Journal of Data Science,17(4). P. 660 - 673,2019
DOI:10.6339/JDS.201910_17(4).0002

GENERALIZED LINEAR DISTRIBUTED LAG MODELS

Hanh Nguyen[1], Qin Shao[2]

[1]*Department of Mathematics and Statistics*
[2]*The University of Toledo*

ABSTRACT

We propose distributed generalized linear models for the purpose of incorporating lagged effects. The model class provides a more accurate statistical measure of the relationship between the dependent variable and a series of covariates. The estimators from the proposed procedure are shown to be consistent. Simulation studies not only confirm the asymptotic properties of the estimators, but exhibit the adverse effects of model misspecification in terms of accuracy of model estimation and prediction. The application is illustrated by analyzing the presidential election data of 2016.

Keywords: Generalized linear distributed lag model, autoregressive time series, multicollinearity, model misspecification.

1. Introduction

We propose a new model class, which is motivated by incorporating lag effects of covariates on the dependent variable. Our paper aims at providing a more accurate statistical analysis for the relationship, for example, between the outcome of an event that occurs once every few years and the covariates that have observations every year. Lag effects have received a great deal of attention since Almon's paper, see Almon (1965). She proposed linear distributed lag models to model the dependence of time series on several regressors from a correlated sequence. In particular, she predicted quarterly capital expenditures in manufacturing industries from present and past appropriations. Greene (2002, Chapter19) is a comprehensive reference for a detailed summary of the development of this model class. Gasparrini et al (2010) extended the ideas to the generalized linear model setting where a dependent variable from an exponential family was linked to covariates from a time series. They modeled the relationship between the mortality count, which follows a Poisson distribution, and temperatures. Rushworth et al (2013) applied a distributed lag model to analyze hydrological data.

We propose to study generalized linear distributed lag models (GLDLM). The proposed model class is closely related to generalized linear models and longitudinal models with discrete response variables. However, the advantage of our model class is that it can be used to describe how much a link, which is function of the mean of a dependent variable y, is linearly explained by a sequence of random covariates $\{x_{t-k}; 0 \le k \le K\}$ and some other covariates $\{z_j\}_{j=1}^{J}$. The covariates $\boldsymbol{z} = (z_1, z_2, \dots, z_J)^T$ can be either fixed or random: in the paper, we assume they are random and independent, and they are also independent of $\boldsymbol{x} = (x_1, x_2, \dots, x_{t-K})^T$. Throughout the paper, we use lower case letters to denote both random variables and their realizations, when the meaning is obvious according to the context. The conditional density function of y belongs to an exponential family defined by

$$f(y \mid \mathbf{x}, \mathbf{z}; \eta) = \exp\{y\eta - b(\eta) + c(y)\}, \tag{1}$$

where the link $\eta = g(\mu)$ is function of $\mu = \mathrm{E}(\mathrm{y}|\mathbf{x}, \mathbf{z})$, $b(\eta)$ and $c(y)$ are respectively functions of η and y only, and the dependence of on the covariates $\boldsymbol{x}$ and $\boldsymbol{z}$ is through η. The density function $f(y \mid \mathbf{x}, \mathbf{z}; \eta)$ satisfies some regularity conditions, which can be found, for example, in Fahrmeir and Kaufmann (1985).

For the density function in (1), we focus on the canonical link function which is a function of the mean of the dependent variable and which is also a linear function of the covariates (c.f. McCullagh and Nelder (1989)):

$$\eta = \omega_0 + \sum_{k=0}^{K} \beta_{k+1} x_{t-k} + \sum_{j=1}^{J} \omega_j z_j . \tag{2}$$

A generalized linear model (GLM) is a special case of (2), in that a GLDLM becomes a GLM when $K = 0$. It happens that the dependent variable is observed only once while a covariate has several data points in the same period. To the best of our knowledge, there is no systematic discussion on analysis of such data, whereas many phenomena exhibit such a pattern. For instance, a college student who drops out due to poor academic performance has usually been

struggling for several semesters; a patient usually has been suffering high blood pressure for a while before a stroke happens; a presidential election happens every four years while some variables that could have a significant impact on the outcome have observations every year. If the model just includes one variable in the sequence, our simulation studies in Section 3 show that the results could be very misleading. The statistical analysis for such data is much more informative and provide a more accurate prediction if the information in previous years is included.

However, it is known that two highly correlated covariates could result in multicollinearity. The adverse consequence of multicollinearity includes, for example, instability of parameter estimates and distortion of standard errors. Cheng and Wu (2006) tackled multicollinearity using partial least squares regression an Kutner et al (2005) is a good reference for this issue in linear regression models. Many researchers have paid attention to the issue and have proposed some methods to overcome it for generalized linear models. For example, Mackinnon and Puterman (1989) proposed a method to detect it, Shen and Gao (2008) provided a solution by a double penalized maximum likelihood approach, and Huanga et al (2016) attempted to solve the problem by a new collinearity diagnostic tool based on variance inflation factor. We utilize the built-in correlated structure of the sequence and propose an estimation procedure so that the estimators not only possess asymptotic normality under very general conditions, but they quantify more accurately how much the dependent variable relies on a covariate in the past as well. Sometimes such information is important for the purpose of avoiding negative outcomes.

The advantages of the new model class are explored by simulation studies, which show that taking the correlated covariates into account can improve accuracy of both prediction and estimation substantially. In most cases of the simulations, the reductions of the standard errors of the estimates and the increases of the relative frequencies of making correct prediction are not trivial.

The paper is organized as follows. The estimation procedure of the model parameters along with theoretical justifications is discussed in Section 2; the intensive simulation studies and data analysis are presented in Sections 3-4; the paper is concluded by remarks in Section 5. The proofs for asymptotics of the estimators in Section 2 is given in the appendix.

