

*Using Estimating Equations for Spatially
Correlated Areal Data*

David Vock, NCSU

December 8, 2009

Introduction

GEEs

Spatial Estimating Equations

Implementation

Simulation

Conclusion

Typical Problem

- Assess the relationship between a spatially varying covariate and areal epidemiologic data; prediction not the focus
- Frequently the response is non-normal
- Should we put in a spatial random effect to smooth?
- Reich et al. (2006) showed using the CAR model can lead to large changes in the posterior mean and variance of fixed effects when spatially varying covariates collinear with spatial random effects

The Real Issues

If the model for the covariance of the response (conditioned on covariates) is wrong

- Often estimators for fixed effects still consistent
- Estimators will be inefficient
- But standard error estimates will be incorrect
- Leads to erroneous inference

Unlikely that one would know the form of the covariance of the conditional response a priori

- Want a method that is robust to misspecification of the covariance
- Only want to assume mean model is correct
- GOAL posit a model for the covariance and hopefully gain efficiency

Liang and Zeger (1986) seminal paper: Generalized Estimating Equations

Some Notation Used

- Observe pairs of independent data $(\mathbf{Y}_i|\mathbf{x}_i)$ for $i = 1, \dots, m$ where \mathbf{Y}_i is $(n_i \times 1)$
- Let $E(\mathbf{Y}_i|\mathbf{x}_i) = \mathbf{f}_i(\mathbf{x}_i, \boldsymbol{\beta})$
- $\text{var}(\mathbf{Y}_i|\mathbf{x}_i) = \mathbf{V}_i(\boldsymbol{\beta}, \boldsymbol{\xi}, \mathbf{x}_i) = \mathbf{T}_i^{1/2}(\boldsymbol{\beta}, \boldsymbol{\theta}, \mathbf{x}_i)\boldsymbol{\Gamma}_i(\boldsymbol{\alpha}, \mathbf{x}_i)\mathbf{T}_i^{1/2}(\boldsymbol{\beta}, \boldsymbol{\theta}, \mathbf{x}_i)$
- \mathbf{T}_i diagonal matrix with elements $\text{var}(Y_{ij}) = \sigma^2 g(\boldsymbol{\beta}, \boldsymbol{\theta}, \mathbf{x}_{ij})$, $\boldsymbol{\Gamma}_i$ correlation matrix

Liang and Zeger (1986) seminal paper: Generalized Estimating Equations

Main Idea

- Posit a working correlation matrix for the multivariate response within subject
- Solve $\sum_{i=1}^m \mathbf{X}_i^T(\boldsymbol{\beta})\mathbf{V}_i^{-1}\{\mathbf{Y}_i - \mathbf{f}_i(\mathbf{x}_i, \boldsymbol{\beta})\} = \mathbf{0}$ to estimate $\boldsymbol{\beta}$
- Solve quadratic estimating equations to obtain estimates for $\boldsymbol{\xi} = (\boldsymbol{\theta}, \boldsymbol{\alpha})$
- Obtain standard error estimates using Sandwich variance estimators

Properties of Generalized Estimating Equations

- Consistent, asymptotically normal estimator of β even if V_i was misspecified
- Sandwich variance estimators are consistent \rightarrow correct inference
- If V_i were correctly specified, most efficient RAL semi-parametric estimator (Tsiatis, 2006)
- Presumably, posit something close to the truth, gain efficiency
- Asymptotic here refers to $m \rightarrow \infty$
- Albert and McShane (1995) (brain imaging) considered repeated measures where there was spatial variability
- Most spatial epidemiology problems have $m = 1$

Extending GEEs to Spatial Epidemiology

- Note on notation: Since we will only consider $m = 1$ for the rest of the talk we will drop the subscript i
- Gotway and Stroup (1997) discuss a Quasi-Likelihood approach which would lead to same form of the estimating equations above with $m = 1$
- But V must satisfy various assumptions not easily verified
- Perhaps can borrow ideas from time series: Zeger (1988)

Development of McShane et al. (1997)

