Multilevel Models

9. Models for Count and Survival Data

Germán Rodríguez

Princeton University

April 23, 2018

Poisson Models

This unit concerns models for count data. We assume that conditional on unobserved random effects the outcomes have a Poisson distribution.

For example in a two-level random intercept model we write

$$Y_{ij}|a_i \sim P(\mu_{ij}) \quad \text{where} \quad \log \mu_{ij} = (\alpha + a_i) + x'_{ij}\beta$$

We will assume that $a_{ij} \sim N(0, \sigma_a^2)$ as we have done for other models. This choice generalizes to more general random-coefficient models but requires quadrature. Stata uses adaptive quadrature in `xtpoisson` and `mepoisson` and R's `glmer()` uses quadrature for one random effect and PQL otherwise.

An alternative with Poisson models is to use a gamma-distributed multiplicative random effect, which can be integrated analytically, but doesn't generalize to correlated random effects. Stata's `xtpoisson` implements gamma as an option.

A Random-Intercept Poisson Model

Our first application is to small area estimation using data on lip cancer from Scotland. The data consist of the number of cases observed in each of 56 counties in 1975-80, and are available at http://www.stata-press.com/data/mlmus3/lips.dta.

We also have information on the expected number of cases based on age-specific lip cancer rates for the whole of Scotland and the age distribution in each county. The ratio of observed to expected counts, usually times 100, is called the Standardized Mortality Ratio (SMR). For example a value of 193.2 denotes almost twice as many cases as expected.

A limitation of crude SMRs is that estimates for counties with small populations are very imprecise. To address this problem we will use Empirical Bayes (EB) estimates based on a random-intercept Poisson model. By adding a random effect at level one we are effectively modeling over-dispersion.

Fitting the Random-Intercept Model

In this model the conditional distribution of the count is Poisson with mean proportional to the expected number of cases

$$Yi|a_i \sim P(\mu_i) \quad \text{with} \quad \log \mu_i = \alpha + a_i + log(e_i) \quad \text{and} \quad a_i \sim N(0, \sigma^2)$$

MALMUS fits the model using `gllammm` (page 724). Using Stata's `mepoisson` we get the same results using the comand

```
mepoisson o, offset(lne) || county:
```

Note that the offset has to be specified as an option in the fixed part of the model. The model can also be fit using R as shown in the computing logs.

Using mean-variance adaptive Gauss-Hermite quadrature with 12 points we get $\hat{\alpha} = 0.0803$ and $\hat{\sigma}^2 = 0.5847$.

The average SMR in this model is 145, obtained by noting that

$$E(Y_i/e_i) = exp(\alpha + a_i) \quad \text{and} \quad E(exp(a_i)) = exp(\sigma^2/2)$$

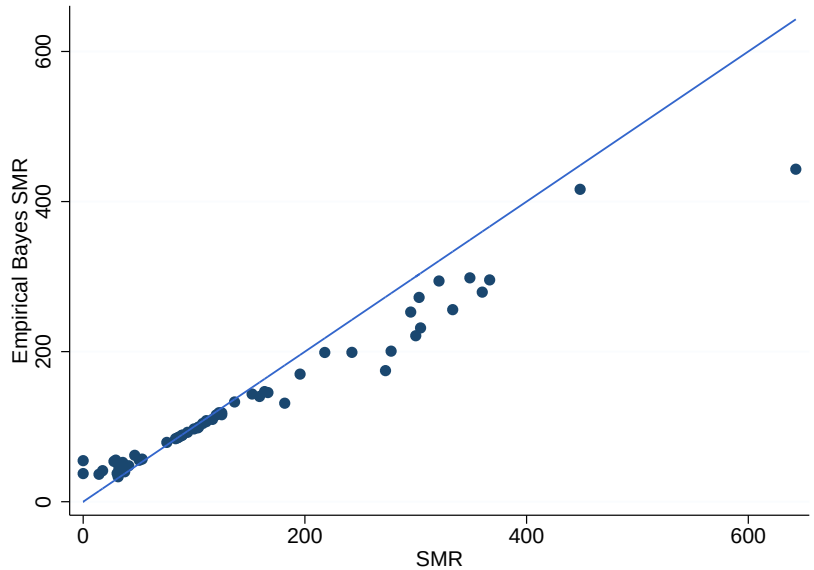
There is, however, substantial variation across counties.

