Journal of Modern Applied Statistical Methods

Volume 10 | Issue 2 Article 7

11-1-2011

Robust Inference for Regression with Spatially Correlated Errors

Juchi Ou
Case Western Reserve University, jxo37@cwru.edu

Jeffrey M. Albert
Case Western Reserve University, jma13@case.edu

Journal of Modern Applied Statistical Methods
November 2011, Vol. 10, No. 2, 462-475

1538 – 9472/11/$95.00

Robust Inference for Regression with Spatially Correlated Errors

Juchi Ou Jeffrey M. Albert
Case Western Reserve University,
Cleveland, OH

A robust variance estimator for a regression model with spatially correlated errors is proposed using the estimated empirical covariogram. Simulations studies show unbiasedness and robustness for the OLS but not for the GLS estimates. The new robust variance estimation method is applied to hospital quality data.

Key words: Ordinary least squares, generalized least squares, robust variance estimation, hospital quality, semivariogram.

Introduction

In observational studies, an objective of interest is to compare the mean response of exposed and unexposed units. Commonly, the effect of an exposure or treatment on an outcome is evaluated via conventional linear regression models that assume independence of errors. For geographical data, observations and corresponding errors may be spatially correlated rather than independent. One unbiased estimator of an exposure effect in a linear regression model is the ordinary least squares estimator (OLS). This estimator is known to be the best linear unbiased estimator (BLUE) when the errors are independent with a constant variance. However, when errors are correlated, this estimator may be inefficient. Furthermore, its standard variance estimator may be biased. To improve precision for correlated data, methods that take into account the correlation structure, such as maximum likelihood (ML) estimation and generalized least squares (GLS) are of interest for evaluating an exposure effect.

Juchi Ou earned a Ph.D. in Biostatistics in 2010 from Case Western Reserve University. Email: jxo37@cwru.edu. Jeffrey M. Albert is an Associate Professor of Biostatistics in the Department of Epidemiology and Biostatistics at the School of Medicine. His research interests include longitudinal data analysis and causal inference. Email: jma13@case.edu.

A number of researchers have studied regression models with serially or spatially correlated errors. For example, Lee & Lund (2004) provided expressions for the OLS variances for autocorrelated errors and proposed confidence intervals based on their derived variance. The empirical coverage probabilities of their confidence intervals were close to the 95%target value when the sample size was large (at least 500). Athough Lee & Lund studied the variance for time series autocorrelation structures, their results require extension to regression models where errors are correlated in a space.

Basu & Reinsel (1994) compared the OLS and GLS estimators when errors follow a spatial unilateral first-order autoregressive moving average model; they found that the difference between variances of the two estimators were small unless the spatial correlation was close to 1. They investigated autocorrelation models; however, regression model errors could follow other spatial structures, such as a spatial Gaussian or spatial exponential model. Mardia and Marshall (1984) developed ML estimators for regression parameters in the spatial context assuming the errors follow a spatial Gaussian distribution.

A limitation of previous methods of inference for spatial data is that they rely on a correct specification of the covariance structure. When the covariance matrix is unknown, methods for variance estimation that are robust to covariance model misspecification are of interest. In the context of longitudinal data, a

well-known robust method to improve variance estimators for correlated data is the sandwich variance estimator (Diggle, et al., 2003). However, this estimator is not suitable for spatially correlated data that involve a single multivariate observation as opposed to multiple independent vectors. Furthermore, previous researches have given little attention to properties of estimators of the variance of effect estimates for spatially correlated errors.

This article develops estimators for mean differences along with robust variance estimators in a regression model with spatially correlated errors. A new robust (sandwich) variance estimator for exposure effects is proposed using the empirical variogram for spatially correlated errors. Although this approach may be applied to the maximum likelihood estimate, the focus here is on the methods of ordinary and generalized least squares. The appeal of the latter is that is has computational advantages over ML estimation and retains equivalent asymptotic efficiency (Charnes, et al., 1976).

The OLS and GLS estimators, along with the proposed versus standard variance estimators, are assessed via simulation studies. Simulation data were generated under either a spatial Gaussian or spatial exponential model, both of which are commonly used to analyze spatial data. As an applied example, data is analyzed to assess the effect of urban versus rural locations on the number of full-time equivalents (FTE) for registered nurses. Previous researchers investigating this question (Rosenblatt, et al., 2006; Jiang, et al., 2006) did not consider the spatial pattern of hospitals in assessing the difference in mean FTE. Therefore, the proposed methods are applied to consider the difference in mean FTE between urban and rural hospitals taking into account spatial correlations among hospitals. The data analyzed are from two databases: hospital financial reports from the Office of Statewide Health Planning and Development, and HCUP State Inpatient Databases (SID).

Methodology

Assume a linear regression model, standard (OLS and GLS) approaches for estimations of regression parameters and that the outcomes (Y(s)) and covariates (X(s)) at location s are linearly related. Also, the errors, e(s), for this linear regression model are allowed to be correlated, where s is an index for a spatial location. This model is as follows:

$$Y(s) = X(s)\beta + e(s);\ \ e(s) \sim N(0; \Sigma), \qquad (1)$$

where Σ represents the variance-covariance matrix for the error vector. The argument, (s), will be dropped for ease of notation.

