SEMIPARAMETRIC POISSON REGRESSION MODEL FOR CLUSTERED DATA

Erniel B. Barrios[1,2] Eiffel A. De Vera[3]
[1]University of the Philippines Diliman, Philippines
[3]Central Luzon State University, Philippines
[2]Corresponding Author: Erniel B. Barrios, ebbarrios@upd.edu.ph

Abstract

A semiparametric Poisson regression is proposed in modeling spatially clustered count data. The heterogeneous covariate effect across the clusters is formulated in the context of nonparametric regression while the random clustering effect is based on a parametric specification. We propose two estimation procedures: (1) the parametric and nonparametric parts are estimated simultaneously via penalized least squares; and (2) the parametric and nonparametric parts are estimated iteratively via the backfitting algorithm. The simulation study exhibited the advantages of these two methods over ordinary Poisson regression and an intrinsically linear model when the aggregate covariate effect is negligible. This happens when sensitivity to the covariate is minimal or the data-generating model is not linear. The two estimation methods are generally more advantageous over the traditional approaches when the linear model fit is poor. In cases where there is a good linear fit, the proposed methods are at par with the traditional methods, but the second approach can still be advantageous when there are several covariates involved since the backfitting algorithm yields computational simplicity in the estimation process.

Keywords: backfitting, generalized additive models, nonparametric regression, random effects

1. Introduction

Epidemiological and environmental studies usually produce variables resulting from counts of the number of times an event happened at a given time or space. Poisson regression analysis is commonly used in analyzing count data generated by such variables as it predicts the average value of the count variable conditional on one or more covariates. When the sample size is large, the central limit theorem is often invoked and classical linear regression is used in modeling the heterogeneous mean. This method is prone to bias and often inconsistent and inefficient because of the stringent assumptions of classical regression. For discrete variables, skewness, nonnegative values of the response, and heteroscedasticity often occurs, violating the assumptions of regression analysis. With malaria incidence data, Ruru and Barrios (2003) demonstrated that poisson regression has edge over classical regression in terms of parsimony since classical regression requires more variables in the equation to achieve as much fit as Poisson regression does (with fewer variables).

The mean of the count data Y is affected by explanatory variables (X_i's) and the heterogeneous Poisson model is $Y_i \sim P_0(\mu + X_i'\beta)$ or $logE(Y_i) = \mu + X_i'\beta$. The expected mean of the response variable in this model is heterogeneous and it depends on the explanatory variables. However, in phenomenon that exhibit spatial dependence like those in epidemiology, the cluster where the observation belongs can further contribute homogeneous effect on the response variable.

Observations in clusters may be correlated because of demographic similarity and other spatial dependency-inducing phenomena. Clusters must be considered in the model because membership within a cluster can have a significant contribution on Y_i. With parsimony in mind, Demidenko(2007) proposed a poisson regression that can account for cluster effect, i.e., $Y_{ij} \sim P_0(\mu_i + X'_{ij}\beta)$ where Y_{ij} refers to the jth observation in the ith cluster and μ_i is the cluster-specific intercept, a random component. The link $logE(Y_i|\mu_i) = \mu_i + X'_{ij}\beta$ implies that the random cluster-specific intercepts and the covariates with fixed coefficients jointly explain the heterogeneous means. However, this did not take into account the possibility that the covariates for each cluster could differ as illustrated in Ruru and Barrios(2003).

It is hypothesized that because of clustering, the effects of X_{ij} vary across the clusters. The model becomes $Y_i^k \sim P_0(\mu_k + X_{ij}^k \beta_j)$ and is highly vulnerable to overparametrization. We propose to resolve this issue by transforming the model into an additive combination of parametric and nonparametric specifications, i.e. $logE(Y_i^k|\mu_k) = \mu_k + f(X_{ij}^k)$. The cluster-specific intercept μ_k is formulated parametrically through the random effects while the covariates are specified in a nonparametric way. The semiparametric model is then estimated iteratively through the backfitting algorithm.