- Developed model for spatially correlated count data
- Main idea: Conceive of some non-negative weakly stationary latent (spatial) process $\boldsymbol{\epsilon} = \{\epsilon(s_1), \dots, \epsilon(s_n)\}$ where s_1, \dots, s_n are the spatial locations, and then assume
$$E(y(s_j)|\boldsymbol{\epsilon}(s_j)) = \exp(\mathbf{x}_j^T \boldsymbol{\beta}) \epsilon(s_j)$$
$$\text{var}(y(s_j)|\boldsymbol{\epsilon}(s_j)) = E(y(s_j)|\boldsymbol{\epsilon}(s_j))$$
 - like a Poisson process
$$\text{cov}(y(s_j), y(s_k)) = \sigma^2 \rho_{\boldsymbol{\epsilon}}(\mathbf{h}, \xi)$$
 - \mathbf{h} is the distance
$$E(\boldsymbol{\epsilon}(s)) = 1$$
- This induces a model for the mean and covariance matrix of the unconditional moments of $\mathbf{y}(s)$ which could be solved for using the law of iterated expectations and covariances

Asymptotic Properties of McShane et al. (1997)

- Zeger showed that if $\epsilon(s)$ is a mixing process then so will $\mathbf{y}(s)$ and the solution to the estimating equations
$$\mathbf{X}^T(\boldsymbol{\beta})\mathbf{V}^{-1}(\boldsymbol{\beta}, \boldsymbol{\xi}, \mathbf{x})\{\mathbf{Y} - \mathbf{f}(\mathbf{x}, \boldsymbol{\beta})\} = \mathbf{0}$$
 will be consistent and asymptotically ($n \rightarrow \infty$) normal for $\boldsymbol{\beta}$
- $\boldsymbol{\xi}$ can be estimated using quadratic estimating equations
- Did not need to consider this model through a latent spatial process
- Could have just directly posited the model for the mean and covariance of $\mathbf{y}(s)$ as long as $\mathbf{y}(s)$ was a mixing process

Variance Estimation in McShane et al. (1997)

- If the covariance matrix \mathbf{V} is correctly specified then asymptotically $\text{var}(\hat{\boldsymbol{\beta}}) = \{\mathbf{X}^T(\boldsymbol{\beta})\mathbf{V}^{-1}(\boldsymbol{\beta}, \boldsymbol{\xi}, \mathbf{x})\mathbf{X}^T(\boldsymbol{\beta})\}^{-1}$
- If in truth the covariance is \mathbf{U} then asymptotically $\text{var}(\hat{\boldsymbol{\beta}}) = (\mathbf{X}^T\mathbf{U}^{-1}\mathbf{X}^T)^{-1}(\mathbf{X}^T\mathbf{U}^{-1}\mathbf{V}\mathbf{U}^{-1}\mathbf{X}^T)(\mathbf{X}^T\mathbf{U}^{-1}\mathbf{X}^T)^{-1}$
- Zeger (1988) only considers the possibility of misspecification for computational purposes
- Note the asymptotic standard errors do not depend on how $\boldsymbol{\xi}$ was estimated
- Suggest that the estimated standard errors are likely to be optimistic
- Unfortunately SAS cannot easily obtain these sandwich-like variance estimators

Implementation in SAS

- PROC GENMOD typically used to GEE models
- Does not have option for spatial working correlation matrix and only uses method of moments estimators for parameters in the working correlation matrix
- PROC GLIMMIX usually thought of as software for Generalized Linear Mixed Models but can also do GEE
- Able to posit spatial working correlation matrices and uses quadratic estimating equations to estimate parameters in the working correlation matrix

Warning on the Interpretation of Parameters

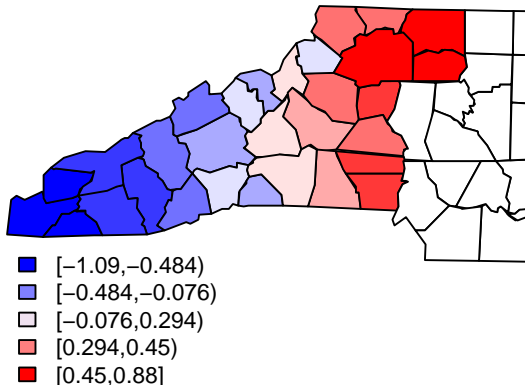
- Repeated measures analysis often think of population-averaged vs. subject-specific models
- Estimating equation approach lends itself to the population average approach
- With a log-linear mean model: $\exp(\beta_1)$ is the factor by which the response changes averaged across all units for a one unit change in x_1
- Likely the perspective preferred by epidemiologists