2. Estimation for Model Coefficients

We assume that $\{x_t\}$ is an autoregressive time series with order one (AR(1)) satisfying

$$x_t - \phi x_{t-1} = \epsilon_t$$

where $\{\epsilon_t\}$ is independent and identically distributed white noise with $E(\epsilon_t) = 0$ and $var(\epsilon_t) = \sigma^2$. Denote $m = K + J + 2$. Note that if the true value ϕ_0 is known, (2) can be written as

$$\eta(\theta) \coloneqq \theta_0 + \sum_{k=0}^{K-1} \theta_{k+1} \epsilon_{t-k} + \theta_{K+1} x_{t-K} + \sum_{j=1}^{J} \theta_{K+1+j} z_j = \mathbf{d}^\top \theta \tag{3}$$

where $\eta(\theta)$ is used to emphasize the dependence of η on $\theta = (\theta_0, \dots, \theta_{m-1})^T$, the $m \times 1$ vectors $d = (1, d_1, \dots, d_{m-1})^T = \left(1, \epsilon_t, \dots, \epsilon_{t-K+1}\ ,\ x_{t-K}, z_1, \dots, z_J\right)^T$ and the parameter θ is determined by $\{\omega_0, \phi, \beta_1, \dots, \beta_{K+1}, \omega_1, \dots, \omega_J\}$. In particular, $\beta_1 = \theta_1$ and for $1 \le k \le K$,

$$\theta_{k+1} = \beta_{k+1} + \phi\theta_k = \sum_{j=0}^{k} \phi^j \beta_{k+1-j}.$$

For the canonical link function, it is known that the density function in (1) implies that the conditional mean of y is $\mu(\theta) = \frac{\partial b(\eta(\theta))}{\partial \eta}$ and the conditional variance $V(\theta) = \frac{\partial^2 b(\eta(\theta))}{\partial \eta^2} = \frac{\partial \mu(\theta)}{\partial \eta}$ given $\{x, z\}$, and thus they depend on the parameter θ.

Compared with (2), model (3) has several advantages. First, (2) confounds the dependence of the link on x_{t-K} due to the correlation of x_{t-K} and $\{x_{t-K}\}_{k=0}^{K}$, whereas (3) reveals how much η relies on x_{t-K}; secondly, the estimates of (3) are much more stable since the design matrix does not have multicollinearity; thirdly, the estimators of (3) are efficient as illustrated later --- they are asymptotically normally distributed.

Suppose there are n independent observations $\{y_i, x_i, z_i\}_{i=1}^{n}$ with $x_i = \left(x_{i,t}, \dots, x_{i,t-K}\right)^T$ and $\boldsymbol{z}_i = \left(z_{i,t}, \dots, z_{i,t-K}\right)^T$, and the canonical link function for the i-th subject is

$$\eta_i(\theta) = \theta_0 + \sum_{k=0}^{K-1} \theta_{k+1}\epsilon_{i,t-k} + \theta_{K+1}x_{i,t-K} + \sum_{j=1}^{J} \theta_{K+1+j}z_{i,j}.$$

Define the $n \times m$ design matrix $\boldsymbol{D} = \left(D_{i,j}\right)$ as follows:

$$\boldsymbol{D} = \begin{pmatrix} 1 & \epsilon_{1,t} & \cdots & \epsilon_{1,t-K+1} & x_{1,t-K} & z_{1,1} & \cdots & z_{1,J} \\ 1 & \epsilon_{2,t} & \cdots & \epsilon_{2,t-K+1} & x_{2,t-K} & z_{2,1} & \cdots & z_{2,J} \\ \vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots \\ 1 & \epsilon_{n,t} & \cdots & \epsilon_{n,t-K+1} & x_{n,t-K} & z_{n,1} & \cdots & z_{n,J} \end{pmatrix}. \tag{4}$$

The maximum likelihood estimates (MLE) $\tilde{\theta}$ for θ can be obtained by maximizing the following objective function $l(\theta)$,

$$l(\theta) = \sum_{i=1}^{n} \left\{ y_i\eta_i(\theta) - b\left(\eta_i(\theta)\right) + c(y_i) \right\} \tag{5}$$

The objective function depends on the linear coefficients θ through $\{\eta_i\}_{i=1}^{n}$. Since (5) is not a linear function of the parameters θ, and there is no explicit formula for $\tilde{\theta}$, the calculation of $\tilde{\theta}$ relies on numerical methods, such as the Newton-Raphson algorithm.

Assumptions:

1. The parameter space $\Theta \subset \mathbb{R}^m$ is an open compact set, and the true value $\theta = (\theta_0, \dots, \theta_{m-1})^T$ is an interior point of Θ.
2. The expectation $E(V(\theta)\boldsymbol{d}\boldsymbol{d}^{\boldsymbol{T}})$ exists and positive definite, the expectation $E_{sup_{\theta\in\Theta_N}}||V(\theta)\boldsymbol{d}\boldsymbol{d}^{\boldsymbol{T}}||)$ exists for a compact neighborhood Θ_N of θ_0, where $\|\cdot\|$is a matrix norm.
3. The autoregressive time series $\{x_t\}$ is causal stationary with a continuous spectral density function; that is, $\sum_{k=0}^{\infty} r(k) < \infty$, where $r(k) = cov(x_t, x_{t-k})$.

The assumption (1) is very typical and ensures the asymptotical normality of the maximum likelihood estimator $\tilde{\theta}$. The assumption (2) implies the consistence of estimators for the linear regression coe cients. The assumption (3) is needed for consistency of $\hat{\phi}$ in (6). The details can be found, for example, in Brockwell and Davis (2002).

