Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

On semiparametric generalized linear models with nonparametric canonical link

Busayasachee Puang-Ngern
Chulalongkorn University
Wandee Wanishsakpong
wandee.w@ku.th
Kasetsart University
Jun Ma
Macquarie University

Research Article

Keywords: Semiparametric generalized linear models, canonical link, maximum likelihood estimation, constrained optimisation

Posted Date: March 8th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-8962870/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

On semiparametric generalized linear models with nonparametric canonical link

Busayasachee Puang-Ngern[1], Wandee Wanishsakpong[2*], Jun Ma[3]

[1]Chulalongkorn Business School, Chulalongkorn University, Bangkok, 10330, Thailand.
[2*]Department of Statistics, Kasetsart University, Bangkok, 10900, Thailand.
[3]School of Mathematical and Physical Sciences, Macquarie University, Sydney, 2109, Australia.

*Corresponding author(s). E-mail(s): wandee.w@ku.th;
Contributing authors: busayasachee@cbs.chula.ac.th; jun.ma@mq.edu.au;

Abstract

Since the seminal paper by Nelder and Wedderburn (1972), generalized linear models (GLMs) have become a popular option for building regression models. Although GLMs offer a rich family of distributions and link functions, modellers often encounter the perplexing challenge of finding a suitable combination of response distribution and link function that yields a satisfactory fit to the data. The semiparametric GLM is a useful and powerful alternative as it offers flexibility in the response distribution, but it still requires specifications of a link function and its maximum likelihood computation is difficult. In this paper, we propose a novel extension of the semiparametric GLM in which the link function is required to be canonical, but its functional form is left unspecified. That is, the link function is unknown (apart from the canonical requirement) and is determined nonparametrically from the data.

Keywords: Semiparametric generalized linear models, canonical link, maximum likelihood estimation, constrained optimisation

1 Introduction

The conventional generalized linear models (GLMs) provide an important parametric framework for constructing a regression model based on observations of a response variable and its predictors. This approach requires two modeling components to be specified explicitly: an error distribution and a link function. Canonical link functions are typically the default choice because they are easy to interpret and generally preserve the constraints that the mean response must satisfy. When a link function is unspecified and determined from the observed data, it is also known as a single-index GLM; see Kakade et al. (2011).

In this paper, we develop an alternative to the semiparametric GLM regression model (see, e.g., Rathouz and Gao (2009) and Huang (2014)), in which a "canonical-type" link function is adopted but its functional form is left unspecified and therefore modeled nonparametrically.

Parametric GLMs are set up as follows. Let Y_i, where $i = 1, \ldots, n$, denote the response variable associated with individual i. Assume there are q predictors whose observed values form an $n \times q$ design matrix, denoted by $\mathbf{X}$, whose i-th row, $\mathbf{X}_i$, contains the predictor values for individual i. It is typically assumed that, conditional on $\mathbf{X}$, the Y_i's are independent. For a GLM, the distribution of Y_i belongs to the exponential family, with density function

$$f(y|\mathbf{X}_i) = f_0(y) \exp\{y\theta_i - b(\theta_i)\}, \qquad (1)$$

where θ_i is called the canonical parameter, $b(\cdot)$ is a known function and $f_0(y) \geq 0$ is also known but not necessarily a density function.

A parametric GLM requires specification of a link function $g(\cdot)$ such that $g(\mu_i) = \eta_i$, where η_i is called the linear predictor given by $\eta_i = \mathbf{X}_i\boldsymbol{\beta}$, where $\boldsymbol{\beta}$ is a q-vector of regression coefficients. If the selected link function g satisfies $g(\mu_i) = \theta_i$, it is then called the canonical link. There are many well known canonical links (see for example McCullagh and Nelder (1989)), such as the log-link for Poisson responses and the logistic-link for Binomial responses.

This paper considers the semiparametric generalized linear model (SPGLM), whose density function retains the form of (1), except that $f_0(y)$ is now treated nonparametrically. However, due to identifiability concerns, in addition to requiring $f_0(y) \geq 0$, it is also necessary to impose $\int f_0(y)dy = 1$; see Rathouz and Gao (2009). Therefore, $f_0(y)$ is assumed to be a density function, although its functional form remains unspecified. Consequently, $f(y|\mathbf{X}_i)$ can be viewed as an exponential tilting of $f_0(y)$. From the relationship given in (2) below, we see that the functional form of $b(\theta_i)$ in the SPGLM density depends on $f_0(y)$, and thus $b(\theta_i)$ is also treated nonparametrically.

Before discussing the estimation of SPGLM parameters in Section 2, we first outline some key features of this model. Additional details on the SPGLM can be found in Rathouz and Gao (2009) and Huang (2014). Let $F_0(y)$ denote the distribution function associated with $f_0(y)$, and let $\mu_i = E(Y_i|\mathbf{x})$ and $\sigma_i^2 = \mathrm{Var}(Y_i \mid \mathbf{X}_i)$. Since $f(y \mid \mathbf{X}_i)$ is a density function, it follows that

$$b(\theta_i) = \log \int \exp\{y\theta_i\}\, dF_0(y), \qquad (2)$$

and this relationship yields the first two derivatives of $b(\theta_i)$ with respect to θ_i:

$$b'(\theta_i) = \mu_i \quad \text{and} \quad b''(\theta_i) = \sigma_i^2 \qquad (3)$$

for any i. Note that the relationship in (2) indicates that the functional form of $b'(\theta_i)$ depends on that of $f_0(y)$. Applicaion of SPGLM typically also requires specification of a parametric link function $g(\cdot)$ such that $g(\mu_i) = \eta_i$. For SPGLMs, the unknown θ_i's must be estimated. Since θ_i depends on f_0 and η_i, its estimation is not straightforward. As explained in Rathouz and Gao (2009) and

Huang (2014), a constrained optimization procedure is involved to estimate $f_0(y)$, $\boldsymbol{\beta}$ and all θ_i simultaneously.

In this paper, we consider a special case of SPGLM in which the link function $g(\cdot)$ is the canonical link; namely,

$$g(\mu_i) = \theta_i = \eta_i. \tag{4}$$

We refer to this model as SPGLM-cl. Although the functional form of g is unspecified, its value at each μ_i equals the linear predictor η_i. This implies that once $\boldsymbol{\beta}$, and thus η_i, is estimated, $g(\mu_i)$ is known. This property explains the computational simplicity when the canonical link is used.

