Spatio–temporal methods in environmental epidemiology

Lecture 23

SPECIAL TOPICS: MEASUREMENT ERROR

Effects and coping techniques

Measurement error

Space- time modelling can mitigate unpredictable, pernicious effects in environmental epidemiology. This lectures reviews error types, effects & modelling

Error types

- missing data
- classical & Berkson
  - **Classical:** x = "true" value measured with error to get X:

    $$X = x + \epsilon$$

    x, ϵ **uncorrelated**
  - **Berkson:** e.g. experimentor sets "control" at level X but output is (unmeasured) x satisfying

    $$x = X + \epsilon$$

    equivalently

    $$X = x + \epsilon$$

    (assuming symmetry of ϵ's distribution) - now x and ϵ are **correlated**!!

- non-differential & differential
  - **non-differential** if conditional on true value x, health outcome Y independent of measured predictor X - otherwise **differential**
- structural & functional
  - **structural** means x is random, otherwise **functional**
- misclassification - error in a binary response

Error effects

- effects of binary exposure variables, x: reduction in apparent effect if *non-differential*
- same with linear regression & continuous exposures (classical error model) but not with *Berkson*
- generally effects vary, hard to predict - best: reduce measurement error by good design
- for nonlinear models effects more subtle

Error effects: nonlinear models

Suppose

- $(Y, x, X) \sim$ normal
- $E[Y \mid x, X] = \exp[\beta x]$

. Thus

- $E[Y \mid X] = E[\exp[\beta x] \mid X] = \exp[\beta\beta_{xX}X + \beta^2\sigma_{x\cdot X}/2]$ if Y,X independent given x

Residual variance $\sigma_{x\cdot X}$ = precision of X.

- if 0 fit $Y = \exp bX$ bias-correct $\hat{\beta} = b/\beta_{xX}$ like linear case
- if $\neq 0$ bias wants to inflate b, imprecision to deflate b

(Large residual variance puts fitted β close to 0.)

Role of spatial prediction

- Basic building blocks: uncorrelated clusters *i*

Clusters	**Data**
subjects eg mice	repeated measures eg tumors
hospitals	daily admission counts
Census Subdivisions (CSD)	auto-correlated daily death counts
years	spatially correlated CSD school absences

- health outcomes (eg deaths) $\{Y_{it}\}$, for timepoint t (eg day), & cluster i
- pollution concentration (& covariate) vectors $\{X_{it} = (X_{it1}, \ldots, X_{itk})\}$ may be hi-pass filtered to unmask blip effects
- effects model

$$E[Y_{it} \mid X_{it}, \mathbf{a_i}] = m_{it} \exp(\mathbf{a}_i^T X_{it})$$

  - m_{it} a fixed factor accounting for population size, day of week & low frequency seasonal components

Effects significant?

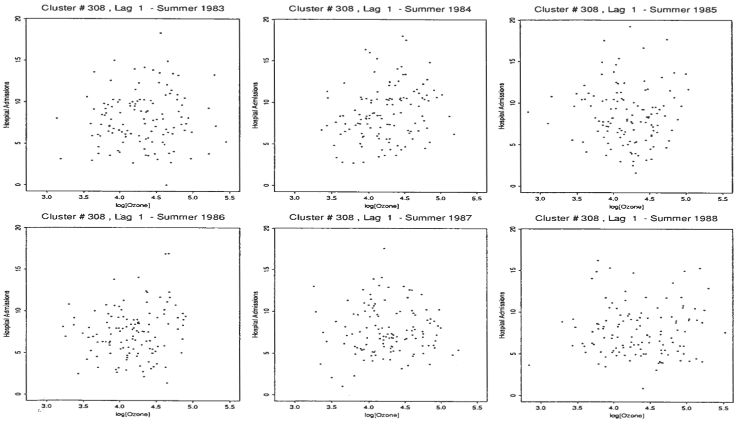
Central issue: Is $a_{kj} = 0$ for pollutant j? for some specific k? All k?

