Accounting for Under Reporting in Disease Counts

The presence only problem

Rodelyn Jaksons, Elena Moltchanova, Beverley Horn, Elaine Moriarty

University of Canterbury, ESR

Table of contents

The presence only problem

The presence only problem

- Information is only known about the presences
- The observed counts are only a subset of the population
- Often encountered in epidemiology, ecology, criminology etc
- It is problematic when we want to estimate population prevalence/incidence

How can we account for this in our estimation procedure?

The Likelihood

The observed counts Y are only a subset of the true number infected Z, from population N:

$$Z \sim Binomial(N, \lambda)$$

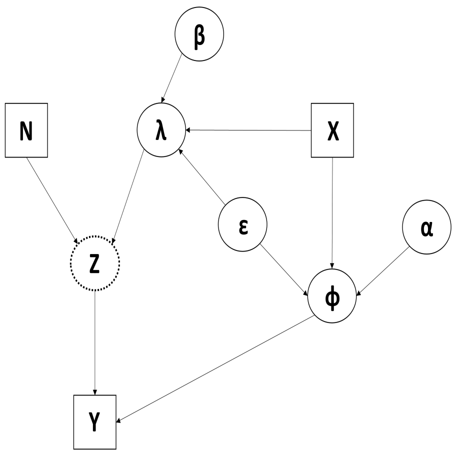
$$Y \sim Binomial(Z, \phi)$$

$$\begin{aligned} f(Y|N, \lambda, Z, \phi) &= \sum_{Z} \binom{N}{Z} \lambda^{Z} (1-\lambda)^{N-Z} \binom{Z}{Y} \phi^{Y} (1-\phi)^{Z-Y} \\ &= \binom{N}{Y} (\lambda\phi)^{Y} (1-\phi\lambda)^{N-Y} \end{aligned}$$

$$Y \sim Binomial(N, \lambda\phi)$$

The only way is Bayes!

Bayesian Hierarchical Models



$$Y \sim Binomial(N, \lambda(X)\phi(X))$$

$$\lambda(X) = \frac{exp\{X\beta + \varepsilon\}}{1 + exp\{X\beta + \varepsilon\}}$$

$$\phi(X) = \frac{exp\{X\alpha + \varepsilon\}}{1 + exp\{X\alpha + \varepsilon\}}$$

- α and β are vectors of regression coefficients
- ε is a spatial residual, with CAR structure

Posterior Distribution

The joint posterior distribution:

$$f(\lambda, \phi | Y, N) \propto f(Y | N, \lambda, \phi) f(\lambda) f(\phi).$$

The posterior distribution for Z:

$$f(Z | \lambda, \phi, Y, N) = \int_0^1 \int_0^1 \binom{N}{Z} \lambda^Z (1-\lambda)^{N-Z} \binom{Z}{Y} \phi^Y (1-\phi)^{Z-Y} f(\lambda) f(\phi) d\phi d\lambda.$$

Simulation Studies

The Scenarios

Underlying Prevalence, $\lambda(X)$:

$$\beta_0 = \{-6.7, -2\}, \beta_1 = \{0, 2.5\}$$

Detection ϕ:

$$\phi = \{0.7, 0.9\}$$

Presence of spatial autocorrelation ε:

$$\varepsilon = \{\varepsilon_1, \varepsilon_2\}$$

with ε_1 indicating spatial autocorrelation and white noise otherwise.

- covariates are considered only for λ
- ϕ is treated as an intercept only model.

In this 2^4 different scenarios were investigated.

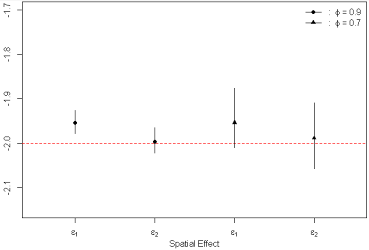
Priors

- $\beta_0, \beta_1 \sim N(0, 0.04)$
- $\phi|\phi = 0.9; \alpha \sim N(logit(0.9), 100)$
- $\phi|\phi = 0.7; \alpha \sim N(logit(0.7), 100)$
- $\varepsilon_{ij} \sim N(\bar{\varepsilon}_{-j}, \tau m_j)$

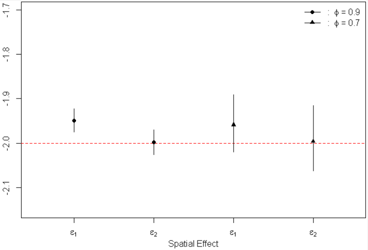
Identifiability

- Placing an informative prior on ϕ was necessary for identifiability and convergence of the model