The rest of this paper is organized as follows. Section 2 provides further details on SPGLM-cl, and also discusses the issue of identifiability. Section 3 describes the maximum likelihood (ML) estimation of $f_0(y)$ and $\boldsymbol{\beta}$, where a Lagrange method is employed to regularize the estimate of $f_0(y)$ so that its integral equals 1. Asymptotic properties of the ML estimates are derived in Section 4. A simulation study and an application of the model to real data are presented in Section 5, and conclusions are given in Section 6.

2 Semiparametric GLM with canonical links

2.1 Canonical link

In Section 1, we explained that the general SPGLM method assumes the response variable follows a semiparametric distribution given by (1), where $f_0(y)$ is a nonparametric density (or probability) function. Note that θ_i in (1) serves a similar role to the canonical parameter in the conventional GLM. Therefore, we also refer to θ_i as the canonical parameter for the SPGLM.

For the SPGLM, a parametric link function $g(\cdot)$ must be specified (Rathouz and Gao 2009). In this paper, however, we study the canonical link function defined in (4), and this special model is referred to as the SPGLM-cl. A remarkable benefit of this model, compared with SPGLM, is that there is no need to estimate the θ_i's, as once $\boldsymbol{\beta}$ is estimated, so are all θ_i $(= \mathbf{X}_i\boldsymbol{\beta})$.

Since, for the SPGLM-cl model, we have $\theta_i = \eta_i$, its density function can be rewritten as

$$f(y|\mathbf{X}_i) = f_0(y)\exp\{y\eta_i - b(\eta_i)\}, \tag{5}$$

where

$$b(\eta_i) = \log \int \exp\{y\eta_i\} dF_0(y) \tag{6}$$

satisfying $b'(\eta_i) = \mu_i$ and $b''(\eta_i) = \sigma_i^2$.

The canonical link function $g(\cdot)$ is not directly specified in SPGLM-cl and is therefore treated nonparametrically. Moreover, this link function is not estimated when we estimate $\boldsymbol{\beta}$ and $f_0(y)$. However, since $g(\mu_i) = \eta_i$, where the values of η_i are available once $\boldsymbol{\beta}$ is estimated, and μ_i can be computed from

$$\mu_i = \int y\exp\{y\eta_i - b(\eta_i)\}\, dF_0(y)$$

after the estimates of η_i and $f_0(y)$ are obtained, we can visualize g by plotting all η_i against all μ_i. Since, for the SPGLM-cl, there is no need to

explicitly specify either the canonical link function or the response distribution, it offers a remarkable advantage. This advantage relieves modellers of the potentially time-consuming and tedious task of searching for suitable response distributions and link functions when fitting regression models. On the other hand, SPGLM-cl can also be viewed as a flexible single-index regression modeling, where there is no need to specify a distribution for the response variable.

2.2 Identifiability

The identifiability issue of SPGLM under a general link function has been discussed in Rathouz and Gao (2009). They demonstrated that the density function in (1) cannot be distinguished between f_0 and f_0^* if f_0^* is a tilted version of f_0. They addressed this problem by constraining the mean of f_0 to an arbitrarily specified value. Therefore, for the SPGLM-cl considered in this paper, we can directly adopt their result to ensure model identifiability. However, we show below that the identifiability issue for the SPGLM-cl can be simplified by imposing the restriction that no intercept is included in the linear predictor. This condition can be easily satisfied by centering each column of the design matrix $\mathbf{X}$, which is a common practice when fitting regression models.

The identifiability issue arises if there exists a pair $(f_0^*(y), \eta_i^*)$, differing from $(f_0(y), \eta_i)$, such that they also produce $f(y \mid \mathbf{X}_i)$, namely

$$f(y \mid \mathbf{X}_i) = f_0^*(y) \exp\{y\eta_i^* - b^*(\eta_i^*)\}, \qquad (7)$$

where $b^*(\eta_i^*) = \log \int \exp\{y\eta_i^*\} f_0^*(y)\, dy$. Such a pair $(f_0^*(y), \eta_i^*)$ can be identified in the following manner. Let $\tilde{\eta}$ be any constant on the real line R and let f_0^* be the tilted version of f_0 using $\tilde{\eta}$, i.e.,

$$f_0^*(y) = f_0(y) \exp\{y\tilde{\eta} - b(\tilde{\eta})\}, \qquad (8)$$

where $b(\tilde{\eta}) = \log \int \exp\{y\tilde{\eta}\} f_0(y)\, dy$. Now, let

$$\eta_i^* = \eta_i - \tilde{\eta}. \qquad (9)$$

Then, after substituting $f_0^*(y)$ from (8) into the expression for $b^*(\eta_i^*)$ given after equation (7), we have $b^*(\eta_i^*) = b(\eta_i) - b(\tilde{\eta})$. Therefore, the pair $(f_0^*(y), \eta_i^*)$, defined respectively by (8) and (9), will also yield $f(y \mid \mathbf{X}_i)$.

When $\mathbf{X}$ has full column rank, one way to avoid this identifiability issue is by setting the intercept in the linear predictor η_i to 0, as this forces $\tilde{\eta} = 0$ (note that η_i^* and η_i differ only by their intercepts). This option should not raise concerns because the intercept is generally not a parameter of interest. With $\tilde{\eta} = 0$ and $\int f_0(y)dy = 1$, we have $f_0^*(y) = f_0(y)$.

3 Maximum likelihood estimation

In this section, we formulate a maximum likelihood (ML) procedure for the simultaneous estimation of $\boldsymbol{\beta}$ and $f_0(y)$. First, we approximate $f_0(y)$ using a finite-dimensional space as a nonparametric $f_0(y)$ is an infinite-dimensional parameter and cannot be estimated directly from a finite number of observations. Similar to Ruppert et al. (2003) and Ma et al. (2014), we

adopt an approximation using an m-dimensional functional space with nonnegative basis functions $\psi_1(y), \ldots, \psi_m(y)$. Therefore,

$$f_0(y) = \sum_{u=1}^{m} \alpha_u \psi_u(y), \tag{10}$$

where $m \leq n$ and $\alpha_u \geq 0$ are coefficients. A special set of nonnegative basis functions, on which we focus in this paper, are indicator functions, leading to a piecewise-constant approximation of $f_0(y)$. Assume the response variable has range $[a, b]$. We divide $[a, b]$ into m mutually exclusive and exhaustive bins $B_1, \ldots, B_m$ using bin edges $w_0, \ldots, w_m$, where $B_u = (w_{u-1}, w_u]$. Here. $w_0 = a$ and $w_m = b$. We note that a particular set of bin edges, given by distinct $y_1, \ldots, y_n$ values, lead to an empirical approximation of the cumulative distribution function $F_0(y)$. Define $\psi_u(y) = 1$ if $y \in B_u$ and $\psi_u(y) = 0$ otherwise. Then, (10) gives $f_0(y) = \alpha_u$ for $y \in B_u$. We define the probability mass as $p_u = \alpha_u \delta_u$, where δ_u denotes the bin width of interval B_u. Let $\mathbf{p}$ denote the m-vector consisting of all p_u. The problem of estimating $f_0(y)$ now reduces to estimating $\mathbf{p}$, where $p_u \geq 0$ for all u and $\sum_u p_u = 1$.