This study proposes an alternative modeling strategy for count clustered data using poisson regression. As clustering of data may cause values of intercepts and coefficients of the covariates of the outcome variable to vary, we provide a parsimonious semiparametric poisson regression model. Real life data are not easy to model not only because of its natural variability but also due to the heterogeneous effect of certain factors across groups of observations. The parametric part of the semiparametric model will take advantage of inherent homogeneity within clusters. While the nonparametric part will induce flexibility into the function to mitigate the overparameterization that can result when dynamic model is used instead.

Considering additivity of the postulated model, backfitting is proposed. This will be advantageous over simultaneous estimation of all the components when more than one predictor is involved. For two or more predictors, simultaneous estimation of the nonparametric functions requires thin plate smoothing splines whose convergence rate declines as the number of predictor increases further. This will not be a problem in backfitting since each term, including the nonparametric functions are estimated one at a time.

2. The Model

Our goal is explain count data Y_i^k in terms of covariates X_{ij}^k in n clusters, i = 1,…, n_k , j = 1,…, p, k = 1,…, n. Existence of clustering among the observations (Y_i^k is the ith observation in the kth cluster, k = 1,…, n, i = 1,…,n_k) implies that within the cluster, covariate effects are homogeneous, but between two different clusters, covariate effect could vary, and Poisson regression will not work. Clustered data are possibly endowed with spatial autocorrelations within the clusters. While observations across the clusters can still be independent, within the clusters they can be spatially correlated, i.e., nearby observations exhibit higher correlations than those farther from each other. It is also possible that correlations are homogeneous among observations within a cluster, e.g., in

epidemiological settings. This is accounted by allowing the covariates parameters to vary across the clusters. Consequently, this leads to substantial overparametrization leading to many constraints on either ordinary least squares or on maximum likelihood estimation. In lieu of varying parameters across the clusters, the effect of X_{ij}^k on Y_i^k is postulated nonparametrically. Relaxing the functional form of the effect of X_{ij}^k on Y_i^k is equivalent to varying effect across clusters.

Given n clusters with n_k elements each, k = 1, …, n, we postulate the model

$$logE\left(Y_i^k|\mu_k\right) = \mu_k + f\left(X_{ij}^k\right), \text{ i = 1,…, } n_k, \text{ j = 1, …, p} \qquad (1)$$

where:

n_k is the cluster size (constant or nonconstant across clusters)

n is the total number of clusters

μ_k is a random variable, $\mu_k = U_k + \varepsilon_k$, $\varepsilon_k \sim N(0, \sigma_\varepsilon^2)$, U_k is the cluster-specific intercept

X_{ij}^k is the value of jth covariate of the ith observation within the kth cluster

$f\left(X_{ij}^k\right)$ is a smooth function of X_{ij}^k (nonparametric)

Y_i^k is the ith value of the response variable within the kth cluster

The following assumptions are further considered:

1. Y_i^k is a count data, $Y_i^k \sim P_0\left(\mu_k + X_{ij}^k \beta_j\right)$
2. X_{ij}^k are the attributes of the observations that may be quantitative or qualitative
3. μ_k is a random cluster intercept addressing the effect of clustering
4. Degree of spatial autocorrelations among observations within the same cluster is assumed to be homogeneous, common in epidemiological settings where all individuals in a neighborhood are either vulnerable or not vulnerable to an epidemic threat.
5. Clusters are independent, implying that only elements within the cluster can exhibit spatial dependencies but elements between clusters are spatially independent.

Model (1) can be extended to more than one covariate as

$$logE\left(Y_i^k|\mu_k\right) = \mu_k + f_1\left(X_{i1}^k\right) + \cdots + f_{ip}\left(X_{ip}^k\right) \qquad (2)$$

where $f_1\left(X_{i1}^k\right), \ldots, f_{ip}\left(X_{ip}^k\right)$ are the smooth functions of the covariates.