Theorem 1 Under the assumptions (1)-(2), the maximum likelihood estimators $\tilde{\theta}$ is asymptotically normally distributed. That is as $n \to \infty$,

$$\sqrt{n}(\tilde{\theta} - \theta_0) \xrightarrow{D} N(0, \Gamma^{-1}),$$

where the (j,k)-entry of the $m \times m$ matrix Γ is $E\left(d_j d_k V(\theta)\right)$ which is the information matrix in one sample.

The proof is straightforward and is given in the appendix for interested readers. The difficulty of the above approach is that $\{\epsilon_{i,t}\}$ in (3) is not observable. Thus we propose a two-step method, in which the error term $\epsilon_{i,t}$ is replaced by the residual $\hat{\epsilon}_{i,t}$. In particular, we follow the approach of Lund et al (2016) and estimated the $AR(1)$ coefficient by:

$$\hat{\phi} = \frac{(K+1)\sum_{i=1}^{n}\sum_{k=0}^{K-1} x_{i,t-k}x_{i,t-k+1}}{K\sum_{i=1}^{n}\sum_{k=0}^{K-1} x_{i,t-k}^2} \tag{6}$$

which under the assumption (3) is asymptotically normally distributed

$$\sqrt{n}(\hat{\phi} - \phi_0) \xrightarrow{D} N(0, 1 - \phi_0^2)$$

where ϕ_0 is the true value. Then the residual of $AR(1)$ $\{\epsilon_{i,t}, 1 \le i \le n\}$ is

$$\hat{\epsilon}_{i,t} = x_{i,t} - \hat{\phi}x_{i,t-1}$$

The MLE $\hat{\theta}$ maximizes the objective function (5) with $\{\epsilon_{i,t}\}$ replaced by $\{\hat{\epsilon}_{i,t}\}$

$$\hat{l}(\theta) = \sum_{i=1}^{n}\{y_i\hat{\eta}_i(\theta) - b(\hat{\eta}_i(\theta)) + c(y_i)\}$$

where

$$\hat{\eta}_i(\theta) = \theta_0 + \sum_{k=0}^{K-1}\theta_{k+1}\hat{\epsilon}_{i,t-k} + \theta_{K+1}\hat{\epsilon}_{i,t-K} + \sum_{j=1}^{J}\theta_{K+1+j}z_{i,j}.$$

The MLE $\hat{\theta}$ is different from $\tilde{\theta}$, however. The following theorem indicates that $\hat{\theta}$ is consistent for the true value θ_0.

Theorem 2 under the assumption (1)-(3), as $n \to \infty$,$\hat{\theta}$ converges to the true value θ_0 in probability; that is, $\hat{\theta} - \theta_0 \xrightarrow{P} 0$.

3. Simulation Studies

Simulations are conducted on binary and Poisson data with a variety of AR(1) models for sample sizes from $n = 100$ to $n = 2000$. The coe cients of AR(1) range from as low as $|\phi_0| = 0.2$ to as high as $|\phi_0| = 0.8$, and the white noise is from the standard normal distribution. Table 1 includes the true values of $(\beta_1, \beta_2, \beta_3, \beta_4)$ along with the corresponding true values $(\theta_{0,1}, \theta_{0,2}, \theta_{0,3}, \theta_{0,4})$ used in the simulations for the models with the lag $K = 3$. It is known that the canon-ical link functions are respectively $\eta = \log\left\{\frac{\pi}{1-\pi}\right\}$ for a Bernoulli distribution with the probability $P(y = 1) = \pi$, and $\eta = \log\mu$ for a Poisson distribution with $E(y) = \mu$. We simulate 1000 sample paths for each model. All the calculations are carried out by the free statistical computing and graphics environment *R* (2015).

Table 1: True Values of Parameters

ϕ_0	Binary Data				Poisson Data			
	$\beta_1 = 15$	$\beta_2 = 10$	$\beta_3 = 2$	$\beta_4 = 1$	$\beta_1 = 1.5$	$\beta_2 = 1$	$\beta_3 = 0.2$	$\beta_4 = 0.1$
	$\theta_{0,1}$	$\theta_{0,2}$	$\theta_{0,3}$	$\theta_{0,4}$	$\theta_{0,1}$	$\theta_{0,2}$	$\theta_{0,3}$	$\theta_{0,4}$
-0.8	15.000	-2.000	-3.600	-1.880	1.500	-0.200	0.3600	-0.188
-0.4	15.000	4.000	0.400	0.840	1.500	0.400	0.040	0.084
-0.2	15.000	7.000	0.600	0.880	1.500	0.700	0.060	0.088
0.2	15.000	13.000	4.600	1.920	1.500	1.300	0.460	0.192
0.4	15.000	16.000	8.400	4.360	1.500	1.600	0.840	0.436
0.8	15.000	22.000	19.600	16.680	1.500	2.200	1.960	1.6680

The simulation results for the estimates are summarized in Tables 2-3. The estimates $\hat{\theta}$ approach the true values with smaller and smaller sample standard errors as the sample size increases, which is in agreement with Theorem 2. Such asymptotic accuracy of estimators is also illustrated by the decreasing size of the boxes in Figure 1. It is tempting to conclude that the estimators are efficient or they have the same asymptotical distribution as $\tilde{\theta}$, which needs to be verified. Such a property of estimators are termed oracle in, for example, Shao and Yang (2017).