For correlated errors, two common estimators of regression parameters (β) are the ordinary least squares (OLS) and the generalized least squares (GLS) estimators. The OLS estimator of regression parameters is

$$\hat{\beta}_{ols} = (X'X)^{-1}X'Y\ ; \qquad (2)$$

and the corresponding naïve variance estimator for $\hat{\beta}_{ols}$ is

$$Var(\hat{\beta}_{ols}) = \hat{\sigma}^2 (X'X)^{-1}\ , \qquad (3)$$

where $\hat{\sigma}^2$ is the sample variance of residuals. Another estimator of regression parameters is the GLS estimator,

$$\hat{\beta}_{gls} = (X'W^{-1}X)^{-1}X'W^{-1}Y\ , \qquad (4)$$

where W is the working matrix and it is equal to the estimated covariance matrix. The corresponding naïve variance estimator is

$$Var(\hat{\beta}_{gls}) = (X'W^{-1}X)^{-1}\ . \qquad (5)$$

Both the OLS and the GLS point estimators are unbiased, but the variance of the GLS estimator is smaller than that of the OLS estimator (Bloomfield & Watson, 1975) when W^{-1} is equal to the true covariance matrix. In the conventional, so-called naïve or model-based, approach, the covariance structure for the OLS variance estimator is assumed to follow the independence model whereas that for the GLS variance estimator is assumed to be proportional

to the working weight matrix W. In the context of longitudinal data, Liang & Zeger (1986) showed that the point estimator for β via generalized estimating equations (GEE) is consistent even if the correlation matrix is misspecified. However, when the assumed covariance structure is different from the true covariance model, the naïve variance estimator is inconsistent.

Robust Variance Estimator

The model-based variance estimators described above may be inadequate when the spatial covariance structure is unknown with the possibility of being misspecified. In the case of longitudinal data, where there are multiple measurements for each subject, a robust (sandwich) variance estimator is available (Diggle, et al., 2003). The robust variance estimator for the generalized least squares estimator $\hat{\beta}_{gls}$ is

$$\mathrm{Var}(\hat{\beta}_{gls})=(X'W^{-1}X)^{-1}X'W^{-1}\hat{V}W^{-1}X(X'W^{-1}X)^{-1}, \quad (6)$$

where $\hat{V}$ is a block-diagonal matrix with non-zero block $\hat{V}_0$ which may be estimated via restricted maximum likelihood estimation (REML). Letting Y_{hij} denote the j^{th} measurement on the i^{th} unit in the h^{th} group, the sample mean for the measurement j in group h is

$$\hat{\mu}_{hj}=\frac{1}{m_h}\sum_{i=1}^{m_h}Y_{hij}, h=1,...,g; i=1,...,m_h; j=1,...,n, \quad (7)$$

and the REML estimator is

$$\hat{V}_0=\left(\sum_{h=1}^{g}m_h-g\right)\sum_{h=1}^{g}\sum_{i=1}^{m_h}(Y_{hi}-\hat{\mu}_h)(Y_{hi}-\hat{\mu}_h)', \quad (8)$$

where $Y_{hi}=(Y_{hi1},...,Y_{hin})'$ and $\hat{\mu}_h=(\hat{\mu}_{h1},...,\hat{\mu}_{hn})'$. For this estimator, no assumption exists regarding the structure of means and covariance matrix.

In the case of longitudinal data where there are independent realizations of the correlated responses, sample estimates of the variance and covariance parameters are generally used to obtain the empirical estimate of V. For spatial data, there is only one (multivariate) observation and the above robust estimator would not be a good estimator. For this case, an empirical covariogram is used in place of the empirical variance-covariance matrix used for longitudinal data.

Variogram

Assume the spatial process to be second-order stationary and isotropic, where stationarity means that absolute coordinates are unimportant and isotropic means that the spatial correlations are the same in different directions (i.e., north-south versus west-east). For a spatial process Y(s): $s \in D \subset R^2$, one common tool to measure spatial correlations is the semivariogram for geostatistical data. The semivariogram $(\gamma^*(s_i, s_j) \equiv \gamma(s_i - s_j) = \gamma(h))$ is defined as a function of the distance (h) of two locations (s_i, s_j),

$$\gamma(h)=\frac{1}{2}\mathrm{Var}[Y(s_i)-Y(s_i+h)]. \quad (9)$$

If the spatial process (Y(s)) is second-order stationary, the semivariogram can be expressed in terms of the covariance function, $C(h)$, and

$$\gamma(h)=C(0)-C(h). \quad (10)$$

There are two important components for a semivariogram: the sill and the spatial range. The sill is defined as the asymptote of the variogram function, and the range is the distance at which the sill is reached.

Two commonly used variogram models are the spatial Gaussian and the spatial exponential models. Their covariance functions are as follows:

1. Gaussian model: $C_g(h)=\sigma^2\exp\{-(h/\alpha)^2\}$, and

2. Exponential model: $C_x(h)=\sigma^2\exp\{-(h/\alpha)\}$,

where α and σ^2 represent the spatial range and the sill, respectively, and h is the distance between two locations. The semivariograms for these two models are shown in Figure 1. As the distance increases, the semivariogram increases. The parameters $\theta \equiv (\alpha, \sigma^2)$ for a variogram model ($\gamma(h,\theta)$) may be estimated by iteratively reweighted least squares (IWLS) to minimize the following expression,

$$\sum |N(h)|(\hat{\gamma}(h)-\gamma(h,\theta))^2 , \qquad (11)$$

where N(h) is the number of distinct pairs of locations at distance h and $\hat{\gamma}(h)$ is an estimate of the semivariogram.