The smooth functions of the X_{ij}^k's are used in lieu of varying coefficients of the predictors between clusters. Ruru and Barrios (2003) illustrates the premise of the postulated model where clustering was apparent and that the covariates vary from one cluster to another. Instead of the separate poisson regression model for each cluster, a semiparametric model is proposed. The term μ_k will take into account the random cluster effect and the term $f\left(X_{ij}^k\right)$ will take into account the varying risk factors for each cluster resulting to one parsimonious Poisson regression model.

Estimation Procedure

Taking advantage of the additive model, backfitting algorithm is used to estimate the parametric and nonparametric parts of Model (1). Two estimation procedures are proposed: Method 1 estimates the parametric and nonparametric parts simultaneously in a semiparametric context; and Method 2 estimates the nonparametric part first, and then the parametric part is estimated from the residuals (backfitting).

Semiparametric Estimation (Method 1)

The generalized additive model (GAM) in Model (1) or (2) makes viable the simultaneous estimation of the parametric and nonparametric components of the model, separating the linear from any general nonparametric trend during the estimation process. GAM fits the parametric linear model to account for the cluster effect and spline nonparametric function to estimate the nonparametric function $f(X_{ij}^k)$.

In Model (2), a thin plate smoothing can be used. It approximates smooth multivariate functions observed with noise. It allows greater flexibility in the form of the regression surface. It also uses the penalized least squares method to fit the data with a flexible model with the advantage that the multidimensional data could be used. Generalized Cross-Validation (GCV) is used as criteria for choosing the best smoothing parameter, see Hardle, et. al. (2004) for further details.

Backfitting of the Semiparametric Model (Method 2)

In Method 2, the parametric and nonparametric parts of the Model (1) are estimated separately in the context of backfitting. First, $f(X_{ij}^k)$ is estimated nonparametrically using spline smoothing. Then the partial residual $\hat{e}_i = \left(Y_i^k - \exp\left[\hat{f}\left(X_{ij}^k\right)\right]\right)$ is computed, this contains information on cluster effect and thus, used to estimate the cluster-specific intercept μ_k. The predicted value of the response variable is the sum of the estimates of the parameters and the nonparametric function, i.e., $\hat{Y}_i^k = \hat{U}_k + exp[\hat{f}(X_{ij}^k)]$.

The advantage of Method 2 over Method 1 is that it is more computationally simpler since the components are estimated one at a time. In Method 1, thin plate smoothing splines should be used once there are two or more predictors involved whose convergence rate decreases as the number of predictor increases.

3. Simulation Studies

A simulation study is conducted to evaluate the performance of semiparametric poisson regression model for clustered data. Each data set was composed of n clusters of size n_k, k = 1, …,n. Y_i^k was generated from the following:

$$logY_i^k = \mu_k + f(X_{ij}^k) + w\epsilon_i \qquad (3)$$

The constant w is used to induce misspecification error. The simulated data are used to compare the proposed semiparametric model (and two estimation procedures) to existing methods like the ordinary poisson regression and classical linear regression through their mean absolute percentage error (MAPE) and root mean square error (RMSE). See Table 1 for the summary of simulation settings.

Table 1. Boundaries of Simulation Study

1. Distribution of μ_k	$\mu_k \sim N(u_k, 2)$, $u_k = 5$, increases by 5 for the succeeding clusters $\mu_k \sim Po(u_k)$, $u_k = 2$, increases by 2 for the succeeding clusters
2. Number of clusters	5, 10, 20
3. Cluster size	Equal: 5, 10, 20 Unequal: small variation– randomly selected from 1 to 5; medium variation– randomly selected from 1 to 10; large

	variation– randomly selected from 1 to 20
4. Covariate	$X_{ij}^{k} \sim U(10,50)$
5. Form of $f(X_{ij}^{k})$	Linear, Exponential
6. Value of β in $f(X_{ij}^{k})$	0.10, 2
7. Error term	$\varepsilon \sim N(0, 1)$
8. Misspecification error	w =1(without), 5(with)

4. Results and Discussion

The predictive performance of the proposed semiparametric poisson regression model estimated using two procedures are compared with ordinary linear regression and ordinary poisson regression in simulated data.