Let n_u be the number of y_i observations contained in bin B_u. Based on the piecewise constant approximation to $f_0(y)$, the log-likelihood function for $y_1, \ldots, y_n$, where the density of each y_i is given by (5), is

$$l(\boldsymbol{\beta}, \mathbf{p}) = \sum_{i=1}^{n} (y_i \eta_i - b(\eta_i)) + \sum_{u=1}^{m} n_u \log p_u, \tag{11}$$

where $\eta_i = \mathbf{X}_i \boldsymbol{\beta}$ and $b(\eta_i)$, according to (6), is

$$b(\eta_i) = \log \sum_{u=1}^{m} (\exp\{w_u \eta_i\} p_u). \tag{12}$$

We estimate $\boldsymbol{\beta}$ and $\mathbf{p}$ by maximizing $l(\boldsymbol{\beta}, \mathbf{p})$ in (11), subject to the constraints $p_u \geq 0$ and $\sum_u p_u = 1$. This constitutes a constrained maximum likelihood problem. We solve it by introducing the Lagrangian for the constraint $\sum_u p_u = 1$ and then applying the Karush–Kuhn–Tucker (KKT) necessary conditions to handle the $p_u \geq 0$ constraints. Further details are provided below.

The Lagrangian for enforcing the constraint $\sum_u p_u = 1$ is

$$\mathcal{L}(\boldsymbol{\beta}, \mathbf{p}) = l(\boldsymbol{\beta}, \mathbf{p}) - \lambda \left(\sum_u p_u - 1 \right), \tag{13}$$

where λ is the Lagrange multiplier. Then, the Karush–Kuhn–Tucker (KKT) conditions for maximizing $\mathcal{L}(\boldsymbol{\beta}, \mathbf{p})$ with respect to $\boldsymbol{\beta}$ and $\mathbf{p}$ (with $p_u \geq 0$) are

$$\frac{\partial \mathcal{L}}{\partial \beta_j} = 0,$$
$$\frac{\partial \mathcal{L}}{\partial p_u} = 0 \ \text{ if } p_u \neq 0, \quad \frac{\partial \mathcal{L}}{\partial p_u} < 0 \ \text{ if } p_u = 0,$$

where $j = 1, \ldots, q$ and $u = 1, \ldots, m$. Let $\mathbf{y}$ and $\boldsymbol{\mu}$ be the n-vectors of y_i's and μ_i's, respectively, and let $d^{(k)}$ be the estimate of d at iteration k. We solve these KKT equations using an alternating iterative algorithm where each iteration involves a combination of the Newton and the multiplicative-iterative (MI) (see, for example, Chan and Ma (2012)) algorithm.

In each iteration of this algorithm, the coefficient vector $\boldsymbol{\beta}$ is updated using the Newton

method. Since

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{\beta}} = \mathbf{X}^T\mathbf{W}(\mathbf{y}-\boldsymbol{\mu})$$
$$\frac{\partial^2 \mathcal{L}}{\partial \boldsymbol{\beta}\,\partial \boldsymbol{\beta}^T} = -\mathbf{X}^T\mathbf{W}\mathbf{X},$$

where $\mathbf{W} = \mathrm{diag}(b''(\eta_1), \ldots, b''(\eta_n))$, the Newton method with line search gives:

$$\boldsymbol{\beta}^{(k+1)} = \boldsymbol{\beta}^{(k)} + \omega_1^{(k)}(\mathbf{X}^T\mathbf{W}^{(k)}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}^{(k)}(\mathbf{y}-\boldsymbol{\mu}^{(k)}), \tag{14}$$

where $\omega_1^{(k)} \in (0, 1]$ is a line search step size, assuring $l(\boldsymbol{\beta}^{(k+1)}, \mathbf{p}^{(k)}) \geq l(\boldsymbol{\beta}^{(k)}, \mathbf{p}^{(k)})$.

The KKT conditions on p_u give $p_u \partial\mathcal{L}/\partial p_u = 0$ for all u, where

$$\frac{\partial \mathcal{L}}{\partial p_u} = -\sum_{i=1}^{n} \frac{\exp\{w_u\eta_i\}}{\exp\{b(\eta_i)\}} + \frac{n_u}{p_u} - \lambda. \tag{15}$$

By solving the equation $\sum_u p_u \partial\mathcal{L}/\partial p_u = 0$ we obtain $\lambda = 0$. Since $b(\eta_i)$ also contains p_u, equation (15) exhibits p_u nonlinearly. We solve the KKT equations for all p_u by the MI algorithm. For Step 1 of this algorithm, we temporarily solve these equations for all p_u by

$$p_u^{(k+\frac{1}{2})} = \frac{n_u}{\sum_{i=1}^{n} \exp\{w_u\eta_i^{(k)} - b(\eta_i^{(k)})\}}, \tag{16}$$

which guarantees $p_u^{(k+\frac{1}{2})} \geq 0$. Then, in Step 2 of the MI algorithm, after a line search step, we have the final updating formula for $\mathbf{p}$:

$$\mathbf{p}^{(k+1)} = \mathbf{p}^{(k)} + \omega_2^{(k)}(\mathbf{p}^{(k+\frac{1}{2})} - \mathbf{p}^{(k)}), \tag{17}$$

where $\omega_2^{(k)} \in (0, 1]$ is a line search step size such that $l(\boldsymbol{\beta}^{(k+1)}, \mathbf{p}^{(k+1)}) \geq l(\boldsymbol{\beta}^{(k+1)}, \mathbf{p}^{(k)})$. More explanations of the MI algorithm can be found in Chan and Ma (2012) and Ma et al. (2014). We make the following comments on formula (17).