Simulation Design - Overview

- Interested in assessing the small-sample properties of these spatial estimating equations
- Assess the effect of adding a spatial random effect on the fixed effects
- Consider a spatially varying covariate x_i over 30 western North Carolina counties
- 500 total diseases with expected number of diseases, E_j , per county proportional to the July, 2008 population estimate
- Observed number of diseases, O_j drawn from a (correlated) Poisson distribution with mean = $E_j \exp(\beta_0 + \beta_1 x_j)$, where $\beta_0 = 0$ and $\beta_1 = -0.4$
- 1,000 Monte Carlo Runs
- Each data set analyzed using independence, spherical, and exponential working correlation matrices

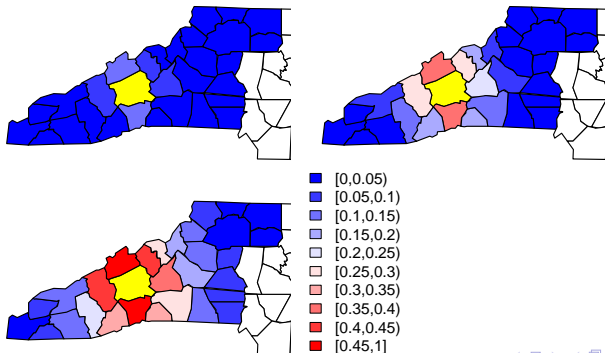
Simulation Design - Spatially Varying Covariate

- Project county centroid to x, y grid
- Spatially varying covariate that was function of x -coordinate plus random error



Simulation Design - Correlation Structure

- Consider three different spatial correlation structures for the response to generate the data: Exponential correlation with range parameter 15 km, 30 km, and 45 km (approximately 45 km, 90 km, and 135 km for effective spatial range)



Simulation Results - Range Parameter 15 km

Working Cor.	% Relative Bias	$\text{var}(\hat{\beta}_1)$	Mean $\{se(\hat{\beta}_1)\}^2$	Ratio	Coverage
Exponential	-0.0118	0.0141	0.0124	0.883	0.936
Independent	-0.0099	0.0110	0.0102	0.924	0.943
Spherical	-0.0067	0.0132	0.0109	0.822	0.928

Table: Maximum standard errors % Relative Bias: 0.0094; $\text{var}(\hat{\beta}_1)$: 0.0006; Mean $\{se(\hat{\beta}_1)\}^2$: 0.0003; Ratio: 0.045; Coverage 0.008

Simulation Results - Range Parameter 45 km

Working Cor.	% Relative Bias	$\text{var}(\hat{\beta}_1)$	Mean $\{se(\hat{\beta}_1)\}^2$	Ratio	Coverage
Exponential	-0.014	0.0219	0.0170	0.774	0.886
Independent	-0.022	0.0254	0.0103	0.404	0.809
Spherical	-0.017	0.0244	0.0221	0.902	0.913

Table: Maximum standard errors % Relative Bias: 0.013; $\text{var}(\hat{\beta}_1)$: 0.0011; Mean $\{se(\hat{\beta}_1)\}^2$: 0.0003; Ratio: 0.045; Coverage 0.012

Conclusion

- We have shown that linear estimating equations can be used for analyzing spatially correlated areal data and those estimators are consistent and asymptotically normal regardless of whether the assumed covariance between the response is correct. We have argued using a simulation study that assuming a spatial covariance will often lead to improved efficiency in the parameter estimates even with a small sample size.

References

- Albert P. S. and McShane L. M. (1995). A generalized estimating equations approach for spatially correlated binary data: Applications to the analysis of neuroimaging data. *Biometrics* **51**, 627-638.
- Gotway, C. A. and Stroup, W. W. (1997). A generalized linear model approach to spatial data analysis and prediction. *Journal of Agricultural Biological and Environmental Statistics* **2**, 157-178.
- Liang K.-Y. and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* **73**, 13-22.
- Lin P.-S. and Clayton, M. K. (2005). Analysis of binary spatial data by quasi-likelihood estimating equations. *Annals of Statistics* **33**, 542-555.
- McShane, L. M., Albert P. S., and Palmatier, M. A. (1997). A latent process regression model for spatially correlated count data. *Biometrics* **56**, 698-706.
- Reich B. J., Hodges, J. S., and Zadnik, V. (2006). Effects of residual smoothing on the posterior of the fixed effects in disease mapping model. *Biometrics* **62**, 1197-1206.
- Tsiatis, A. A. (2006). *Semiparametric Theory and Missing Data*. New York: Springer.
- Zeger, S. L. (1988). A regression model for time series of counts. *Biometrika* **74**, 621-629.