Supplemental material for "The Relationship between Tobacco Retailer Density and Neighborhood Demographics in Ohio"

Chiche Adibe[1], **Peter F. Craigmile**[2], **Nathaniel Onnen**[2], **Elli Schwartz**[3], **Megan E. Roberts**[3].

[1] College of Liberal Arts and Social Sciences, DePaul University, Chicago, IL, USA

[2] Department of Statistics, The Ohio State University, Columbus, OH, USA

[3] College of Public Health, The Ohio State University, Columbus, OH, USA

Last updated December 5, 2018

In this supplement we provide more detail about the statistical model that we fit to the data, and how we account for spatial dependence to provide estimates of the covariance of the model coefficients.

A marginal negative binomial model, that accounts for spatial dependence

Suppose that there are m census tracts. Let Y_i denote the number of establishments in census tract i ($i = 1, \ldots, m$), and let P_i denote the population (in 1000s) in tract i. In our marginal model for the counts Y_i, we assume that the expectation is

$$E(Y_i) = \mu_i,$$

where for a set of covariates of length p, $\boldsymbol{x}_i$, and coefficients $\boldsymbol{\beta}$,

$$\eta_i = \log(\mu_i) = \log(P_i) + \boldsymbol{x}_i^T \boldsymbol{\beta}.$$

Equivalently our model for the log retailer rate is

$$\log(\mu_i/P_i) \;=\; \boldsymbol{x}_i^T \boldsymbol{\beta}.$$

With a negative binomial model, the variance depends on the mean μ_i through the expression

$$\mathrm{var}(Y_i) = V(\mu_i) = \mu_i + \frac{\mu_i^2}{\theta},$$

where $\theta > 0$ is known as the *dispersion parameter.*

To account for the spatial dependence for the counts over the census tracts we assume that

$$\mathrm{cov}(Y_i, Y_j) = \sqrt{V(\mu_i)V(\mu_j)}\,\rho_{ij},$$

where ρ_{ij} parameterizes the correlation in the counts between tracts i and j. We assume a *conditional autoregressive (CAR)* model for $\boldsymbol{R}$, the $m \times m$ correlation matrix with (i, j) element ρ_{ij}. To define this correlation matrix, we first specify a $m \times m$ spatial proximity matrix $\boldsymbol{W}$ as follows: the (i, j) element of $\boldsymbol{W}$ is equal to 1 if census tract j is a neighbor of tract i, and zero otherwise (we assume each tract cannot be a neighbor of themselves). Next, let $\boldsymbol{C}$ denote a diagonal $m \times m$ matrix, where the ith diagonal element is equal to the number of neighbors that census tract i has. Then

$$\boldsymbol{R} \;=\; \boldsymbol{C} - \alpha \boldsymbol{W},$$

where $-1 < \alpha < 1$ is known as the *spatial dependence parameter.*

An estimate of the covariance of $\boldsymbol{\beta}$, assuming residual spatial dependence

We fit a negative binomial model to the Y_i values, assuming a working covariance of independence between the counts. (We will correct this assumption later in the section.) Let $\widehat{\boldsymbol{\beta}}$ denote the estimate of the model coefficients $\boldsymbol{\beta}$ under this assumption. We use a sandwich estimator to correct the covariance of $\widehat{\boldsymbol{\beta}}$ for the spatial dependence in the counts.

Let $\boldsymbol{X}$ denote an $m \times p$ design matrix with ith row $\boldsymbol{x}_i$ or (i, j) element x_{ij}, and let $\boldsymbol{V}$ denote a diagonal matrix with (i, i) element $V(\mu_i)$. Let $\boldsymbol{D}$ be an $m \times p$ matrix with (i, j) element

$$D_{ij} \quad = \quad \frac{\partial \mu_i}{\partial \beta_j} \quad = \quad \frac{\partial \mu_i}{\partial \eta_i}\frac{\partial \eta_i}{\partial \beta_j} \quad = \quad \mu_i \, x_{ij},$$

since

$$\frac{\partial \mu_i}{\partial \eta_j} \quad = \quad \left[\frac{\partial \eta_i}{\partial \mu_i}\right]^{-1} \quad = \quad \left[\frac{\partial \log(\mu_i)}{\partial \mu_i}\right]^{-1} \quad = \quad \mu_i.$$

Then the sandwich estimator of the covariance is

$$\text{cov}(\widehat{\boldsymbol{\beta}}) \quad = \quad (\boldsymbol{D}^T\boldsymbol{V}^{-1}\boldsymbol{D})^{-1}\boldsymbol{D}^T\boldsymbol{V}^{-1}\text{cov}(\boldsymbol{Y})\boldsymbol{V}^{-1}\boldsymbol{D}(\boldsymbol{D}^T\boldsymbol{V}^{-1}\boldsymbol{D})^{-1}.$$

To estimate $\text{cov}(\boldsymbol{Y})$ for each $i = 1, \ldots, m$, let $r_i = (Y_i - \widehat{\mu}_i)/\sqrt{V(\widehat{\mu}_i)}$ denote the *Pearson residuals* of the model and note that when our statistical model is true

$$\text{cov}(r_i, r_j) \quad \approx \quad \rho_{ij},$$

for each i and j. We estimate α from the Pearson residuals via maximum likelihood (ML). let $\widehat{\alpha}$ denote the ML estimate of α, and with this estimate let $\widehat{\rho}_{ij}$ denote the resulting estimate of ρ_{ij}. Then our estimate of $\text{cov}(\boldsymbol{Y})$ is

$$\widehat{\text{cov}}(Y_i, Y_j) = \sqrt{V(\widehat{\mu}_i)V(\widehat{\mu}_j)}\;\widehat{\rho}_{ij}.$$

Letting $\boldsymbol{J} = \boldsymbol{D}^T\boldsymbol{V}^{-1}\boldsymbol{D}$, our sandwich estimator of the covariance of the estimated model coefficients can be written as

$$\widehat{\text{cov}}(\widehat{\boldsymbol{\beta}}) \quad = \quad \boldsymbol{J}^{-1}\boldsymbol{B}^T(\boldsymbol{C} - \widehat{\alpha}\boldsymbol{W})\boldsymbol{B}\boldsymbol{J}^{-1},$$

where for this model

$$\boldsymbol{B} = \text{diag}\left(\frac{\widehat{\mu}_i}{\sqrt{V(\widehat{\mu}_i)}}\right)\boldsymbol{X}.$$