1. Algorithm (17) can be expressed as a gradient algorithm:

$$\mathbf{p}^{(k+1)} = \mathbf{p}^{(k)} + \omega_2^{(k)}\mathbf{S}^{(k)}\frac{\partial \mathcal{L}(\boldsymbol{\beta}^{(k+1)}, \mathbf{p}^{(k)})}{\partial \mathbf{p}}, \tag{18}$$

   where $\mathbf{S} = \mathrm{diag}(s_1, \ldots, s_m)$ with $s_u = p_u / \sum_i \exp\{c_u\eta_i - b(\eta_i)\}$. Since matrix $\mathbf{S}$ is non-negative definite, the search direction for updating $\mathbf{p}$ is uphill and therefore the line search step will guarantee non-decreasing of the log-likelihood function.

2. If (17) converges, it will converge to a solution satisfying the KKT conditions on p_u (e.g. Chan and Ma (2012)). Note that equation (16) shows that $p_u^{(k+\frac{1}{2})} \geq 0$, where the equality holds only when $n_u = 0$. Thus $\mathbf{p}^{(k+1)} \geq 0$ if $\mathbf{p}^{(k)} \geq 0$, where the inequalities for vectors are interpreted elementwisely. When $n_u = 0$ then clearly $\partial\mathcal{L}/\partial p_u < 0$.

Thus, in each iteration of this proposed algorithm, (14) is first used to update $\boldsymbol{\beta}$, and then (17) is applied to update $\mathbf{p}$. Under certain regularity conditions (Chan and Ma 2012), this algorithm produces a sequence $\{\boldsymbol{\beta}^{(k)}, \mathbf{p}^{(k)}\}$ that converges to a solution satisfying the KKT conditions as well as the constraint $\sum_u p_u = 1$. Moreover, it is possible that some estimates $\widehat{p}_u = 0$, which means that the ML solution may lie on the boundary of the solution space. This fact makes the traditional asymptotic results for maximum likelihood

estimation less applicable. In the next section, we develop asymptotic results for the constrained maximum likelihood estimates of $\boldsymbol{\beta}$ and $\mathbf{p}$.

4 Asymptotic results

In this section, we develop asymptotic normality results for the parameter estimates of the proposed SPGLM-cl method. The results are derived under the assumption that f_0 is parameterized by $\mathbf{p}$ with a fixed m, so that m remains in the asymptotic expressions. This assumption lies between parametric and nonparametric modeling, as it allows flexibility in the choice of m. Under this setting, the maximum likelihood estimates of $\boldsymbol{\beta}$ and $\mathbf{p}$ still achieve a $\sqrt{n}$ convergence rate.

Let the vector $\boldsymbol{\gamma} = \left(\boldsymbol{\beta}^T, \mathbf{p}^T\right)^T$, which has dimension $q+m$. Its maximum likelihood estimate is denoted by $\widehat{\boldsymbol{\gamma}}$. Since it is often the case that some of the constrained maximum likelihood estimates of p_u are exactly zero, implying the existence of active constraints, we adopt the approach of Moore et al. (2008) to develop asymptotic normality results for $\widehat{\boldsymbol{\gamma}}$. Suppose there are m_1 constraints of the form $p_u = 0$. Note that the constraint $\sum_u p_u = 1$ is always active for all p_u. Therefore, there exist $m_1 + 1$ active constraints.

The Fisher information matrix for $\boldsymbol{\gamma}$ is

$$\mathcal{I}(\boldsymbol{\gamma}) = -\mathrm{E}\left(\frac{\partial^2 l(\boldsymbol{\gamma})}{\partial\boldsymbol{\gamma}\partial\boldsymbol{\gamma}^T}\right) = \begin{bmatrix} \mathcal{I}_{\boldsymbol{\beta}\boldsymbol{\beta}} & \mathcal{I}_{\boldsymbol{\beta}\mathbf{p}} \\ \mathcal{I}_{\mathbf{p}\boldsymbol{\beta}} & \mathcal{I}_{\mathbf{pp}} \end{bmatrix}. \quad (19)$$

where the details of this matrix elements are given in the Appendix.

To simplify the discussion, we assume that the first m_1 constraints of the form $p_u \geq 0$ are active. Define

$$\xi_0 = \sum_u p_u, \quad \xi_1 = p_1, \ldots, \xi_{m_1} = p_{m_1}.$$

Let $\mathbf{G}$ denote the matrix of derivatives of $\xi_0, \xi_1, \ldots, \xi_{m_1}$ with respect to $\boldsymbol{\gamma}$. This matrix has dimension $(m_1 + 1) \times (q + m)$ and is given by

$$\mathbf{G} = \begin{bmatrix} \mathbf{0}_{1\times q} & \mathbf{1}_{1\times m_1} & \mathbf{1}_{1\times(m-m_1)} \\ \mathbf{0}_{m_1\times q} & \mathbf{I}_{m_1\times m_1} & \mathbf{0}_{m_1\times(m-m_1)} \end{bmatrix}, \quad (20)$$

where $\mathbf{0}$ and $\mathbf{1}$ denote respectively the matrices of zeros and ones. Let $\mathbf{U}$ be a $(q + m) \times (q + m - m_1 - 1)$ matrix whose columns form an orthonormal bases for the null space of $\mathbf{G}$, so that $\mathbf{GU} = \mathbf{0}_{(m_1+1)\times(q+m-m_1-1)}$ and $\mathbf{U}^T\mathbf{U} = \mathbf{I}_{(q+m-m_1-1)\times(q+m-m_1-1)}$. Then according to Moore et al. (2008), the asymptotic covariance matrix for $\widehat{\boldsymbol{\gamma}}$ is given by

$$\mathbf{F}(\boldsymbol{\gamma})^{-1} = \mathbf{U}(\mathbf{U}^T\mathcal{I}(\boldsymbol{\gamma})\mathbf{U})^{-1}\mathbf{U}^T. \quad (21)$$

The asymptotic normality results in Theorem 1 require some regularity conditions as specified below:

Theorem 1. *Consider the maximum likelihood estimates of $\boldsymbol{\beta}$ and $\mathbf{p}$ for the SPGLM-cl model with a fixed m. Suppose that m_1 of the $p_u \geq 0$ constraints are active. Assuming that Assumptions 1–7 in the Appendix hold, as $n \to \infty$, the quantity $\sqrt{n}(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0)$ converges in distribution to a multivariate normal distribution $N(0, \mathbf{F}(\boldsymbol{\gamma}_0)^{-1})$, where $\mathbf{F}(\boldsymbol{\gamma}_0)^{-1}$ is given in equation (21).*

Here, we have omitted the proof of this theorem, as it can be readily derived by following

Theorem 2 of Moore et al. (2008). The simulation results reported in Section 5 indicate that the asymptotic variances computed using (21) are accurate.