In addition, Tables 2-3 also include the information for the impact of model misspecification. In particular, the true model or M_1 in (7) contains $\{x_{t-k}\}_{k=0}^{3}$, while the fitted model or M_2 in (8) is mis-specified as only one covariate x_t:

$$M_1: \quad \eta = \omega_0 + \beta_1 x_t + \beta_2 x_{t-1} + \beta_3 x_{t-2} + \beta_4 x_{t-3} \tag{7}$$

$$M_2: \quad \eta = \omega_0 + \beta_1 x_t \tag{8}$$

Table2: Parameter Estimates and Standard Errors for Binary Data

ϕ_0	n	M_1				M_2
		$\hat{\theta}_{0,1}$	$\hat{\theta}_{0,2}$	$\hat{\theta}_{0,3}$	$\hat{\theta}_{0,4}$	$\hat{\theta}_1$
−0.8	100	12.637 ± 4.393	−1.679 ± 1.026	3.017 ± 1.408	−1.597 ± 0.880	1.438 ± 0.279
	500	14.968 ± 2.645	−2.004 ± 0.509	3.597 ± 0.732	−1.877 ± 0.428	1.436 ± 0.126
	1000	14.977 ± 1.806	−1.990 ± 0.354	3.596 ± 0.496	−1.869 ± 0.293	1.432 ± 0.091
	2000	14.929 ± 1.232	−1.991 ± 0.237	3.585 ± 0.338	−1.873 ± 0.209	1.440 ± 0.063
−0.4	100	13.009 ± 4.434	3.457 ± 1.529	0.355 ± 0.747	0.715 ± 0.711	2.012 ± 0.413
	500	15.183 ± 2.581	4.079 ± 0.834	0.400 ± 0.323	0.859 ± 0.319	1.995 ± 0.176
	1000	14.914 ± 1.684	3.984 ± 0.566	0.401 ± 0.210	0.839 ± 0.209	1.996 ± 0.121
	2000	15.030 ± 1.236	4.017 ± 0.410	0.410 ± 0.149	0.838 ± 0.153	1.995 ± 0.087
−0.2	100	12.794 ± 4.954	5.971 ± 2.438	0.542 ± 0.778	0.738 ± 0.747	2.256 ± 0.448
	500	15.072 ± 2.806	7.040 ± 1.388	0.631 ± 0.346	0.882 ± 0.353	2.255 ± 0.205
	1000	15.045 ± 1.802	7.043 ± 0.901	0.610 ± 0.233	0.886 ± 0.235	2.256 ± 0.140
	2000	15.003 ± 1.214	6.992 ± 0.619	0.604 ± 0.162	0.886 ± 0.165	2.258 ± 0.099
0.2	100	10.985 ± 4.059	9.467 ± 3.515	3.320 ± 1.554	1.433 ± 0.870	2.774 ± 0.579
	500	14.987 ± 3.306	12.992 ± 2.934	4.616 ± 1.172	1.932 ± 0.576	2.739 ± 0.233
	1000	14.958 ± 2.019	12.956 ± 1.790	4.593 ± 0.725	1.924 ± 0.368	2.748 ± 0.166
	2000	14.989 ± 1.276	12.998 ± 1.142	4.602 ± 0.470	1.923 ± 0.245	2.743 ± 0.113

0.4	100	9.889 ± 3.510	10.450 ± 3.601	5.470 ± 2.116	2.806 ± 1.229	2.973 ± 0.614
	500	15.000 ± 3.565	15.959 ± 3.734	8.350 ± 2.012	4.340 ± 1.106	2.982 ± 0.253
	1000	15.000 ± 2.176	15.970 ± 2.313	8.362 ± 1.289	4.342 ± 0.716	2.987 ± 0.177
	2000	15.002 ± 1.495	15.984 ± 1.602	8.385 ± 0.894	4.351 ± 0.508	2.985 ± 0.125
0.8	100	6.043 ± 2.109	8.811 ± 2.970	7.802 ± 2.661	6.580 ± 2.217	3.512 ± 0.908
	500	14.967 ± 5.815	21.946 ± 8.577	19.528 ± 7.738	16.596 ± 6.589	3.493 ± 0.352
	1000	15.252 ± 3.345	22.349 ± 4.844	19.932 ± 4.354	16.941 ± 3.721	3.472 ± 0.232
	2000	15.060 ± 2.051	22.093 ± 3.009	19.689 ± 2.698	16.761 ± 2.309	3.492 ± 0.165

It is worth pointing out that for the true models M_1, when the sample size is as small as 500, the estimates are very close to the true values for all $AR(1)$ models. On the other hand, for the mis-specified models M_2, the discrepancy of the true value of $\theta_{0,1}$ and the estimate $\hat{\theta}_1$ could be dramatic even when the sample size is 2000. For example, for binary data $\hat{\theta}_1 = 14.929$ in the true model M_1, and $\hat{\theta}_1 = 1.440$ in the mis-specified model M_2 for $\phi_0 = -0.8$ at $n = 2000$. While the true value $\theta_{0,1} = 15$.