To avoid a parametric assumption regarding the spatial model, the moment-based empirical semivariogram could be used to estimate the semivariogram. The empirical (Matheron) semivariogram ($\hat{\gamma}$) for two observed measurements ($Y(s_i),Y(s_j)$) with distance h between two different locations (s_i,s_j) is

$$\hat{\gamma}(h)=\frac{1}{2|N(h)|}\sum_{N(h)}(Y(s_i)-Y(s_j))^2 , \qquad (12)$$

where |N(h)| is the number of measurement pairs with distance h. The corresponding empirical covariogram estimator for the covariance function, C(h) is as follows

$$\hat{C}(h)=\frac{1}{|N(h)|}\sum_{N(h)}(Y(s_i)-\bar{Y})(Y(s_j)-\bar{Y}) , \qquad (13)$$

where $\bar{Y}$ is the average of all Y(s). In this study, the empirical covariogram estimator is used to estimate the variance-covariance matrix.

Simulation Study

Data Generation

Using a 10x10 grid, two different covariance structures for the errors in Model 1 were studied: spatial Gaussian and spatial exponential. In general, the sill for a covariance structure varies from 0.01 to over 100. Therefore, the sill for both covariance structures was set to 9 in this study. The spatial ranges were set to 2, 5 or 10 in order to compare weak, modified and strong correlations between locations on a 10x10 grid. A binary covariate (X, with values 0 and 1) was generated from the binomial distribution with probability of X = 1 equal to 0.5 and the outcome (Y) was generated from the linear model

$$Y = 2X + \varepsilon, \qquad (14)$$

that is, the outcome was linearly related with the binary covariate with slope 2.

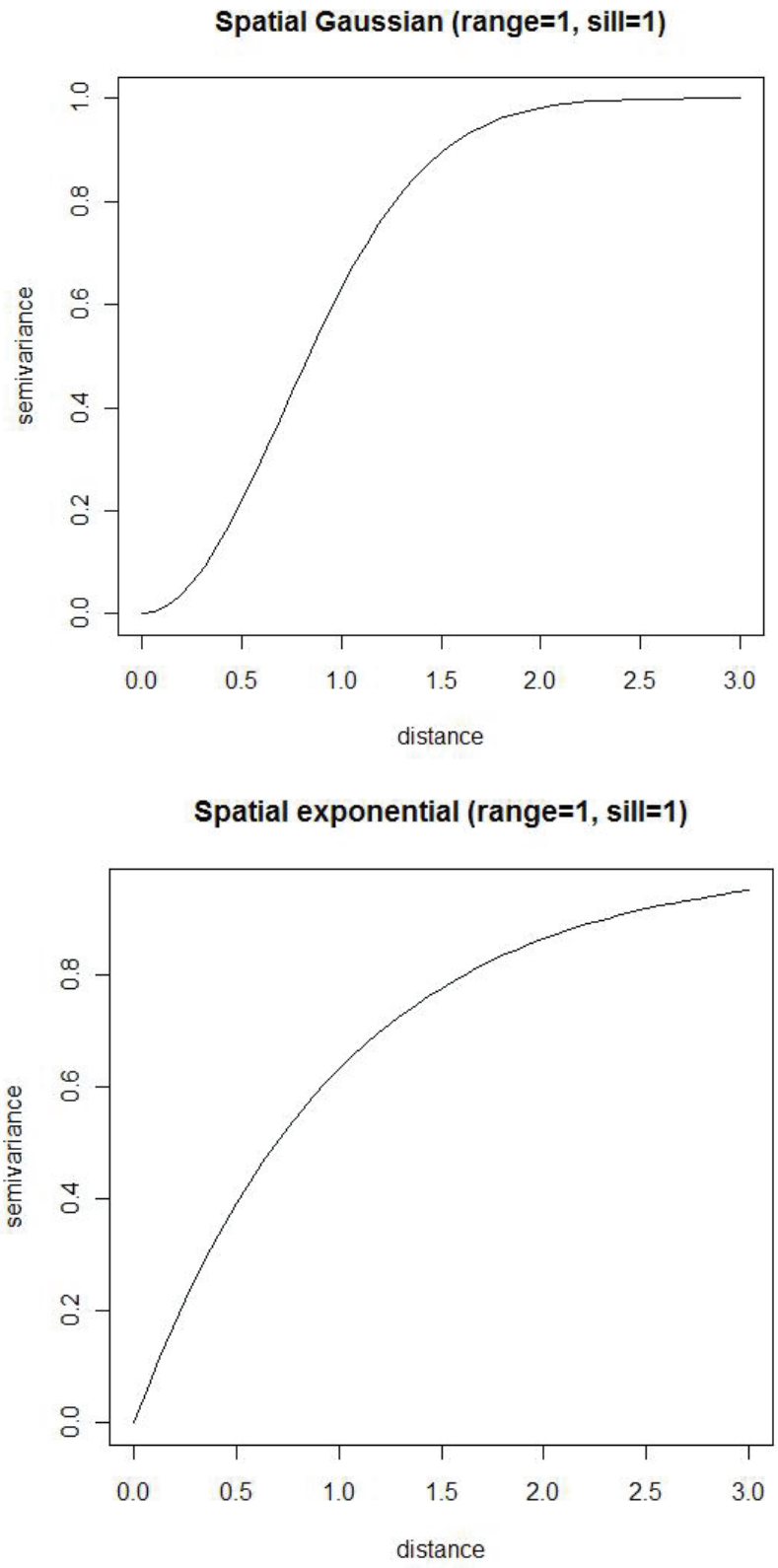
Estimator of the Exposure/Treatment

Two point estimators for the exposure/treatment effect were studied, namely, OLS (ordinary least squares) and GLS (generalized least squares) estimators. In addition, the working matrix of the GLS estimator was estimated based on either independence (OLS residuals), spatial Gaussian or spatial exponential.