5 Results

We assess the performance and accuracy of the proposed SPGLM-cl model by a simulation study and a real data application.

5.1 Simulation

In this simulation study, we generated samples of sizes $n = 100$, 500, and 10,000. These sample sizes allow us to examine the performance of SPGLM-cl for small, medium, and large samples. For each sample size, the Monte Carlo results were obtained from $N = 1{,}000$ replicated samples.

To study the performance of SPGLM-cl with different response types and their canonical link functions, we generated response data using the following GLMs: (i) Poisson distribution with a log link, (ii) normal distribution with an identity link, and (iii) Binomial distribution with a logit link.

Specifically, we generated samples of size $n = 100$ for the Poisson data, $n = 500$ for the normal data, and $n = 10{,}000$ for the Binomial data. For model fitting with SPGLM-cl, we set $n_u = 1$ (the number of observations in each bin) for all u when there were no ties. All simulations were conducted using the R software. We included two covariates, x_1 and x_2, in our simulations, with their values generated from the standard uniform distribution. The true regression coefficients used for x_1 and x_2 were $[\beta_1, \beta_2] = [-0.3, 0.5]$ for the normal response data, and $[\beta_1, \beta_2] = [0.3, 0.2]$ for both the Poisson and Binomial response data.

We compared the proposed SPGLM-cl with the GLM in terms of their estimates of the regression coefficients β_1 and β_2, where the GLMs assumed both the correct response distribution and the correct link function. The accuracy of the estimates was assessed by comparing their mean estimates (MEAN), bias (BIAS), average asymptotic standard error (AASE), Monte Carlo standard error (MCSE), and mean squared error (MSE). Note that the MSE values were calculated using the AASE. The results are presented in Table 1.

We also examined the accuracy of inference on the regression coefficients. In particular, we tested the following hypotheses: $H_0 : \beta_1 = -0.3$ versus $H_1 : \beta_1 \neq -0.3$ for the normal data, and $H_0 : \beta_1 = 0.3$ versus $H_1 : \beta_1 \neq 0.3$ for both the Poisson and Binomial data, all at the 5% significance level. We calculated and reported Type I error rates based on (i) the Wald test using the AASE of $\widehat{\beta}_1$ (WALD-AASE), and (ii) the likelihood ratio test with χ^2_1 distribution (LR). The results are presented in Table 2.

The results in Table 1 show that SPGLM-cl gives similar results to GLM particularly when sample sizes are large. It is expected that GLM provides smaller biases for small to moderate sample sizes as the correct response distribution and link function were correctly specified for the GLM.

The type I error results in Table 2 reveal that, for both SPGLM-cl and GLM, type I errors are close to the nominal rate of 0.05. These findings suggest that the coefficient estimates and inferences of the proposed SPGLM-cl are similar to those of the GLM when the distribution and link functions are correctly specified for the latter. However, SPGLM-cl offers the advantage that neither the response distribution nor the link function needs to be specified, which is a remarkable improvement over the conventional GLM.

Using the simulated normal response data, we further investigated the impact of the number of observations for the bins in the piecewise constant approximation to $f_0(y)$, where equal numbers of observations were adopted for each bin. Since all n_u are equal in this context, we denote them by n_0 in the following discussions. For each simulated data set, we fitted the SPGLM-cl model with $n_0 = 1, 5, 20, 25, 50$, and 100, and calculated the MSEs for each β_j under these different n_0 values. The MSE plots are shown in Figure 1. In this figure, the MSEs of $\hat{\beta}_1$ (solid line) and $\hat{\beta}_2$ (dashed line) are plotted against n_0 from 0 to 100. In general, smaller n_0 values (less than 25) yield better MSEs than larger ones. Although the optimal MSE occurs around $n_0 = 23 \sim 25$ for both coefficients, the improvement over the MSE at $n_0 = 1$ is minimal. The plots also indicate that large n_0 values should be avoided, as the MSE of $\hat{\beta}_1$ increases rapidly when $n_0 > 30$. Our experience with other data sets similarly suggests that a small n_0, such as $n_0 = 1$, is often a better choice. In practice, if one wishes to select an optimal n_0, AIC or other information criteria can be used.

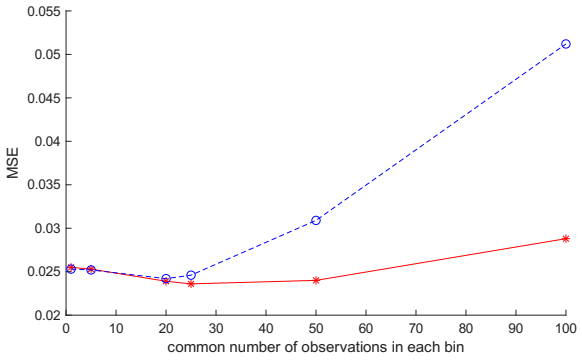


Fig. 1: Plots of MSE of $\hat{\beta}_1$ (solid-star red) and $\hat{\beta}_2$ (dashed-circle blue) using different number of observations in each bin for normal response with $n = 500$.

5.2 Application to CD4 count data

In this section, we apply the SPGLM-cl model to the CD4 count data reported in Wade and Ades (1994). This dataset has previously been used to illustrate the GAMLSS modeling approach in Stasinopoulos and Rigby (2007) and is available in the R package `gamlss.data`.

Our objective is to model the relationship between the mean CD4 count of uninfected children born to HIV-1 infected mothers (`cd4`) and the age of the child (`age`). The scatter plot in Figure 2 suggests that the relationship is likely to be non-linear. In fact, the preliminary analysis in Stasinopoulos and Rigby (2007) indicates that this relationship is not straightforward to model. In this section, we apply the SPGLM-cl model to fit this dataset.

Table 1: Simulation results for Poisson ($n = 100$), Normal ($n = 500$) and Binomial ($n = 10,000$) responses.