Table 3: Parameter Estimates and Standard Errors for Poisson Data

ϕ_0	n	M_1				M_2
		$\hat{\theta}_1$	$\hat{\theta}_2$	$\hat{\theta}_3$	$\hat{\theta}_4$	$\hat{\theta}_1$
−0.8	100	1.507 ± 0.078	−0.201 ± 0.079	0.365 ± 0.078	−0.190 ± 0.060	0.763 ± 0.143
	500	1.501 ± 0.028	−0.201 ± 0.033	0.360 ± 0.032	−0.188 ± 0.026	0.771 ± 0.079
	1000	1.499 ± 0.018	−0.199 ± 0.023	0.361 ± 0.023	−0.188 ± 0.018	0.774 ± 0.064
	2000	1.500 ± 0.012	−0.200 ± 0.016	0.360 ± 0.015	−0.189 ± 0.012	0.774 ± 0.046
−0.4	100	1.502 ± 0.075	0.396 ± 0.102	0.044 ± 0.069	0.083 ± 0.062	1.099 ± 0.219
	500	1.499 ± 0.029	0.400 ± 0.046	0.042 ± 0.026	0.083 ± 0.025	1.114 ± 0.121
	1000	1.501 ± 0.019	0.401 ± 0.031	0.040 ± 0.019	0.083 ± 0.016	1.121 ± 0.093
	2000	1.500 ± 0.013	0.400 ± 0.022	0.040 ± 0.013	0.084 ± 0.012	1.121 ± 0.073
−0.2	100	1.502 ± 0.076	0.698 ± 0.110	0.066 ± 0.068	0.087 ± 0.062	1.266 ± 0.252
	500	1.500 ± 0.024	0.698 ± 0.044	0.061 ± 0.026	0.086 ± 0.024	1.293 ± 0.151
	1000	1.500 ± 0.017	0.699 ± 0.031	0.060 ± 0.018	0.087 ± 0.017	1.300 ± 0.120
	2000	1.500 ± 0.012	0.701 ± 0.022	0.060 ± 0.012	0.088 ± 0.011	1.302 ± 0.082
0.2	100	1.505 ± 0.058	1.295 ± 0.103	0.456 ± 0.102	0.194 ± 0.066	1.655 ± 0.370
	500	1.500 ± 0.019	1.298 ± 0.043	0.459 ± 0.045	0.193 ± 0.026	1.671 ± 0.208
	1000	1.500 ± 0.013	1.301 ± 0.031	0.461 ± 0.032	0.193 ± 0.019	1.681 ± 0.180
	2000	1.499 ± 0.009	1.300 ± 0.021	0.460 ± 0.023	0.192 ± 0.014	1.696 ± 0.147
0.4	100	1.502 ± 0.049	1.589 ± 0.095	0.829 ± 0.124	0.433 ± 0.101	1.857 ± 0.465
	500	1.500 ± 0.015	1.598 ± 0.040	0.838 ± 0.056	0.436 ± 0.045	1.898 ± 0.284
	1000	1.500 ± 0.010	1.599 ± 0.028	0.840 ± 0.039	0.437 ± 0.031	1.907 ± 0.256
	2000	1.500 ± 0.006	1.599 ± 0.020	0.838 ± 0.028	0.435 ± 0.022	1.906 ± 0.195
0.8	100	1.500 ± 0.014	2.191 ± 0.050	1.941 ± 0.109	1.645 ± 0.147	2.381 ± 0.935
	500	1.500 ± 0.002	2.199 ± 0.021	1.957 ± 0.048	1.665 ± 0.065	2.354 ± 0.597
	1000	1.500 ± 0.001	2.199 ± 0.015	1.959 ± 0.033	1.666 ± 0.046	2.381 ± 0.627
	2000	1.500 ± 0.001	2.199 ± 0.010	1.959 ± 0.023	1.667 ± 0.032	2.375 ± 0.517

We also demonstrate the prediction power of the true models and the mis-specified models by misclassification which occurs when the fitted model fails to correctly identify the true category. We compare the estimated probabilities at the cutoff

$\pi = 0.5$ for binary data. We study the performance of M_1 in (7) and M_2 in (8) via three indicators: the true positive rate (TPR) or the relative frequency of being correctly identified as $y = 1$; the true negative rate (TNR) or the relative frequency of being correctly identified as $y = 0$; the accuracy (ACC) or the relative frequency of correct predictions. We randomly divide the observations into two parts: the training data set which consists of 70% of the observations, and the test data set. We t a model using the training data, and then calculate the three indicators (TPR, TNR and ACC) using the test data. Table 4 shows that for all values of ϕ_0, all the three indicators of M_1 are larger than 95% and are at least 10% higher than those of M_2. For each value of ϕ_0, we also compare M_1 and M_2 by the receiver operating characteristic (ROC) curves based on one sample path of size $n = 1000$. An ROC curve is obtained by the true positive rates and the false positive rates at various cutoff values of . All the ROC curves in Figure 2 from M_1 are above those from M_2, which implies that M_1 is better than M_2.

Table4: Prediction Accuracy for Binary Data

ϕ_0	n	M1			M2		
		TPR	TNR	ACC	TPR	TNR	ACC
−0.8	500	0.964	0.970	0.962	0.800	0.806	0.800
	1000	0.964	0.977	0.964	0.800	0.811	0.800
−0.4	500	0.963	0.973	0.963	0.785	0.792	0.784
	1000	0.964	0.973	0.964	0.790	0.794	0.788
−0.2	500	0.964	0.989	0.964	0.794	0.813	0.793
	1000	0.966	0.968	0.966	0.796	0.797	0.796
0.2	500	0.970	0.988	0.971	0.820	0.837	0.821
	1000	0.970	0.988	0.971	0.820	0.837	0.821
0.4	0.844	0.974	0.988	0.975	500	0.845	0.854
	1000	0.975	0.985	0.975	0.844	0.851	0.843
0.8	500	0.985	0.997	0.984	0.905	0.915	0.904
	1000	0.986	0.988	0.986	0.905	0.907	0.905

4. Real Data Application

We illustrate the application of the proposed method by modeling the relationship between the presidential election outcome and the unemployment rate. A presidential election is held once every four years, while the factors that have impact on it have observations every year. It is reasonable to think that the performance of the first term of the president Obama from 2009-2012 did not a affect the choice of a voter between Hillary Clinton and Donald Trump much, whereas his second term from 2013-2016 played a role in a voter's decision. The county election data in Figure 3 was downloaded from https://public.opendatasoft.com, and the unemployment data from https://www.census.gov. The binary dependent variable y is the final election outcome of a county in 2016 ($y = 1$ if Hillary Clinton was voted for). The covariate x_t is the difference between the log transformed unemployment rate (UR_{year}) (year = 2016; 2015; ⋯ ; 2012) of adjacent two years and can be considered as the relative change of the unemployment rate. That is,

$$x_t = \log UR_{2016} - \log UR_{2015},$$
$$x_{t-1} = \log UR_{2015} - \log UR_{2014},$$
$$x_{t-2} = \log UR_{2014} - \log UR_{2013},$$
$$x_{t-3} = \log UR_{2013} - \log UR_{2012},$$

The $AR(1)$ cefficient estimate is $\hat{\phi} = 0.289$, and the estimates are by the relative change of the unemployment rate as early as in the year of 2013, which was the first year of Obama's second term. According to the analysis, the adverse impact of the factor was as much as -3.619. There could be some other economic factors and demographic factors that could have contributed to the outcome, and thus a more comprehensive analysis should be conducted.