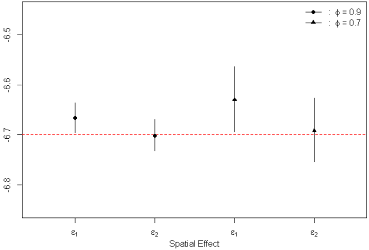
Results



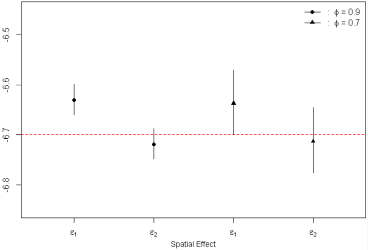
(a) $\hat{\beta}_0|\beta_0 = -2, \beta_1 = 0$



(b) $\hat{\beta}_0|\beta_0 = -2, \beta_1 = 2.5$

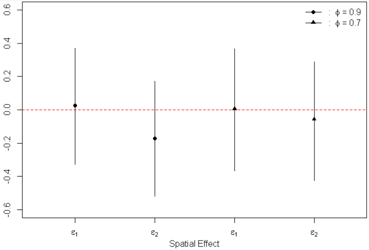


(c) $\hat{\beta}_0|\beta_0 = -6.7, \beta_1 = 0$

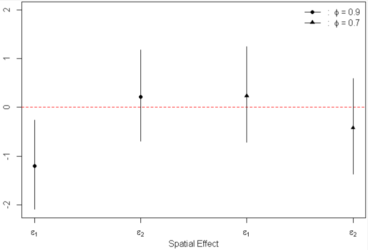


(d) $\hat{\beta}_0|\beta_0 = -6.7, \beta_1 = 2.5$

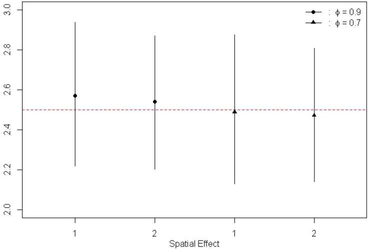
Results



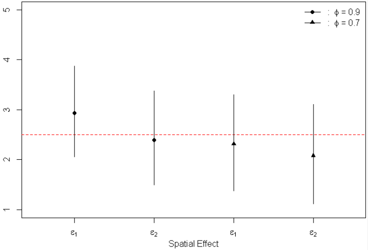
(a) $\hat{\beta}_1|\beta_0 = -2, \beta_1 = 0$



(b) $\hat{\beta}_1|\beta_0 = -6.7, \beta_1 = 0$



(c) $\hat{\beta}_1|\beta_0 = -2, \beta_1 = 2.5$



(d) $\hat{\beta}_1|\beta_0 = -6.7, \beta_1 = 2.5$

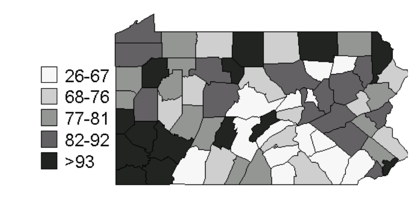
Pennsylvania Lung Cancer Data Set

Pennsylvania Lung Cancer Data Set

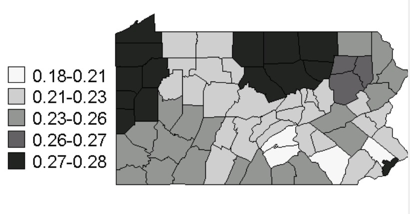
- Pennsylvania Lung Cancer Data available from the SpatialEpi package in R.
- Comprises of lung cancer cases and population counts at the county level, with $n = 67$.
- County-specific smoking rates.
- Population counts were obtained from the 2000 decennial census
- Stratified on race (white vs non-white), gender and age (Under 40, 40-59, 60-69 and 70+).

For simplicity, we aggregated the data to county specific level only.

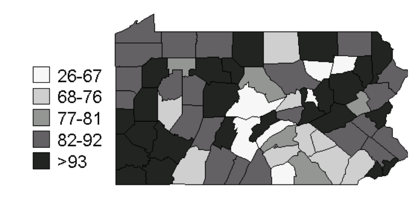
Pennsylvania Lung Cancer Data Set: Results



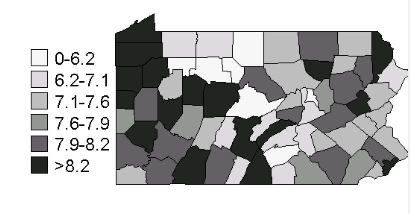
(a) Observed Prevalence



(b) Proportion of Smokers



(c) Predicted Prevalence



(d) Predicted - Observed

Conclusion and Future Work

Summary

- Can be used to estimate true population parameters
- Potential to uncover areas of under detection/ under reporting

Future Work

- Application to kidney stones data set
- Investigate the correlations between hierarchies
- Covariates on detection

Questions?