NOTES:

- Effects subtle not significant for specific k
- X_{kt} unmeasured for many k-**need spatial prediction!!**

Subtle effects

Not apparent: strong association between ozone and admissions in plots of daily log O_3 vs hospital admissions in this census subdivision of Toronto, 1983 (upper left)-1988



Borrowing strength

Small insignificant effects for each cluster i can be significant in the aggregate if pattern consistent.

$\boxed{\textbf{How to aggregate?}}$

- Use random effects model:

$$\mathbf{a}_i = \beta + \mathbf{b}_i, \ \mathbf{b}_i^{vector} \sim N(0, D)$$

- Here $\mathbf{b}_i$ is a random effects vector, the deviations for cluster i from the population levels, β

Ambient monitoring

Not all monitoring sites measure the same pollutants.

- networks set up for different purposes often amalgamated
- some purposes not foreseen & hence no gauges originally attached (eg importance of $PM_{2.5}$ only recently recognized)

Ambient pollution levels are unmonitored over many large urban areas
Hence:

$\boxed{\textbf{ambient levels can be a poor surrogate for exposure}}$

$\Rightarrow$ need for spatial predictive methodology

Using spatial prediction

Assumptions:

- $E[Y_{it} \mid X_{it}, \mathbf{a}_i] = \zeta(\mathbf{a}_i^T X_{it})$
- $Var[Y_{it} \mid X_{it}, \mathbf{a}_i] = \phi\zeta(\mathbf{a}_i^T X_{it})$
    - ϕ = overdispersion parameter
    -
- Y_{it_1}, Y_{it_2} $t_1 \neq t_2$ independent conditional on $\mathbf{a}_i$,$X_{it_1} X_{it_2}$ a (“working assumption”)

GEE approach

Only the spatial predictive distribution's mean and variance are needed in this approach[1]!!

- By a similar linearization of the nonlinear mean and variance functions of $\{a_k\}$ additional simplification is gained. Now only their means and variances are needed!
- Assuming $\{Y_{kt}\}$ are (quasi) normal (**GEE approach!!**) enables estimation of the β, $\{b_k\}$ etc.

[1] Liang and Zeger [1986]

Covariance approximation[2]

$$
\begin{aligned}
Cov\left(Y_{it_1}, Y_{it_2} \mid \mathbf{a_i}\right) &\approx \Lambda_{it_1t_2}(\mathbf{a}_i) \text{ where} \\
\Lambda_{it_1t_2}(\mathbf{a}_i) &= \delta_{it_1t_2}\phi E\left(Y_{it_1} \mid \mathbf{a}_i\right) \\
&\quad + \zeta'\left(\mathbf{a}_i^T \mathbf{z}_{it_1}\right)\zeta'\left(\mathbf{a}_i^T \mathbf{z}_{it_2}\right)\mathbf{a}_i^T \mathbf{G}_{it_1t_2}\mathbf{a}_i.
\end{aligned}
$$

[2] Zidek et al. [1996]

Lindstrom-Bates Approximation

Suppose $\mathbf{a}_i \simeq \mathbf{a}_i^{o\ fixed} \equiv \beta_i^o$. Then

$$E(Y_{it}|\mathbf{a}_i) \simeq \zeta(\mathbf{a}_i^{oT} z_{it}) + \hat{Z}_{it}(\mathbf{a}_i - \mathbf{a}_i^o) + \cdots \equiv \eta(\mathbf{a}_i)$$

where[3]

- $\hat{Z}_{it} = \zeta'(\mathbf{a}_i^{oT} z_{it}) z_{it}^T$

Further with $E(\mathbf{a}_i) = \beta$ & $\mathrm{Cov}(\mathbf{a}_i) = \mathrm{D}$,

$$E(Y_{it}) \simeq \mu(\mathbf{a}_i^o) \equiv \zeta(\mathbf{a}_i^{oT} z_{it}) + \hat{Z}_{it}(\beta - \mathbf{a}_i^o)$$

$$+\frac{1}{2}\zeta''(\mathbf{a}_i^{oT} z_{it}) \left\{ z_{it}^T [D + (\beta - \mathbf{a}_i^o)(\beta - \mathbf{a}_i^o)^T] z_{it} \right\}$$

Similar approximations for conditional & unconditional covariances.

