Binomial geostatistical models

Dr Emanuele Giorgi
Lancaster University
e.giorgi@lancaster.ac.uk

Toowoomba 21-25 October 2019

Binomial geostatistical models

- Binomial sampling, extra-binomial variation.
- Binomial generalized linear model with spatially correlated and uncorrelated random effects.
- Maximum likelihood estimation.
- Plug-in and Bayesian prediction.

Binomial sampling and extra-binomial variation

- Bernoulli trial: binary random variable, $Y = 1/0$ with probabilities p and $1 - p$, respectively.

- binomial distribution: discrete random variable , $Y \sim \mathrm{Bin}(m, p)$ is the sum of m independent Bernoulli trials

  Mean and variance:

  $$\mu = \mathrm{E}[Y] = p, \quad \mathrm{Var}(Y) = mp(1 - p) = \mu(1 - \mu/m)$$

- extra-binomial variation: discrete random variable Y is the sum of m binary outcomes, $\mathrm{E}[Y] = \mu$, $\mathrm{Var}(Y) > \mu(1 - \mu/m)$

Blue: how many boys are there in a family of twins?

How does extra-binomial variation arise?

Binary random variables $X_i : i = 1, ..., m$,

$$Y = \sum_{i=1}^{m} X_i$$

Heterogeneity: X_i are mutually independent, with $\mathrm{P}(X_i = 1) = p_i$

Dependence: $\mathrm{P}(X_i = 1) = p$ for all i, but $\mathrm{Corr}(X_i, X_j) = \rho > 0$

Exercise: In either case, prove that $\mathrm{Var}(Y) > \mathrm{E}[Y](1 - \mathrm{E}[Y]/m)$

Binomial logistic geostatistical model

- Latent spatial process

$$S(x) \sim \mathrm{SGP}\{0, \sigma^2, \rho(u))\}$$

$$\rho(u) = \exp(-|u|/\phi)$$

- Linear predictor

$$d(x) = \text{environmental variables at location } x$$

$$\eta(x) = d(x)'\beta + S(x)$$

$$\eta(x) = \log[p(x)/\{1 - p(x)\}]$$

- Error distribution

$$Y_i|S(\cdot) \sim rmBin\{m_i, p(x_i)\}$$

Parameter estimation

For convenience, write

$$T_i = d(x_i)'\beta + S(x_i) \quad T = (T_1, ..., T_n)$$

Then, all model parameters, θ say, are contained within $[T]$.

Data: $(m_i, y_i) : i = 1, ..., n$

Model:

$$T = \mathrm{MVN}(\theta) \quad [Y|T] = \prod_{i=1}^{n}[Y_i|T_i] \quad [Y_i|T_i] = \mathrm{Bin}(m_i, p_i),\ p_i = p(T_i)$$

Likelihood:

$$L(\theta) = \int_T [T;\theta] \times [Y|T]dT$$

Problem ... how to evaluate the high-dimensional integral?

Monte Carlo maximum likelihood: a general method for parameter estimation with intractable likelihoods

Scenario

$L(\theta) = f(y;\theta) = c(\theta)g(y;\theta)$, with $g(\cdot)$ known but $c(\cdot)$ intractable.

Method

1. Note first that $c(\theta)^{-1} = \int g(y;\theta)dy$

2. Choose any fixed value θ_0, and let $E_0[\cdot]$ mean "expectation when $\theta = \theta_0$" Then,

$$E_0\left[\frac{g(y;\theta)}{g(y;\theta_0)}\right] = \int \frac{g(y;\theta)}{g(y;\theta_0)}c(\theta_0)g(y;\theta_0)dy = \frac{c(\theta_0)}{c(\theta)}$$

Monte Carlo maximum likelihood: continued

3. Write the likelihood ratio wrt θ_0 as

$$LR(\theta) = \frac{f(y;\theta)}{f(y;\theta_0)} = \frac{c(\theta)g(y;\theta)}{c(\theta_0)g(y;\theta_0)} = \frac{g(y;\theta)}{g(y;\theta_0)} / E_0\left[\frac{g(y;\theta)}{g(y;\theta_0)}\right]$$