Prediction of SMRs by County

We now consider predicting the SMR in each county using EB posterior means or modes. Stata's `mepoisson` uses means, but has an option for modes; `gllamm` uses means, and R uses modes.

We first predict the random effects using `predict a, reffects` or `ranef()` to obtain $\hat{a}_i$ for each county, and then add the constant but leave out the offset, computing the predicted SMR as $100 \exp\{\hat{\alpha} + \hat{a}_i\}$.

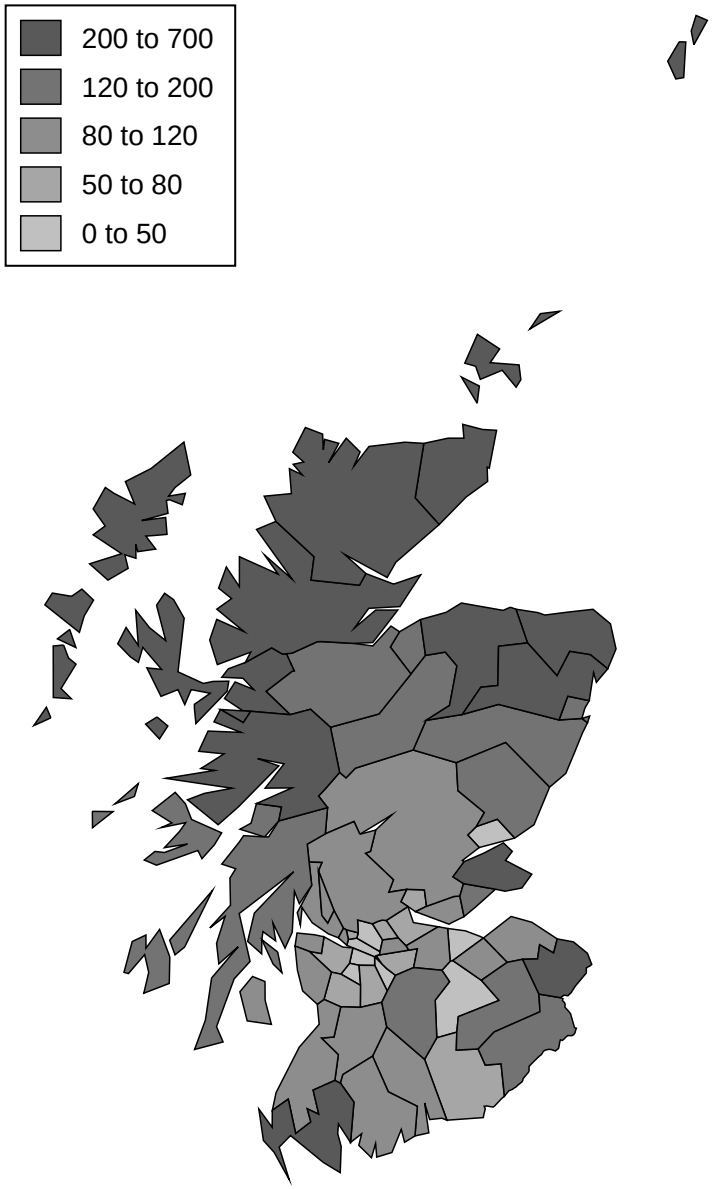
The figure on the right shows the EB estimates plotted against the crude SMRs and exhibits the usual shrinkage towards the overall mean, see MALMUS figure 13.3.



A Choropleth Map

The map on the right shows the counties of Scotland with shading representing the EB estimate of the SMR, reproducing MALMUS Figure 13.2.

The computing log shows how to reproduce this graph using Stata code available from Stata press or our own R code. The incidence of lip cancer is higher in coastal places, particularly in the north.



Health-Care Reform in Germany

MALMUS examines the extent to which the health-care reform in Germany reduced the number of doctor visits, using panel data for women working full time before and after the reform.

Here is a comparison of effect estimates from three models, all including controls for age, education, married, bad-health, log-income and summer

Model	Poisson	R-Intercept	R-Slope
Reform	0.8690	0.9547	0.9023
σ_a	-	0.9051	0.9541
σ_b	-	-	0.9303

The random-intercept model shows substantial unobserved heterogeneity in doctor visits among women with the same observed attributes; a one std dev increase in "frailty" results in 2.5 times as many visits.

The random-slope model allows the effect of the reform to vary across women. The effect for the average woman is now a 10% reduction, but varies substantially across women. The correlation between intercept and slope is -0.491.