Variance Estimator of the Treatment Effect

The naïve variance estimators as well as the sandwich variance estimators were evaluated. For the sandwich variance estimator, the variance-covariance matrix could be the spatial Gaussian $\left(\hat{C}_g\right)$, spatial exponential $\left(\hat{C}_x\right)$ or the spatial empirical covariance structure $\left(\hat{C}\right)$. The variance estimators for the OLS point estimator are as follows: independence, $\hat{\sigma}^2(X'X)^{-1}$; empirical, $(X'X)^{-1}X'\hat{C}X(X'X)^{-1}$; Gaussian, $(X'X)^{-1}X'\hat{C}_gX(X'X)^{-1}$; and Exponential: $(X'X)^{-1}X'\hat{C}_xX(X'X)^{-1}$.

Figure 1: Semivariogram Models



Where $\hat{C}$, $\hat{C}_g$ and $\hat{C}_x$ represent the spatial empirical covariance, the estimated spatial Gaussian covariance and the estimated spatial exponential covariance matrices. The variance estimators for the GLS point estimator are naïve, $(X'W^{-1}X)^{-1}$, and empirical, $(X'W^{-1}X)^{-1}X'W^{-1}\hat{C}W^{-1}X(X'W^{-1}X)^{-1}$, where W^{-1} would be either the spatial Gaussian or the spatial exponential covariance matrix, and $\hat{C}$ is the empirical covariance matrix.

The bias and MSE of the OLS and GLS point estimators of the regression coefficient were computed. The bias and MSE for 1,000 replications are obtained as

$$\text{Bias} = \frac{1}{1000}\sum(\hat{\beta}_i - 2), \qquad (15)$$

$$\text{MSE} = \frac{1}{1000}\sum(\hat{\beta}_i - 2)^2. \qquad (16)$$

In addition, the relative bias for each estimator ($\hat{\theta}$, that is, $\hat{\beta}$ or $\hat{V}(\hat{\beta})$) was calculated. This relative bias is defined as

$$\text{RB} = \frac{\hat{\theta}-\theta}{\theta}. \qquad (17)$$

Results

Spatial Gaussian Errors Data: OLS

The bias of the ordinary least squares estimator (OLS) and its corresponding variance estimator, in the case where the errors are spatially correlated over a 10 * 10 grid, are shown in Table 1. When the covariance matrix for errors is spatial Gaussian distributed, the bias of the OLS estimator is smaller (closer to 0.01) for all examined spatial ranges. The corresponding MSE decreases as the spatial range increases. Among the four variance estimators, the estimator using the independence covariance structure has the largest difference from the true variance for each spatial range. As the strength of spatial correlation (that is, the range) increases, the bias of the independence variance estimator increases. Both the empirical and the Gaussian variance estimators underestimate the variance. In addition, the empirical estimated variance is closer to the true value than the two estimators based on incorrect covariance models (independence and exponential) and has similar bias to the estimator using the correct covariance model (Gaussian), over varying range values.

Spatial Gaussian Errors Data: GLS

Working weight matrices for the GLS estimator based on the Gaussian and the exponential spatial covariance models were considered. The results for the Gaussian and exponential working matrices are shown in Table 2. For the Gaussian working matrix, the bias of the estimated effect is small for the each strength of the spatial correlations. The bias for the Gaussian working matrix is reduced at least 80% from the OLS estimators. The bias of the naïve estimated variance is smaller than that of the empirical estimator when the true working matrix (Gaussian model) was fit. However, as the spatial correlation increases, the relative bias of the naïve and empirical variance become more similar. When the exponential working matrix is used for the spatial Gaussian errors data, the biases of the GLS estimated effect are also small, and the bias is reduced at least 46.4% from the OLS estimators. In this case, the naïve and empirical variance estimators both have large biases which are similar in magnitude.

Spatial Exponential Errors Data: OLS

A second simulation involved the generation of spatial exponential errors. The bias and MSE for the ordinary least squares estimators (OLS) and its corresponding variance estimators are shown in Table3. The bias of the estimated effect is smaller than 0.005 for all examined spatial ranges. The independence estimator overestimates the variance of the effect for all examined spatial ranges and the spatial empirical estimator slightly underestimates the variance. The spatial empirical estimated variance is closer to the true value than the other estimated variances. The exponential variance estimator for the OLS estimator, though it uses the correct covariance model, underestimates the variance for all examined spatial ranges. The Gaussian variance estimator overestimates the variance when the spatial range is larger than 5.

Spatial Exponential Errors Data: GLS

For the spatial exponential errors data, two working weight matrices for the generalized least squares (GLS) estimator are considered: the spatial Gaussian and the spatial exponential covariance models. The results for the GLS effect estimators are shown in Table 4. For both Gaussian and exponential working matrices, the biases of estimated effects are smaller than 1% for all examined spatial ranges. When data are spatial exponential correlated across a study space (spatial range at 10), the biases of the GLS effect estimators are smaller than that of the OLS estimator. The bias reduction is 37.1% for a strongly spatial correlation. For the spatial exponential errors data, the relative bias decreases as the spatial range increases. When the spatial correlation (spatial range) increases, the MSE decreases.