4. Write $r(y;\theta) = g(y;\theta)/g(y;\theta_0)$
5. Simulate $y_1, y_2, ..., y_s$ from the distribution of Y when $\theta = \theta_0$.
6. For any value of θ compute $\bar{r}(\theta) = \sum_{k=1}^{s} r(y_k;\theta)$
7. Compute $\hat{\theta}$ to maximise

$$MClogL(\theta) = \log g(y;\theta) - \log \bar{r}(\theta)$$

Monte Carlo maximum likelihood for the binomial geostatistical model

$$
\begin{aligned}
L(\theta) &= \int_T [T;\theta] \times [Y|T]\, dT \\
&= \int_T [T;\theta] \times \frac{[T;\theta_0]}{[T;\theta_0]} \times [Y|T]\, dT \\
&= \int \frac{[T;\theta]}{[T;\theta_0]} \times [Y,T;\theta_0]\, dT \\
&= [Y;\theta_0] \int \frac{[T;\theta]}{[T;\theta_0]} \times [T|Y;\theta_0]\, dT \\
&= [Y;\theta_0] \times E_0\left[\frac{[T;\theta]}{[T;\theta_0]}\right],
\end{aligned}
$$

where $E_0[\cdot]$ denotes expectation wrt the conditional distribution of T given $Y = y$ when $\theta = \theta_0$.

Now replace $E_0[\cdot]$ by sample mean over simulations of Y at $\theta = \theta_0$, as before.

Sampling from $[T|Y]$: Markov chain Monte Carlo

A Markov chain is a sequence of random variables $Y_1, Y_2, ..., Y_n, ...$ with the property that, for all n,

$$[Y_n|Y_1, ..., Y_{n-1}] = [Y_n|Y_{n-1}]$$

Markov chain Monte Carlo (MCMC) methods are now widely used in applied statistics approximate evaluation of intractable expectations.

Metroplis MCMC methods are useful for problems involving distributions that are known up to a constant of proportionality.

Gamerman and Lopez (2010)

Metropolis MCMC: to simulate a sample from a distribution with pdf $f(y) = c \times g(y)$

Let $r(y, y^*) = f(y)/f(y^*) = g(y)/g(y^*)$... known for all y and y^*

A proposal distribution, $p(y, y^*)$, is any conditional distribution for Y^* given Y, with the property that $[Y^*|Y] = [Y|Y^*]$

Metropolis update: $Y_n \rightarrow Y_{n+1}$

1. sample a candidate X from any proposal distribution $p(Y_n, X)$
2. calculate the acceptance probability

$$a = \min\left\{1, \frac{p(Y_n, X)}{p(X, Y_n)}\right\}$$

3. sample U from the uniform distribution on $(0, 1)$
4. if $U \leq a$, set $Y_{n+1} = X$, otherwise set $Y_{n+1} = Y_n$

Theorem: in the limit $n \rightarrow \infty$, for any Y_1 and any proposal distribution $p(\cdot)$, the distribution of a sequence of Metropolis updates converges to the distribution with pdf $f(y)$

Prediction in the binomial model

Recall that S denotes the underlying signal at the data-locations, while S^* denotes the signal at all prediction locations of interest, typically a finely spaced grid to cover the region of interest, A

Plug-in prediction

- Estimate $\hat{\theta}$ by Monte Carlo maximum likelihood
- Sample from $[S|Y;\hat{\theta}]$ by MCMC
- Sample from multivariate Normal distribution $[S^*|S]$

Bayesian prediction

- Extend hierarchical representation of model to

$$[Y, S, \theta] = [\theta] \times [S|\theta] \times [Y|S, \theta]$$

- Sample from $[S, \theta|Y]$ by MCMC
- Sample from multivariate Normal distribution $[S^*|S]$

9. Prevalence mapping

- Mapping prevalence: exceedance probability maps.
- Extensions: combining data form multiple surveys, zero- inflation, spatio-temporal models.

A non-spatial model for prevalence survey data

Design

- Sample communities $i = 1, ..., n$.
- In community i, sample m_i individuals of whom Y_i test positive for disease of interest.
- Associated covariates w_i

Model

- $p_i =$ probability that a randomly sampled individual
  in community i will test positive
- $\log\{p_i/(1 - p_i)\} = \alpha + w_i'\beta$
- $Y_i \sim \mathrm{Binomial}(m_i, p_i)$, mutually independent

A spatial model for prevalence survey data

Design

- Sample communities $i = 1, ..., n$ at locations x_i
- In community i, sample m_i individuals of whom Y_i test positive for disease of interest.
- Associated covariates $w_i = w(x_i)$

Model

- $\rho_i =$ probability that a randomly sampled individual in community i will test positive
- $\log\{P_i/(1 - P_i)\} = \alpha + w(x_i)'\beta + S(x_i)$
- $Y_i \sim \mathrm{Binomial}(m_i, P_i)$, conditionally independent given $S(\cdot)$

A spatial model for prevalence survey data (continued)

Two kinds of covariates

- $w(x_i)$ an intrinsic property of the location x_i
- $w(x_i)$ a property of the people who live at location x_i

Practical implication: when mapping prevalence we need to be able to assign a value $w(x)$ to every location in the study-region.