Infant and Child Mortality in Kenya

An important application of Poisson models is to multilevel survival analysis via the connection with piecewise exponential survival.

I illustrate this approach with an analysis of infant and child mortality using the Kenya DHS, with an abridged version in "Multilevel Models in Demography" and full details in my chapter of the *Handbook of Multilevel Analysis*.

Let $\lambda_{ijk}(t)$ denote the hazard at age t for the i-th child of the j-th mother in the k-th community. We consider a three-level model

$$\lambda(t|x_{ijk}, a_{jk}, a_k) = \lambda_0(t) \exp\{x'_{ijk}\beta + a_{jk} + a_k\}$$

where $\lambda_0(t)$ is the baseline hazard, β is a vector of fixed parameters representing effects of observed covariates, and $a_{jk} \sim N(0, \sigma_2^2)$ and $a_k \sim N(0, \sigma_3^2)$ are random effects representing unobserved family and community frailty.

Estimation Using Poisson Regression

We assume the hazard is constant in intervals with cutpoints τ_d. After some exploratory work I chose cutpoints 0,1,6,12,24 and 60 months. I then split each observation into one episode per interval visited, and count events and exposure, obtaining 48,094 episodes.

Predictors include one variable at the community level (urban or rural), one at the mother level (years of education) and five at the level of the child, all well-known risk factors (gender, cohort, age of mother, birth order, and length of the previous birth interval). I'll show how these are represented when I display the coefficients.

To fit the piecewise exponential model we treat the death indicator as Poisson with the log of exposure time as an offset. Estimation using mean-variance adaptive Gaussian quadrature is implemented in Stata's `mepoisson`. (Unfortunately R's `glmer` in the `lme4` package uses PQL for three-level models. Fortunately there is a good interface to Stan for Bayesian estimation.)