For simulated data with exponential errors, the naïve (based on the correct working covariance matrix) and empirical variance estimates have positive biases for all examined spatial ranges. The bias of the naïve estimated variance is smaller than that of the empirical estimated variance. For all examined spatial correlations, the MSE of the GLS with incorrect (Gaussian) working matrix is larger than corresponding MSE of the GLS with correct (exponential) working matrix for the spatial exponential errors data.

Table 1: OLS-Bias and Variance Estimator for Spatial Gaussian Errors for 1,000 Replications

Range	OLS-Bias	MSE			Variance		
			TRUE	Indep*	Em*	Gau* (correct)	Ex*
2	0.0069	0.339	0.354	0.334	0.343	0.346	0.342
5	0.0108	0.136	0.146	0.222	0.146	0.132	0.142
10	0.0103	0.033	0.033	0.096	0.031	0.030	0.060

*indep: independent; Em: empirical; Gau: Gaussian; Ex: exponential

Table 2: GLS Bias and Variance Estimator for Spatial Gaussian Errors for 1,000 Replications

Range	Gaussian Working Matrix (Correct)					
	GLS-Bias	RB*	MSE	Variance		
				True(sim)*	Naïve	Em*
2	0.00138	-80.00%	0.0091	0.0091	0.0256	4.4886
5	0.00004	-99.60%	0.0020	0.0019	0.0019	0.9598
10	0.00089	-91.40%	0.0004	0.0004	0.0007	0.0333

Range	Exponential Working Matrix (Incorrect)					
	GLS-Bias	RB*	MSE	Variance		
				True(sim)*	Naïve	Em*
2	-0.0037	-46.40%	0.0238	0.0238	0.0928	0.0982
5	0.0005	-95.40%	0.0014	0.0014	0.0565	0.0314
10	0.0003	-97.10%	0.0008	0.0008	0.0403	0.0121

*RB: relative bias; True(sim): simulated variance; Em: empirical

Example

Background

A common cause of adult hospitalization is pneumonia. Several pneumonia inpatient management measures are provided by the Centers for Medicare & Medicaid Service. Among these quality measures, a blood culture prior to first antibiotic administration is recommended (Waterer & Wunderink, 2001; Metersky, et al., 2004). For care services in the hospitals, nurse staffing plays an important role. Kovner, et al. (2000, 2002) found that lower nurse staffing levels resulted in significantly higher rates of pneumonia. Rosenblatt, et al. (2006) and Jiang, et al. (2006) showed that the full-time equivalent (FTE) for registered nurses were significantly different between rural and urban community health centers in the US. However, although these studies assumed the hospital outcomes to be independent, they did not take into account possible spatial correlations among hospitals.

Data Source and Sample

This research is interested in examining the association between the FTEs for registered nurses and hospital location (urban versus rural). In general, one FTE represents 2,080 work hours

Table 3: OLS-Bias and Variance Estimator for Spatial Exponential Errors for 1,000 Replications

Range	OLS-Bias	MSE	Variance				
			TRUE	Indep*	Em*	Gau*	Ex* (correct)
2	0.0026	0.277	0.307	0.317	0.302	0.301	0.300
5	0.0041	0.171	0.185	0.223	0.185	0.187	0.177
10	0.0035	0.099	0.106	0.143	0.106	0.110	0.104

*indep: independent; Em: empirical; Gau: Gaussian; Ex: exponential

Table 4: GLS Bias and Variance Estimator for Spatial Exponential Errors for 1,000 Replications

Range	Gaussian Working Matrix (Incorrect)					
	Bias	RB*	MSE	Variance		
				True(sim)*	Naïve	Em*
2	-0.0037	42.30%	0.135	0.135	0.147	0.187
5	-0.0042	2.40%	0.056	0.056	0.061	0.091
10	-0.0032	-8.60%	0.029	0.029	0.030	0.050

Range	Exponential Working Matrix (Correct)					
	Bias	RB*	MSE	Variance		
				True(sim)*	Naïve	Em*
2	-0.0033	26.90%	0.130	0.130	0.146	0.187
5	-0.0031	-24.40%	0.054	0.054	0.066	0.092
10	-0.0022	-37.10%	0.027	0.027	0.034	0.050

*RB: relative bias; True(sim): simulated variance; Em: empirical

within a year to a fulltime worker. Here, the outcome of interested was FTEs for registered nurse per occupied bed. Data for this outcome, available in hospitals financial reports, was provided by the Office for Statewide Health Planning and Development (OSHPD). The binary predictor, hospital location (urban/rural), was taken from the Healthcare Cost and Utilization Project (HCUP) California State Inpatient Database (SID); this predictor was denoted as location. In addition, the report for pneumonia quality measures of inpatient management was provided by the Centers for Medicare & Medicaid Service. Data was merged from these three sources restricting the sample to hospitals in the State of California in 2004. The resulting dataset included 186 hospitals that reported: the above pneumonia quality measure, the number of registered nurse FTEs per occupied bed and hospital location.

The spatial correlation for each model variable was assessed via the test by Diblasi & Bowman (2001). The semivariograms of the response (FTE) and predictor (location) with their corresponding p-value of the spatial correlation test are shown in Figure 2. Both variables were spatially correlated across hospitals in California in 2004.