Distribution		SPGLM-cl		GLM	
		$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_1$	$\hat{\beta}_2$
Poisson	MEAN	0.3265	0.1970	0.3114	0.1866
	BIAS	0.0265	-0.0030	0.0114	-0.0134
	AASE	0.4060	0.4001	0.3927	0.3874
	MCSE	(0.3941)	(0.4065)	(0.3740)	(0.3842)
	MSE	0.1655	0.1601	0.1544	0.1503
Normal	MEAN	-0.3023	0.5090	-0.2990	0.5032
	BIAS	-0.0023	0.0090	0.0010	0.0032
	AASE	0.1597	0.1589	0.1573	0.1545
	MCSE	(0.1659)	(0.1587)	(0.1629)	(0.1524)
	MSE	0.0255	0.0253	0.0247	0.0239
Binomial	MEAN	0.3682	0.1858	0.3682	0.1858
	BIAS	0.0682	-0.0142	0.0682	-0.0142
	AASE	0.7096	0.7266	0.7096	0.7266
	MCSE	(0.7301)	(0.7537)	(0.7301)	(0.7537)
	MSE	0.5081	0.5281	0.5081	0.5281

Table 2: Type I errors for testing β_1.

Distribution		SPGLM-cl	GLM
Poisson	WALD-AASE	0.041	0.038
	LR	0.045	0.041
Normal	WALD-AASE	0.059	0.063
	LR	0.060	0.060
Binomial	WALD-AASE	0.048	0.048
	LR	0.054	0.054

Apart from SPGLM-cl, we also fitted two standard GLMs to this data set: a log-linear model assuming Poisson-distributed responses (Poi-log) and a linear model with an identity link assuming normally distributed responses (Nor-ide). We evaluated the goodness of fit by comparing their prediction MSEs obtained using the leave-one-out cross-validation method (MSE-CV), and assessed the underlying distributional assumptions of these models using Probability-Probability (P-P) and histogram plots of the nonrandomized probability integral transform (PIT) quantities (e.g., Czado et al. (2009)).

Table 3: The model fitting results of the CD4 data set.

	Parameter	
	intercept	age
SPGLM-cl		
Estimates	-	-0.0016
ASE	-	0.0001
Poi-log		
Estimates	7.2479	-0.3956
ASE	0.0039	0.0017
Nor-ide		
Estimates	1,007.70	-171.72
ASE	36.78	12.56

The coefficient estimates and their asymptotic standard errors (ASEs) are presented in Table 3. Note that the regression coefficients are not directly comparable as these models adopt different link functions. The fitted curves in Figure 2 indicate that both SPGLM-cl (dashed) and Poi-log (dash-dotted) provide better fit to the data than Nor-ide (solid). Moreover, by comparing the prediction MSE-CV values in Table 4, we observe that SPGLM-cl provides the best fit. The P-P plots in Figure 3 show that SPGLM-cl (dashed) is very close to the benchmark line (dotted), whereas the other two methods deviate substantially. In addition, the PIT histograms in Figures 4 indicate that only SPGLM-cl employs an appropriate response distribution for model fitting, as its histogram is the only one that closely resembles the standard uniform distribution.

Table 4: The prediction MSEs (MSE-CV).

	SPGLM-cl	Poi-log	Nor-ide
MSE-CV	150,835	152,739	164,459

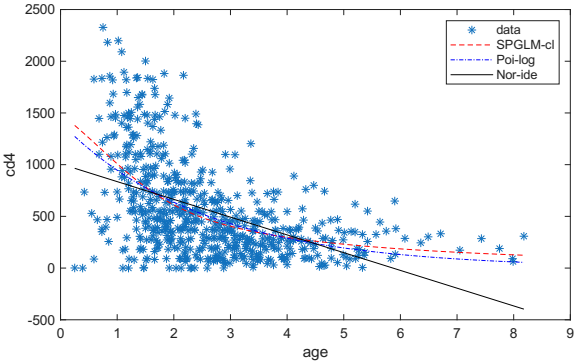


Fig. 2: The scatter plot of `cd4` and `age` and the fitted curves for SPGLM-cl (dashed red), Poi-log (dashed-dotted blue) and Nor-ide (solid black).

It is often informative to visualize the link function. For SPGLM-cl, this can be done by plotting the fitted linear predictor against the fitted mean as explained previously in Section 2. Figure 5 presents such a plot for the CD4 count data, providing a visualization of the nonparametric canonical link for this data set.

5.3 Application to Auto Insurance Claims Data

The Auto Insurance Claims dataset provided by Ndinguri (2024) is employed to illustrate performance of the proposed model. The dataset

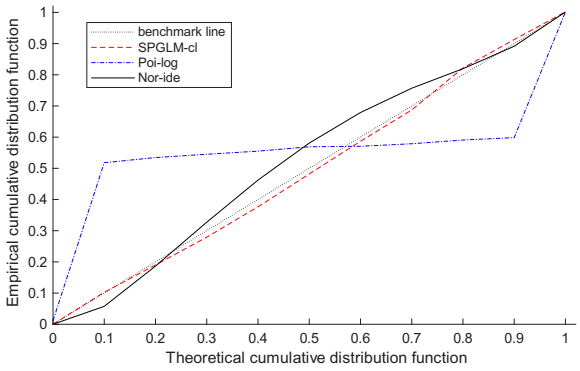


Fig. 3: The P-P plots of PITs for SPGLM-cl (dashed red), Poi-log (dashed-dotted blue) and Nor-ide (solid black) with the CD4 data set.

contains detailed insurance claim records, including policyholder characteristics and claim information, making it suitable for modelling claim amounts and identifying key risk factors.

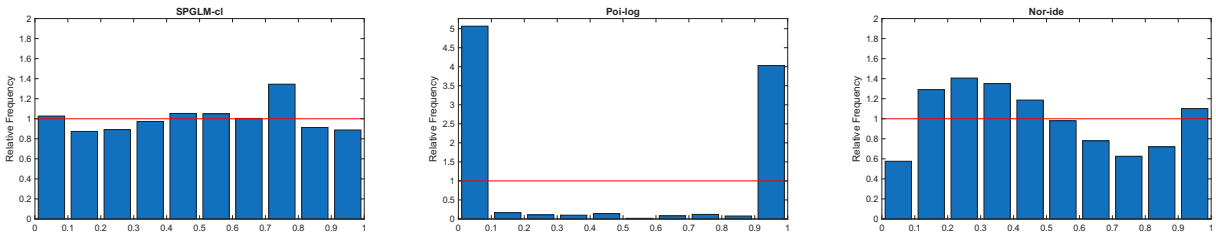
This dataset consists of 9,000 observations. The relationship between claim amount (`claim`) and a set of explanatory variables is examined, including insured age (`age`), sex (`sex`), education level (`education`), and occupation (`occupation`).

Education levels are grouped into three categories—low, medium, and high—to ensure sufficient sample sizes within each group and to enhance the stability of parameter estimation. The low education category includes individuals with a high school education. The medium education category comprises those holding associate or bachelor's degrees. The high education category consists of individuals with postgraduate or professional degrees, including master's, Juris Doctor (JD), Doctor of Medicine (MD), and Doctor of Philosophy (PhD) degrees. The low education (high school) category is used as the reference category in the regression analysis.