5. Concluding Remarks

In this paper, we have extended the idea of generalized linear models so as to include lagged effects. We utilized the built-in correlated structure of covariates to estimate the otherwise confounded effects on the dependent variable, and proposed an estimation procedure for model coefficients along with the theoretical justification for an $AR(1)$ sequence. Such a model could be a candidate for analysis of the relation between the dependent variable and a sequence of covariates. Our simulations indicated that the lagged effects are important and model misspecification could lead to misinterpretations of model coefficients and inaccurate predictions. We focused on canonical link functions and our conjecture is that similar results can be obtained for some other link functions. Our limited simulations for $AR(2)$ which are not included suggested that GLDM might work well for higher order autoregressions.

6. Appendix

We use C to denote a constant, the value of which varies according to the context. We characterize the magnitude of a random variable by $O_p(\cdot)$ if it is bounded in probability and $O_p(\cdot)$ if it is of smaller order in probability. The detailed definitions can be found, for example, in Fuller (1995). Define the $m \times 1$ vector of the score functions $\nabla l_j(\theta) = \partial l(\theta)/\partial\theta$ with the j-th entry being

$$\nabla l_j(\theta) = \sum_{i=1}^{n} D_{i,j}(y_i - u_i(\theta)), \tag{9}$$

where $\mu_i(\theta) = \partial b(\eta_i(\theta))/\partial\eta$ and the $m \times m$ Hessian matrix $\nabla^2 l(\theta) = \partial^2 l(\theta)/\partial\theta\partial\theta^T$ with the (j,k)-th entry being

$$\nabla^2 l_{j,k}(\theta) = -\sum_{i=1}^{n} D_{i,j} D_{i,k} V_i(\theta), \tag{10}$$

where $V_i(\theta) = \frac{\partial^2 b(\eta_i(\theta))}{\partial\eta^2} = \partial\mu_i/\partial\eta$. Note that for any fixed $0 \le i \le n$ and $1 \le j,k \le m$,

$$E\left\{\sum_{i=1}^{n} D_{i,j}(y_i - u_i)\right\} = 0$$
$$E\left\{\sum_{i=1}^{n} D_{i,j} D_{i,k} V_i(\theta)\right\} = E(d_j d_k V(\theta))$$

Hence the assumptions (1) and (2) imply as $n \to \infty$

$$\frac{1}{n}\sum_{i=1}^{n} D_{i,j}\, D_{i,k} V_i(\theta_0) \xrightarrow{P} E(d_j d_k V(\theta_0))$$

and

$$\frac{1}{\sqrt{n}} \nabla l(\theta_0) \xrightarrow{p} N(0, \Gamma^{-1})$$

The Taylor expansion implies that

$$\sqrt{n}\left\{-\frac{1}{n}\nabla^2 l(\theta^*)\right\}\left(\tilde{\theta} - \theta_0\right) = \frac{1}{\sqrt{n}} \nabla l(\theta_0)$$

Proof of Theorem 1. Theorem 1 is the direct result of Fahrmeir and Kaufmann (1985),

$$\frac{1}{\sqrt{n}}\left(\tilde{\theta} - \theta_0\right) \xrightarrow{p} N(0, \Gamma^{-1})$$

The proof is complete.

Note that (9) and (10) respectively become with $\epsilon_{i,t}$ replaced by $\hat{\epsilon}_{i,t}$

$$\nabla \hat{l}_j(\theta) = \sum_{i=1}^{n} \widehat{D}_{i,j}\left(y_i - \hat{u}_i(\theta)\right)$$

and

$$\nabla^2 \hat{l}_{j,k}(\theta) = -\sum_{i=1}^{n} \widehat{D}_{i,j}\widehat{D}_{i,k} V_i(\theta)$$

with $\hat{u}_i(\theta) = \partial b(\hat{\eta}_i(\theta))/\partial\eta$ and $\hat{\eta}_i(\theta) = \sum_{j=0}^{m-1} \widehat{\mathrm{D}}_{i,j+1}\, \theta_j$. Furthermore, $\widehat{\mathrm{D}}_{i,j+1} - D_{i,j+1}$ is $\left(\phi_0 - \hat{\phi}\right)x_{i,t-j}$ for $1 \le j \le K$ and 0 otherwise. Therefore, it is straight forward that for $1 \le j \le K$,

$$\textstyle\sum_{i=1}^{n}\left(\widehat{\mathrm{D}}_{\mathrm{i,j+1}} - \mathrm{D}_{\mathrm{i,j+1}}\right)^2 = \sum_{i=1}^{n}\left(\phi_0 - \hat{\phi}\right)^2 x_{i,t-j}^2 = O_p\,(1),$$

and for $\delta = 1,2$,

$$\sup_{\theta}\left|\sum_{i=1}^{n}\left(\hat{\eta}_i(\theta) - \eta_i(\theta)\right)^{\delta}\right|$$

$$= \sup_{\theta}\left|\textstyle\sum_{i=1}^{n}\left\{\sum_{j=2}^{K+1}\theta_j\left(\widehat{\mathrm{D}}_{i,j} - \mathrm{D}_{i,j}\right)\right\}^{\delta}\right|$$

$$= \left|\phi_0 - \hat{\phi}\right|^{\delta} \sup_{\theta}\left|\sum_{i=1}^{n}\left\{\sum_{j=2}^{K+1}\theta_j x_{i,t-j+1}\right\}^{\delta}\right| = O_p(1) \tag{11}$$