OLS Result

The effect of hospital location on the number of FTEs for registered nurses was estimated using the ordinary least squares (OLS). OLS estimates, the independence variance estimate,and three spatial variance estimates (empirical, spatial Gaussian, exponential structure) are shown in Table 5, along with standardized effect estimates (estimated effect divided by the square root of the estimated variance). The OLS estimated mean difference for FTE between urban and rural hospitals was 0.3018. The independence and spatial empirical variance estimates were close and both were less than 0.1. These two variance estimators both provided standardized effect estimates greater than 3.9. The spatial Gaussian and exponential variance estimates were larger, and their respective standardized estimates of 2.2 and 1.99, smaller than the other two estimates. Thus, all methods indicated an effect of the hospital locations on FTE with higher mean FTEs at the urban hospitals. The standardized effects based on the spatial Gaussian and spatial exponential estimated variances suggested marginal evidences; by contrast, the standardized effects based on independence and the empirical estimated indicated strong evidences of a location effect. The conclusions, based on California hospitals, are substantially the same as previous study results for United States health centers.

The semivariograms of OLS residuals are shown in Figure 3. The line in the left figure is the fitted spatial Gaussian structure with estimated spatial range and sill equal to 0.43 and 0.08. The line in the right figure is to the fitted spatial exponential structure with estimated range and sill equal to 0.50 and 0.11. Both theoretical semivariogram models (i.e., Gaussian and exponential) were close to empirical semivariogram when the distance was smaller than 2. However, these two models were far from empirical semivariogram when the distance was larger than 2.

GLS Result

For comparison, GLS estimators were considered under the same models as examined for the OLS estimators. Thus, estimated spatial Gaussian and exponential structures were used as the working weight matrices for GLS estimators. The results for the point and variance estimates are shown in Table 6. Compared to the OLS estimated effects, the two GLS estimated effects were larger. For each working weight matrix, both the naïve and the empirical variance estimates were less than 0.01. The empirical variance estimate was smaller than the naïve estimated variance for both the Gaussian and exponential working matrices. All three GLS standardized effect estimates were greater than 3.5 and one of them was as high as 3.73. All GLS standardized effect estimates indicated strong evidences of an effect of location on FTE, with a higher mean FTE at urban hospitals. Thus, the conclusion based on the GLS estimators with either a spatial Gaussian or exponential working matrices, agree with that given above for the OLS estimators.

Figure 2: Semivariograms of Response and Predictor

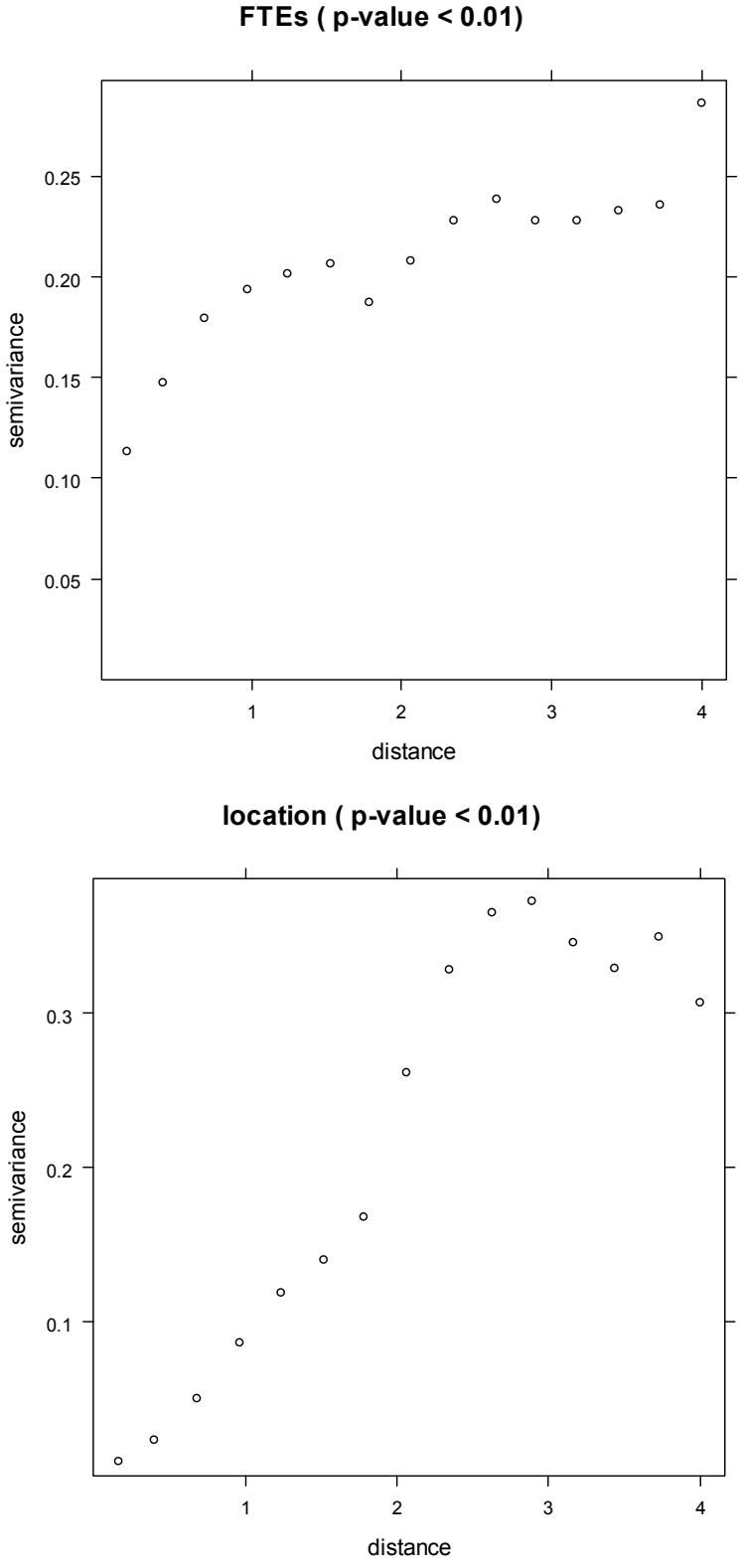


Table 6: GLS Effect Estimator and Its Estimated Variance (STD)*

	Working Matrix	
	Gaussian	Exponential
Estimated Effect	0.3255	0.3396
Naïve Variance	0.0081(3.62)	0.0089(3.60)
Empirical Variance	0.0076(3.73)	0.0085(3.68)