What is $S(x)$?

- an unobserved spatially varying stochastic process
- a proxy for unmeasured, spatially structured covariates

Practical implication: in any application where $S(x)$ turns out to be important, it is worth asking what the missing covariate(s) might be.

Person or place?

Extend spatial model to

$$\log\{P_i/(1-P_i)\} = \alpha + \{w(x_i)'\beta + S(x_i)\} + \{d_i'\gamma + U_i\}$$

- $w(x)$: measured properties of location x
- $S(x)$: stochastic process, proxy for unmeasured properties of x
- d_i : measured properties of ith community
- U_i : independent random variables, proxy for unmeasured properties of ith community

Exploratory analysis: empirical logits

- fitting the binomial logistic model is computationally demanding, and requires judgement:
  - convergence of iterative algorithms
  - judicious choice of approximations
- empirical logit transform:

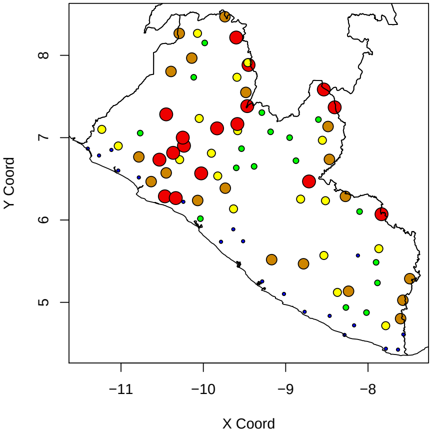
$$Z_i = \log\{(Y_i + 0.5)/(n_i - Y_i + 0.5)\}$$

- fit linear model with Z_i as response

Onchocerciasis (river blindness) in Liberia

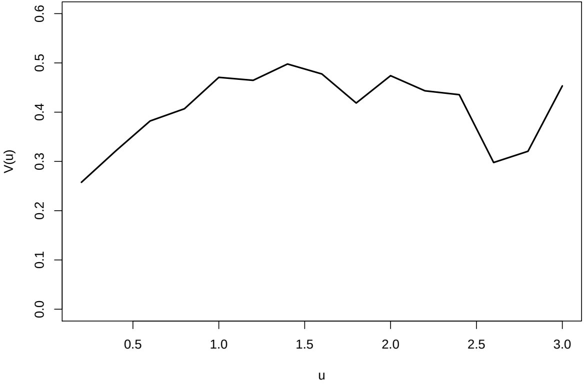
- prevalence data from 90 villages in Liberia
- sample sizes 40 to 50
- empirical prevalences 0% to 35%
- use empirical logit transformation for exploratory analysis

Exploratory analysis of onchocerciasis data



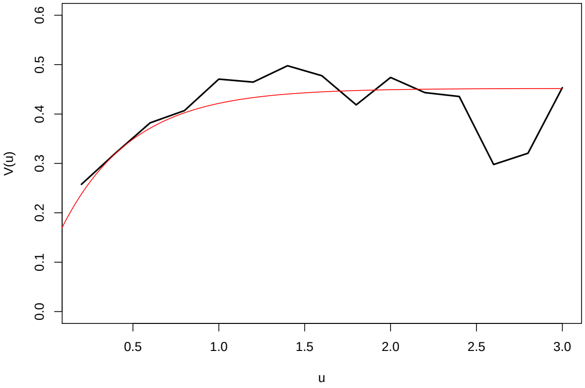
- patches of high and low prevalence
- increasing trend away from coast?

Exploratory analysis of onchocerciasis data (2)



- residual variogram after fitting linear trend surface

Exploratory analysis of onchocerciasis data (3)



- fitted Matérn model with $\kappa = 0.5$

Fitting the binomial logistic model

- likelihood function involves intractable high-dimensional integral
- need to use Monte Carlo methods
- Monte Carlo maximum likelihood or Bayesian estimation according to choice
- for large data-sets, algorithms need careful tuning to preserve accuracy while remaining computationally feasible