Effect of Clustering

The simulation scenarios included varying cluster sizes as well as constant and non-constant elements within the cluster. The average MAPE values under equal cluster sizes shown in Table 2 illustrates the advantage of semiparametric and backfitting methods over ordinary poisson and linear regression for all levels of cluster size. The two estimation methods are also fairly robust to cluster size. Furthermore, in large cluster sizes, the backfitting method has a lower MAPE than the semiparametric method. Similar is true in unequal cluster size as shown in Table 3.

Table 2. MAPE for Equal Cluster Size

Cluster Size	Semiparametric Method	Backfitting Method	Poisson Regression	Linear Regression
Small (n_k = 5)	14.60	14.99	17.66	36.29
Medium (n_k = 10)	16.69	16.32	19.58	40.73
Large (n_k = 20)	16.40	15.09	19.24	39.99

Table 3. MAPE for Unequal Cluster Size

Cluster Size	Semiparametric Method	Backfitting Method	Poisson Regression	Linear Regression
Small (n_k from 1 to 5)	13.70	16.01	16.70	22.13
Medium (n_k from 1 to 10)	14.22	15.13	16.40	22.08
Large (n_k from 1 to 20)	16.74	16.26	18.51	26.28

The MAPE values (Table 4) of semiparametric, backfitting and ordinary Poisson regression are comparable but lower than those from ordinary linear regression for all number of clusters categories. As the number of clusters increases, MAPE of all methods decreases, confirming the observations of Arceneaux and Nickerson (2009) that adding clusters, and not increase of cluster size, leads to increase in efficiency in poisson regression. Also, backfitting method has lower MAPE than semiparametric method if the number of clusters is large.

Table 4. MAPE for Varying Number of Clusters

No. of Clusters	Semiparametric Method	Backfitting Method	Poisson Regression	Linear Regression
Small (n = 5)	16.22	17.18	19.32	36.93
Medium (n = 10)	15.59	15.77	17.97	31.25
Large (n = 20)	13.99	13.23	16.04	23.85

Covariate Effect

Given a linear function of the covariates, MAPE for all methods are comparable. Ordinary linear regression is advantageous when the linear effect of the covariate is dominant. The proposed methods however are at par with the existing methods even in cases where these existing methods are known to perform optimally. However, in an exponential function of the covariates, semiparametric, backfitting and Poisson regression are relatively advantageous over linear regression. As expected, the semiparametric and backfitting methods are more flexible in capturing nonlinear forms of the relationship between the covariates and the count data Y_i^k.

Model Misspecification

Misspecification was simulated using a constant multiplier to the error terms. Model misspecification usually leads to residuals that exhibit large variance. When a constant is multiplied into the simulated error terms, the bigger part of the variation in Y cannot be explained by the covariates. Whether there is misspecification or not, the two proposed methods are always superior over the Poisson and linear regression models, possibly indicating robustness of the proposed methods to misspecification errors, inherent advantages of nonparametric models.

5. Conclusions

The semiparametric Poisson regression model for spatially clustered count data given n clusters with n_k observations given by $logE\left(Y_i^k|\mu_k\right) = \mu_k + f\left(X_{ij}^k\right)$ can account for the varying coefficient of the covariates across the clusters. The simulated data exhibited the advantages of the model along with the two estimation methods over Poisson and linear regression models when the covariate effect is negligible, i.e., sensitivity to the covariate is minimal or the data-generating model is not linear. The two methods are generally advantageous over the traditional approaches when the linear model is inadequate. As the number of clusters increases, the predictive ability of the two methods also increases.

References:

Arceneaux, K. Nickerson, D., (2007) "Modeling certainty with clustered data: a comparison of methods," *Political Analysis*, 17, 177-190.

Demidenko, E., (2007) "Poisson regression for clustered data," *International Statistical Review*, 75, 96-113.

Hardle, W, Muller, M, Sperlich, S, Werwatz, A., (2004) *Nonparametric and Semiparametric Models*, Springer, Berlin.

Ruru, Y, Barrios, E., (2003) "Poisson regression models of malaria incidence in Jayapura, Indonesia," *The Philippine Statistician*, 52, 27-38.