*Standardized effect estimates are in parentheses

Conclusion

This article addresses the problem of estimating exposure (or treatment) effect in a regression models with spatially correlated errors. Considering both OLS and GLS estimators, a new robust variance estimator was presented based on the estimated semivariogram. In order to evaluate the OLS and GLS estimators or their corresponding variance estimators under spatial correlated errors, simulation studies were conducted. Two different spatial correlation models were considered: spatial Gaussian and spatial exponential.

For spatial Gaussian and exponential simulated data, neither the OLS nor GLS estimators showed evidence of bias. When the spatial range increased, the true variance decreased. For the OLS estimator, the bias of the naïve (independence) estimated variance was smallest at spatial range 2 among three spatial ranges. The empirical estimated variance for the OLS estimator was closer to the true value than the other three estimated variances. For the GLS estimator, the naïve estimated variance was closer to the simulated variance than the empirical estimated variance. However, when the GLS estimator used an incorrect working matrix, the naïve estimated variance would be far from the simulated variance (e.g., GLS with an exponential working matrix for spatial Gaussian errors data). In addition, even when the correct working matrix is used, the estimated variance of the GLS estimate sometimes varied substantially from the true (simulation) value. Therefore, estimating exposure effects via ordinary least squares (OLS) with the empirical variance estimator is recommended when the data exhibit spatial patterns.

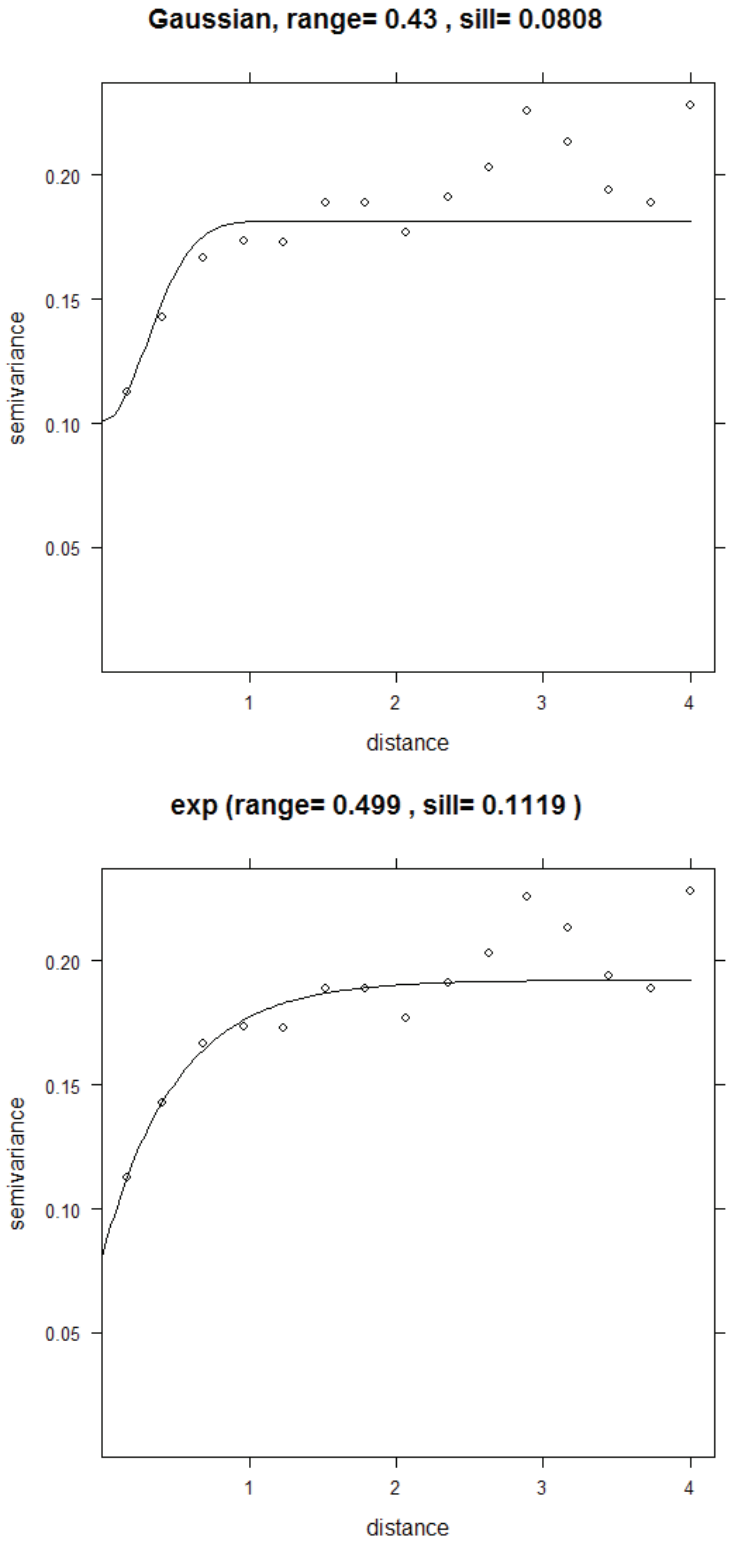
The effect of hospital locations on FTE where both variables exhibited spatial patterns (based on their empirical semivariogram and spatial correlation test) across California in 2004