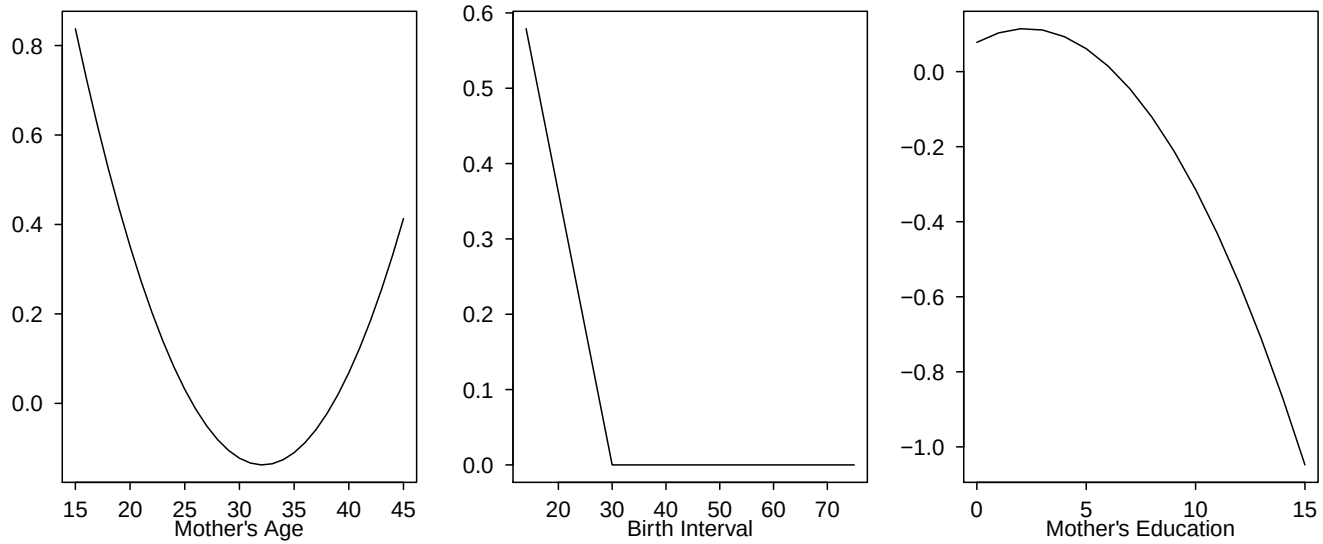
Parameter Estimates

Variable	Term	Coefficient	Standard Error	Hazard Ratio
		Fixed Coefficients		
Constant	1	-4.588	0.118	-
Age	1–5	-1.642	0.089	0.194
	6–11	-1.998	0.097	0.136
	12–23	-2.822	0.106	0.059
	24–59	-3.362	0.109	0.026
Sex	male	0.087	0.068	1.091
Cohort	1993+	0.173	0.069	1.189
Mother's	$a - 25$	-0.047	0.011	0.954
age	$(a - 25)^2$	0.003	0.001	1.003
Birth	$o - 3$	0.043	0.039	1.044
order	$(o - 3)^2$	0.004	0.005	1.004
Interval	$(30 - i)_+$	0.036	0.006	1.037
Mother's	$e - 7$	-0.068	0.015	0.934
education	$(e - 7)^2$	-0.007	0.003	0.993
Residence	urban	0.040	0.142	1.041
		Variance Parameters		
Family	σ_2	0.613	0.086	1.846
Community	σ_3	0.680	0.055	1.973

Hazard Ratios

The fixed coefficients can be interpreted in the usual fashion. Children born after 1993 have 19% *higher* risk that those born earlier, after adjusting for all other factors.

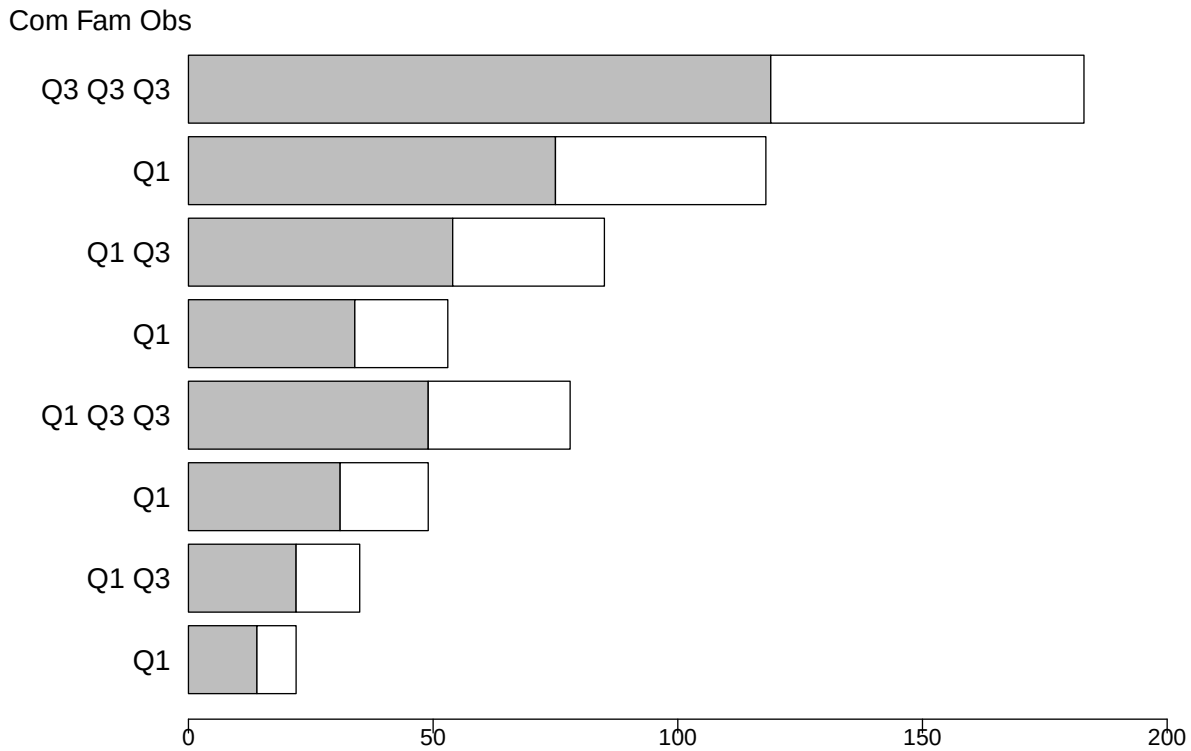
For variables represented using a quadratic or a spline a graph is always helpful:



The most remarkable feature of the results, however, is the extent to which we have unobserved heterogeneity at the family and community level.

Predicted Probabilities

A nice way to present results is to compute conditional and marginal probabilities of death by age one and five.
Here are estimated conditional (or subject-specific) probabilities for quartiles 1 and 3 of observed and unobserved risk:



The marginal (or population-average) probabilities can be obtained using Gauss-Hermite quadrature.