Proof of Theorem 2. Notice that the Taylor expansions and (11) imply that

$$\sup_{\theta}\left|\sum_{i=1}^{n}\left(b(\hat{\eta}_i(\theta)) - b(\eta_i(\theta))\right)\right| = \sup_{\theta}\left|\sum_{i=1}^{n}\frac{\partial^2 b\left(\eta_i^*\right)}{\partial\eta^2}\left(\hat{\eta}_i(\theta) - \eta_i(\theta)\right)^2\right| = O_p(1)$$

where η_i^* is between $\hat{\eta}_i(\theta)$ and η_i. Therefore, we conclude

$$\sup_{\theta} n^{-1}|\hat{l}(\theta) - l(\theta)| = n^{-1}\sup_{\theta}\left|\sum_{i=1}^{n}\left\{y_i\hat{\eta}_i(\theta) - b\left(\hat{\eta}_i(\theta)\right) - y_i\eta_i(\theta) + b(\eta_i(\theta))\right\}\right|$$

$$\leq n^{-1}\sup_{\theta}\left|\sum_{i=1}^{n} y_i\{\hat{\eta}_i(\theta)-\eta_i(\theta)\}\right| + n^{-1}\sup_{\theta}\left|\sum_{i=1}^{n}\{b(\eta_i(\theta))-b(\hat{\eta}_i(\theta))\}\right|$$

$$\leq n^{-1}\sup_{\theta}\left(\sum_{i=1}^{n} y_i^2\right)^{\frac{1}{2}}\left[\sum_{i=1}^{n}\{\hat{\eta}_i(\theta)-\eta_i(\theta)\}^2\right]^{\frac{1}{2}} + n^{-1}\sup_{\theta}\left|\sum_{i=1}^{n}\{b(\eta_i(\theta))-b(\hat{\eta}_i(\theta))\}\right|$$

$$= O_p(1)$$

which implies $\hat{\theta}-\theta_0 \xrightarrow{P} 0$ according to Amemiya (1985). The proof is complete.

References

[1] Almon, S. (1965). The distributed lag between capital appropriations and expenditures. Econometrica 33, 178–196.

[2] Brockwell, P. and Davis, R. (1991) Time Series Theory Methods (2nd), Springer. Cheng, B. and Wu, X. (2006) An Modified PLSR Method in Prediction. Journal of Data Science 4, 257-274.

[3] Cheng, B. and Wu, X. (2006) An Modified PLSR Method in Prediction. Journal of Data Science 4,257-274.

[4] Fahrmeir, L. and Kaufmann, H. (1985) Consistency and asymptotic normality of the maximum likelihood estimator in generalized linear models. The Annals of Statistics 13, 342-368.

[5] Fuller, A. W. (1995) Introduction to Statistical Time Series (2nd), Wiley.

[6] Greene, W. H. (2002) Econometric Analysis (5th ed), Pearson, New York.

[7] Gasparrini, A., Armstrong, B. and Kenward, M. G. (2010) Distributed lag non- linear models. Statistics in Medicine 29, 2224-2234.

[8] Huanga, C. L., Joub, Y. and Choa, H. (2016) A new multicollinearity diagnostic for generalized linear models. Journal of Applied Statistics 43, 2029–2043.

[9] Kutner, M., Nachtsheim, C. J., Neter, J. and Li, W. (2005) Applied Linear Statistical Models (5th), McGraw-Hill, Irwin.

[10] Lund, R., Liu, G., and Shao, Q. (2016) A new approach to ANOVA methods for autocorrelated data. The American Statistician 70, 55-62.

[11] Mackinnon, M. J. and Puterman, M. L. (1989) Collinearity in generalized linear models. Communications in Statistics: Theory and Methodology 18, 3463- 3472.

[12] McCullagh, P. and Nelder, J. A. (1989) Generalized Linear Models (2nd), Chap- man & Hall/CRC Monographs on Statistics & Applied Probability.

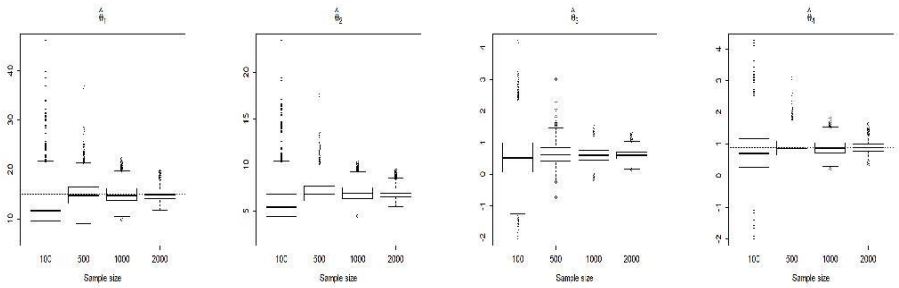
[13] R Core Team (2015) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org.

[14] Rushworth, A. M., Bowman, A. W., Brewer, M. J. and Langan, S. J. (2013) Distributed lag models for hydrological data. Biometrics 69, 537–544.

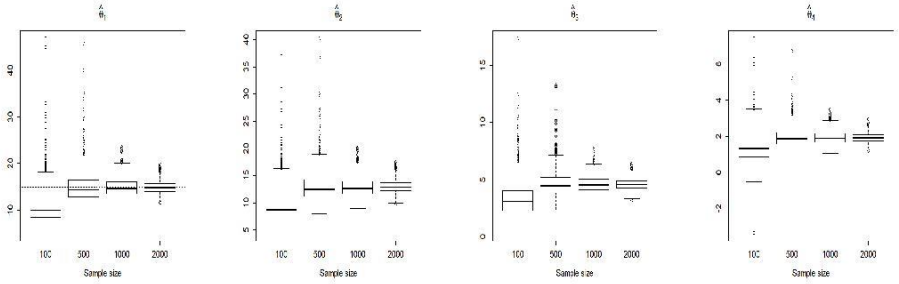
[15] Shao, Q. and Yang, L. (2017) Oracally efficient estimation and consistent model selection for auto-regressive moving average time series with trend. Journal of the Royal Statistical Society Series B 79, 507-524.