Insured occupations are similarly classified into three categories—manual, service, and professional—based on similarities in job characteristics and risk exposure. The manual category includes craft-repair, farming-fishing, handlers-cleaners, machine-operator-inspection, and transport-moving occupations. The service category comprises other-service, private household service, protective service, sales, technical support, and administrative clerical occupations. The professional category consists of executive managerial, professional specialty, and armed forces occupations. The service category is selected as the reference group, as it represents an intermediate level of occupational risk and income characteristics.

The distribution of claim amounts, measured in thousand dollars ('000), is illustrated in Figure 6. The histogram reveals a high concentration (but a smaller portion) of small claims, followed by a bell-shaped distribution for larger claim amounts. This bimodal distribution indicate potential difficulties when adopting a conventional GLM model.

Three models are applied to the claim amount data. These include the proposed SPGLM-cl model and two standard generalized linear models (GLMs): a gamma regression model with a reciprocal link function (Gam-rec), and a normal linear model with an identity link function (Nor-ide). The estimated regression coefficients and their corresponding asymptotic standard errors (ASEs)



(a) (b) (c)

Fig. 4: PIT histogram of different models for the CD4 data set: (a) SPGLM-cl, (b) Poi-log and (c) Nor-ide.

Table 5: The model fitting results of the auto insurance claims data set.

	SPGLM-cl			Gam-rec			Nor-ide		
	Estimate	ASE	p-value	Estimate	ASE	p-value	Estimate	ASE	p-value
age	0.0003	0.0000	<0.001	0.9098	0.0180	<0.001	0.0004	0.0000	<0.001
sex (F)	0.0025	0.0008	0.0023	5.7981	0.5585	<0.001	0.0018	0.0002	<0.001
edu (med)	-0.0026	0.0013	0.0421	6.8273	0.8538	<0.001	0.0048	0.0003	<0.001
edu (high)	0.0026	0.0011	0.0243	10.4400	0.7472	0.0016	0.0035	0.0002	<0.001
occ (man)	0.0073	0.0009	<0.001	8.7792	0.6455	<0.001	0.0007	0.0002	<0.001
occ (pro)	0.0057	0.0010	<0.001	8.1686	0.7290	<0.001	0.0013	0.0003	<0.001

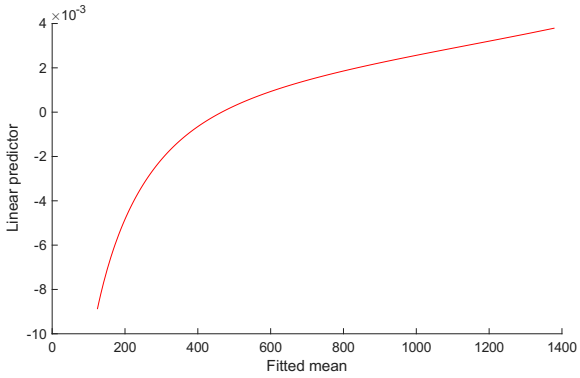


Fig. 5: The fitted mean curve of SPGLM-cl model for the CD4 data set.

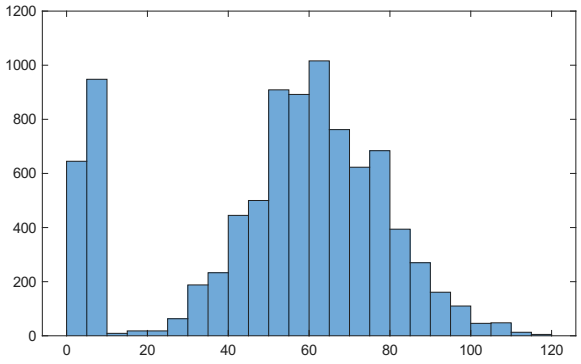


Fig. 6: The histogram of `claim`.

are reported in Table 5. In general, the signs of the regression coefficient estimates are the same across the three models, except for `education`

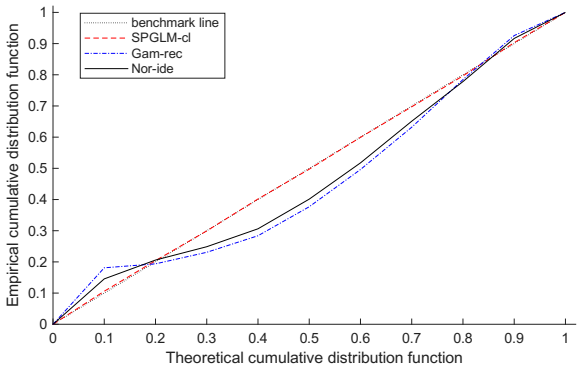


Fig. 7: The P-P plots of PITs for SPGLM-cl (dashed red), Gam-rec (dashed-dotted blue) and Nor-ide (solid black) with the auto insurance claims data set.

`(medium)`, where SPGLM-cl yields a negative estimate while the other models produce positive estimates.

Model adequacy is further assessed using probability–probability (P–P) plots and probability integral transform (PIT) histograms. The P–P plots in Figure 7 show that the SPGLM-cl model (dashed) closely follows the benchmark line (dotted), whereas the two parametric GLMs exhibit substantial deviations. Consistently, the PIT histograms in Figure 8 indicate that only the SPGLM-cl model yields a distribution that closely resembles the standard uniform distribution, suggesting its more appropriate specification of the response distribution for claim severity modelling.

As a complementary diagnostic, we visualize the implied link function by plotting the fitted linear predictor against the fitted mean. The resulting plot, shown in Figure 9, provides insight into the estimated nonparametric canonical link function for the auto insurance claims data.