[3] Lindstrom and Bates [1990]

Random effect estimates

Assume conditional on $\{\mathbf{a}_i\}$, $\{Y_i\}$(vector form) normally distributed. Given $Y_i = y_i$ (observed for all i) & β, log posterior density of the $\mathbf{b}_i$ is

$$\propto -\frac{1}{2}\mathbf{r}_i^T \Lambda_i^{-1} \mathbf{r}_i - \frac{1}{2}\mathbf{b}_i^T D^{-1} \mathbf{b}_i$$

with ($\mathbf{r}_i \equiv y_i - \eta_i(\beta + \mathbf{b}_i)$). Its mode solves

$$\mathcal{W}_i \equiv \hat{Z}_i^T \Lambda_i^{-1} \mathbf{r}_i + D^{-1} \mathbf{b}_i = 0$$

to yield $\hat{\mathbf{b}}_i$. [Λ fixed at previous iteration & η linear]

Computing random effects estimates

To solve the estimating equations iteratively using "Fisher's scoring algorithm" requires the gradient of $\mathcal{W}_i$ w.r.t. $\mathbf{b}_i$, i.e. $\mathbf{A} = -\hat{Z}_i^T \Lambda_i \hat{Z}_i - D^{-i}$. Getting the next value of $\mathbf{b}_i$ in an iterative solution of the estimating equation involves finding $\mathbf{b}_i^*$ as solution of

$$\mathcal{W}_i + \mathbf{A}[\mathbf{b}_i^* - \mathbf{b}_i] = 0$$

That is

$$\mathbf{b}_i^* = D\hat{Z}_i^T \Sigma_i^{-1} \hat{\mathbf{r}}_\mathbf{i}$$

where $\hat{\mathbf{r}}_\mathbf{i} = y_i - \eta(\beta + \hat{\mathbf{b}}_i) + \hat{Z}_i \hat{\mathbf{b}}_i$

Estimating β

After estimating the $\{\mathbf{b}_i\}$ for fixed β at stage K, update the latter's estimated by "marginalizing out" the $\{\mathbf{b}_i\}$ & maximizing the marginal posterior. Result: estimating equations solved numerically [analogous to the random effects] The result:

$$\hat{\beta}^* = \hat{\beta} + [\sum_i \hat{X}_i^T \Sigma_i^{-1} \hat{X}_i]^{-1} [\sum_i \hat{X}_i^T \Sigma_i^{-1} \mathbf{r}_i]$$

is fixed & used for the $\{\beta_i^o\}$ in stage $K+1$ to get revised versions of the $\{\mathbf{b}_i\}$ & so on.

Likewise $\mathbf{D}$ and ϕ may be estimated by the maximizing the quasi log likelihood[a]. Robust estimates of the covariance matrix of the coefficient estimates vector can also be found.

[a] Burnett et al. [1994]

Summary

- Measurement error comes in a **variety of forms.** Each type can has its own special impacts
- Those **impacts are hard to predict** when nonlinear regression models are used in environmental epi
- **Spatial prediction needed** to reduce those effects
- **generalized estimating equations approach simplifies analysis** - only means & variances of spatial predictive distribution needed (tho more complicated alternatives OK)

Richard T Burnett, Robert E Dales, Mark E Raizenne, Daniel Krewski, Peter W Summers, Georgia R Roberts, May Raad-Young, Tom Dann, Jeff Brook, et al. Effects of low ambient levels of ozone and sulfates on the frequency of respiratory admissions to ontario hospitals. *Environmental research*, 65(2):172, 1994.

Kung-Yee Liang and Scott L Zeger. Longitudinal data analysis using generalized linear models. *Biometrika*, 73(1):13–22, 1986.

Mary J Lindstrom and Douglas M Bates. Nonlinear mixed effects models for repeated measures data. *Biometrics*, pages 673–687, 1990.

JAMES V Zidek, HUBERT Wong, ND Le, and RICK Burnett. Causality, measurement error and multicollinearity in epidemiology. *Environmetrics*, 7(4):441–451, 1996.