[16] Shen, J. and Gao, S. (2008) A solution to separation and multicollinearity in multiple logistic regression. Journal of Data Science 6, 515-531.

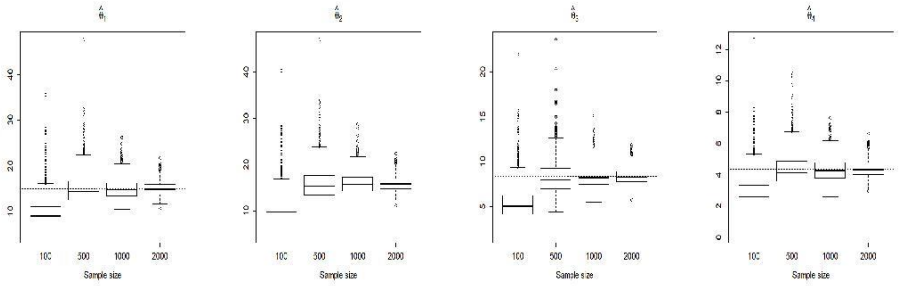
$\phi_0 = -0.2$



$\phi_0 = 0.2$



$\phi_0 = 0.4$



$\phi_0 = -0.4$

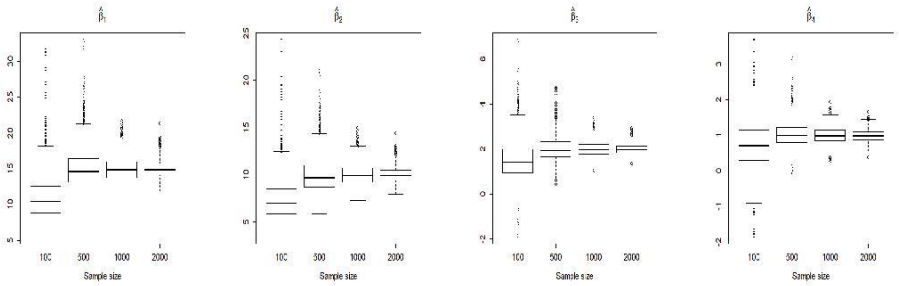


Figure 1: Box-plots for Coefficient Estimates of Binary Data

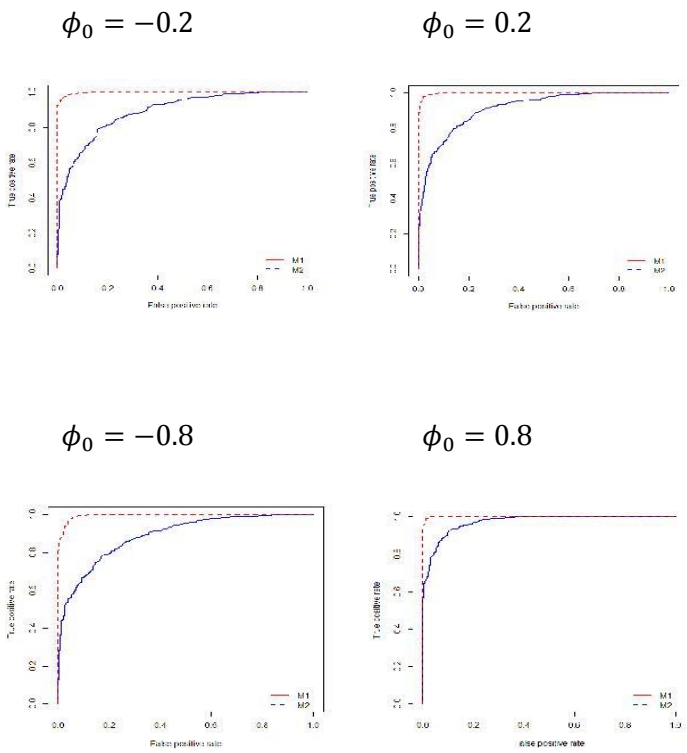


Figure 2: ROC Curves for M_1 and M_2

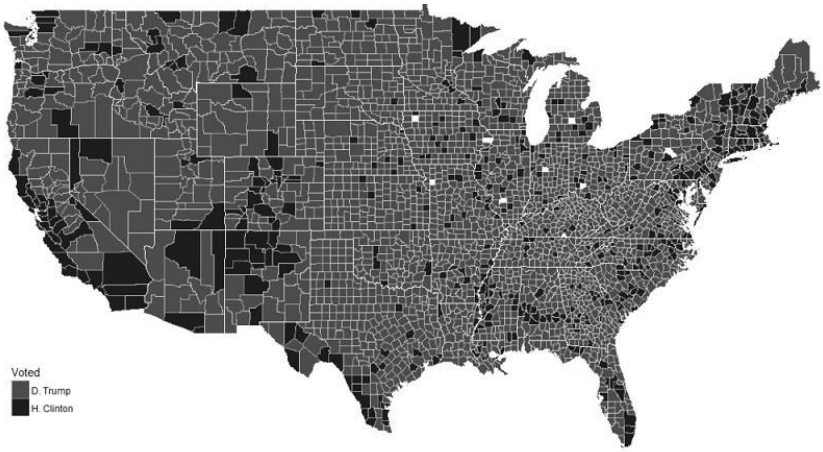


Figure 3: Presidential Election Map for Counties