exposure group relative to external standard, i.e.,

$$\gamma = \log \frac{O_+}{E_+^*} = \log \text{SMR}$$

and

$$\hat{\psi}_k = \text{SMR}_k = \frac{O_k}{E_k^*} = \exp(\hat{\beta}_k).$$

Note that this model has fewer parameters,  $\gamma$  and  $\beta$ , rather than  $\alpha_1, \dots, \alpha_J$  and  $\beta$ , because we are not required to model the underlying shape of the hazard.

On the other hand, we rely on the proportionality assumption.

Typically the number of parameters in an internally standardized model can be minimized by fitting a parametric form to the  $\alpha_j$ .

## Relative Risk Models (1)

Relative risk or risk ratio

$$RR(x) = \frac{\lambda_j(x)}{\lambda_j(0)}$$

- Multiplicative models

$$\lambda_j(x) = \lambda_j(0) \exp(x\beta)$$

$$RR(x) = \exp(x\beta).$$

- Additive models

$$\lambda_j(x) = \alpha_j + x\beta = \alpha_j(1 + x\beta^*)$$

$$RR(x) = 1 + x\beta^* \text{ where } \beta^* = \beta/\alpha_j$$

Alternatively

$$\lambda_j(x) = \exp(\alpha_j)(1 + x\beta)$$

$$RR(x) = 1 + x\beta$$

## Relative Risk Models (2)

More generally

$$\lambda_j(x) = \exp(\alpha_j)r(x; \beta)$$

where

$$\lambda_j(0) = \exp(\alpha_j),$$

and so that

$$RR(x) = r(x; \beta).$$

$$\log r(x; \beta) = \begin{cases} x\beta & \rho = 0 \\ \frac{(1+x\beta)^\rho - 1}{\rho} & \rho \neq 0 \\ \log(1 + x\beta) & \rho = 1 \end{cases}$$

$$\log r(x; \beta) = \frac{\log(1 + \rho x\beta)}{\rho}, \quad 0 \leq \rho \leq 1$$

$$r(x; \beta) = \{\exp(x\beta)\}^\rho (1 + x\beta)^{1-\rho}, \quad 0 \leq \rho \leq 1$$

$$r(x; \beta) = (x_0 + x)^\beta \text{ or } (1 + x)^\beta$$

or

$$\log r(x; \beta) = \beta \log(x_0 + x) = \beta x^*$$

where  $x^* = \log(x_0 + x)$  is *transformed* dose and  $x_0$  is background exposure accounting for the spontaneous incidence.

## Relative vs Excess Risk Models

- Relative risk models

$$\lambda_{jk} = \lambda_j^* \exp(\alpha + x_{jk}\beta)$$

$$RR(x_{jk}) = \exp(\alpha + x_{jk}\beta)$$

$$\begin{aligned} \log \mu_{jk} &= \log \lambda_j^* n_{jk} + \alpha + x_{jk}\beta \\ &= \log E_{jk} + \alpha + x_{jk}\beta \end{aligned}$$

- Excess risk models

$$\lambda_{jk} = \lambda_j^* + \exp(\alpha + x_{jk}\beta)$$

$$\begin{aligned} \mu_{jk} &= \lambda_j^* n_{jk} + n_{jk} \exp(\alpha + x_{jk}\beta) \\ &= E_{jk} + n_{jk} \exp(\alpha + x_{jk}\beta) \end{aligned}$$

Excess mortality ratio (EMR)

$$EMR_k = \frac{\text{Excess risk at } x_{jk}}{\text{Excess risk at 0}} = \exp(x_{jk}\beta).$$

## Regression Analysis using Standard Rates

$\lambda_{jk}$ , disease rate in stratum  $j$  and exposure level  $k$

$\lambda_j^*$ , known disease rate in standard population

$x_{jk}$ , exposure/covariates

- Multiplicative models (SMR)

$$\lambda_{jk} = \lambda_j^* \theta_{jk} = \lambda_j^* \exp(\alpha + x_{jk} \beta)$$

$e^\alpha =$  baseline SMR

$\beta = \log RR$

- Additive models (EMR)

$$\lambda_{jk} = \lambda_j^* + \theta_{jk} = \lambda_j^* + \exp(\alpha + x_{jk} \beta)$$

$e^\alpha =$  baseline excess risk

$e^\beta =$  EMR

## SMR Regression

$d_{jk}$ , number of deaths in stratum  $j$  and at exposure level  $k$

$n_{jk}$ , person-years observation time

$e_{jk} = \lambda_j^* n_{jk}$ , expected number of deaths

$$E[d_{jk}] = \lambda_{jk} n_{jk} = \exp(\log e_{jk} + \alpha + x_{jk} \beta)$$

$\log e_{jk}$ , offset

## EMR Regression

$$E[d_{jk}] = e_{jk} + n_{jk} \exp(\alpha + x_{jk}\beta)$$

$$\mu = e + n \exp(\eta)$$

$$\eta = \log(\mu - e) - \log n$$

$$\frac{d\eta}{d\mu} = \frac{1}{\mu - e}$$

$\mu$ , fitted value

$\eta$ , linear predictor