Table 5: OLS Effect Estimate, Variance Estimates and Standardized Effect Estimates (STD)*

	Effect	Variance			
		Indep**	Empirical	Gaussian	Exponential
Estimate	0.3018	0.0059	0.0044	0.0184	0.0231
STD		3.9291	4.5498	2.2249	1.9857

*STD: the effect estimate divided by the square root of the variance estimate;
**Indep: independence covariance structure

Figure 3: Semivariograms of OLS Residuals



was examined. The linear relationship between hospital location (urban/rural) and full-time-equivalents (FTE) for registered nurse adjusted by the number of occupied beds was assessed via the OLS and the GLS estimators. From the semivariogram of the OLS errors, the OLS errors exhibited a spatial pattern. Therefore, the OLS estimated effect with corresponding empirical variance was preferred. Based on OLS, the estimated difference between urban and rural hospitals was 0.3 FTE. The empirical estimated variance for the OLS estimator was around 0.004 and the ratio of estimated effect to the square root of empirical variance was 4.55. This result, corroborating the previous findings, suggests that there is a significant difference in FTE for urban versus rural hospitals.

The robust approach proposed could be used with the maximum likelihood estimates, though results are expected to be similar to GLS. A limitation of this study is that it assumed the spatial field to be stationary. For a non-stationary field, semivariogram models are not valid as the semivariogram is not defined for non-stationary correlation structures. Another limitation is that the outcome was assumed to be continuous and normally distributed. For a categorical or other non-normally distributed outcome, the linear regression would not be suitable. It will be necessary to use the logistic regression or to do a Box-Cox transformation for such outcomes. In addition, for some extreme values, the Cressie-Hawkins robust estimator could be considered for the estimation of the semivariogram (Cressie & Hawkins, 1980) instead of the Matheron estimator. The empirical covariogram used is a biased estimator of the covariance function; therefore, the problem of the biased estimator of the covariogram will need to be solved in the future.

Acknowledgement

The authors would like to thank Mireya Diaz for directing the first author to this area of research and for helpful comments on an earlier draft.

References

Basu, S., & Reinsel, G. C. (1994). Regression models with spatially correlated errors. *Journal of the American Statistical Association*, *89*, 88-99.

Bloomfield, P., & Watson, G. S. (1975). The inefficiency of least squares. *Biometrika*, *62*, 88-116.

Cressie, N., & Hawkins, D. M. (1980). Robust estimation of the variogram: I. *Journal of the International Association for Mathematical Geology*, *12*, 115-125.

Charnes, A., Frome, E. L., & Yu, P.L. (1976). The equivalence of generalized least squares and maximum likelihood estimates in the exponential family. *Journal of the American Statistical Association*, *71*, 214-222

Diblasi, A., & Bowman, A. W. (2001). On the use of the variogram in checking for independence in spatial data. *Biometrics 57*, 211–218.

Diggle, P. J., Heagerty, P., Liang, K-Y, & Zeger, S. L. (2003). *Analysis of Longitudinal Data* (*2nd Ed.*). Oxford Statistical Science Series.

HCUP State Inpatient Databases (Sid). (2003-2004). *Healthcare Cost And Utilization Project* (*HCUP*). Agency for Healthcare Research and Quality, Rockville, MD. www.hcupus.ahrq.gov/sidoverview.jsp.

Jiang, H. J., Stocks, C., & Wong, C. J. (2006). Disparities between two common data sources on hospital nurse staffing. *Journal of Nursing Scholarship*, *38*, 187-193.

Kovner, C., Jones, C., Zhan, C., Gergen, P. J., & Basu, J. (2002). Nurse staffing and postsurgical adverse events: An analysis of administrative data from a sample of U.S. hospitals, 1990-1996. *Health Services Research*, *37*, 611-629.

Kovner, C., Mezey, M., & Harrington, C. (2000). Research priorities for staffing, case mix and quality of care in the U.S. nursing homes. *Journal of Nursing Scholarship*, *32*, 77-80.

Lee, J., & Lund, R. (2004). Revisiting simple linear regression with autocorrelated errors. *Biometrika*, *91*, 240-245.

Liang, K., & Zeger, S. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, *73*, 13-22.

Mardia, K. V., & Marshall, R. J. (1984). Maximum likelihood estimation of models for residual covariance in spatial regression. *Biometrika*, *71*,135-46.

Metersky, M. L., Ma, A., Bratzler, D. W., & Houck, P. M. (2004). Predicting bacteremia in patients with community-acquired pneumonia. *American Journal of Respiratory and Critical Care Medicine*, 169, 342-347.

Rosenblatt, R. A., Andrilla, C. H. A., Curtin, T., & Hart, L. G. (2006). Shortatges of medical personnel at community health centers. *Journal of the American Medical Association*, *295*, 1042-1049.

Waterer, G. W., & Wunderink, R. G. (2001). The influence of the severity of community-acquired pneumonia on the usefulness of blood cultures. *Respiratory Medicine*, 95, 78-82.