6 Conclusions

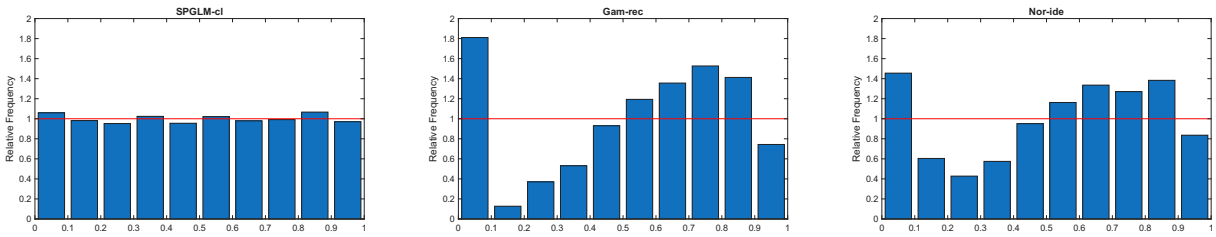
In this paper, we developed a novel semiparametric GLM method that adopts the canonical link function. The canonical link function is fully data-dependent and therefore must be estimated jointly with the other model parameters. A distinctive advantage of this method is that neither the response distribution nor the link function needs to be specified. Another benefit of this model is that the canonical parameters θ_i reduce to the linear predictors η_i, making the computations required to fit the model considerably less demanding than those for the conventional semiparametric GLMs of Rathouz and Gao (2009) and Huang (2014).

A maximum likelihood estimation procedure is developed in this paper. This algorithm is efficient for data sets with large sample sizes. Asymptotic results are derived, and simulation studies demonstrate that the asymptotic standard deviations closely match the Monte Carlo standard deviations. The reported simulation results and real-data application provide evidence that the proposed model-fitting method can deliver satisfactory results, particularly when traditional modeling approaches, such as GLM, encounter implementation difficulties.

Declarations

Competing interests. The authors declare no competing interests.

Data availability. The CD4 count data is available in the R package `gamlss.data`.



(a) (b) (c)

Fig. 8: PIT histogram of different models for the auto insurance claims data set: (a) SPGLM-cl, (b) Gam-rec and (c) Nor-ide.

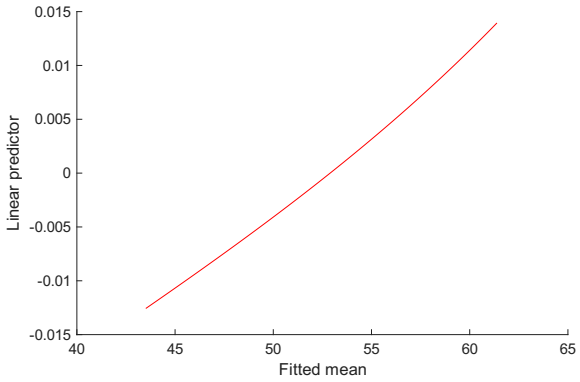


Fig. 9: The fitted mean curve of SPGLM-cl model for the auto insurance claims data set.

The Auto Insurance Claims dataset can be obtained from the following link: https://doi.org/10.5281/zenodo.13381118

Appendix

Elements of $\boldsymbol{I(\lambda)}$

Details of this matrix are:

$$\mathcal{I}_{\boldsymbol{\beta\beta}} = \mathbf{X}^T\mathbf{W}\mathbf{X} \quad \text{and} \quad \mathcal{I}_{\mathbf{p}\boldsymbol{\beta}} = \mathcal{I}^T_{\boldsymbol{\beta}\mathbf{p}}.$$

The elements of $\mathcal{I}_{\mathbf{pp}}$ are given by

$$-\mathrm{E}\left(\frac{\partial^2 l}{\partial p_u \partial p_v}\right) = -\sum_i \exp\{(w_u + w_v)\eta_i - 2b(\eta_i)\} + I_{u=v} n_u / p_u^2,$$

where $I_{u=v} = 1$ if $u = v$, and $I_{u=v} = 0$ otherwise. For $\mathcal{I}_{\boldsymbol{\beta}\mathbf{p}}$, its elements are

$$-\mathrm{E}\left(\frac{\partial^2 l}{\partial \boldsymbol{\beta} \partial p_u}\right) = \sum_i (w_u - \mu_i) \exp\{w_u\eta_i - b(\eta_i)\}\mathbf{X}_i.$$

Regularity conditions

The regularity conditions for Theorem 1 are:

1. The random vectors $(Y_i, \mathbf{X}_i)$, $i = 1, \ldots, n$, are independent and identically distributed. Moreover, the distribution of $\mathbf{X}_i$ is independent of $\boldsymbol{\gamma}$.
2. The covariates matrix $\mathbf{X}$ has full column rank.
3. Let Γ be the domain of $\boldsymbol{\gamma}$. Assume Γ is a compact subset of R^{q+m}.
4. $l(\boldsymbol{\gamma})$ is continuous over Γ.
5. Let $\boldsymbol{\gamma}_0$ be the true parameter $\boldsymbol{\gamma}$. Assume $l(\boldsymbol{\gamma})$ is twice differentiable in a neighborhood of $\boldsymbol{\gamma}_0$.

6. $\mathrm{E}_{\boldsymbol{\gamma}_0}[n^{-1}l(\boldsymbol{\gamma})]$ exists and has a unique maximum at $\boldsymbol{\gamma}_0 \in \Gamma$.
7. The Fisher information matrix at $\boldsymbol{\gamma}_0$, that is $\mathcal{I}(\boldsymbol{\gamma}_0)$, exists.

References

Czado, C., Gneiting, T., Held, L.: Predictive model assessment for count data. Biometrics **65**, 1254–1261 (2009)

Chan, R.H., Ma, J.: A multiplicative iterative algorithm for box-constrained penalized likelihood image restoration. IEEE Trans. Image Processing **21**, 3168–3181 (2012)

Huang, A.: Joint estimation of the mean and error distribution in generalized linear models. J. Amer. Statist. Assoc. **109**, 186–196 (2014)

Kakade, S., Kalai, A.T., Kanade, V., Shamir, O.: Efficient Learning of Generalized Linear and Single Index Models with Isotonic Regression. https://arxiv.org/abs/1104.2018

Ma, J., Heritier, S., Lô, S.: On the maximum penalized likelihood approach for proportional hazard models with right censored survival data. Computational Statistics and Data Analysis **74**, 142–156 (2014)

McCullagh, P., Nelder, J.A.: Generalized Linear Models (2nd Edition). Chapman and Hall, London (1989)

Moore, T.J., Sadler, B.M., Kozick, R.J.: Maximum-likelihood estimation, the Cramér – Rao bound, and the method of scoring with parameter constraints. IEEE Trans. Sig. Proc. **56**, 895–908 (2008)

Ndinguri: Auto Insurance Claims. https://doi.org/10.5281/zenodo.13381118 . https://doi.org/10.5281/zenodo.13381118

Nelder, J.A., Wedderburn, R.W.M.: Generalized linear models. J. Roy. Statist. Soc. B **135**, 370–384 (1972)

Rathouz, P.J., Gao, L.: Generalized linear models with unspecified reference distribution. Biostatistics **10**, 205–218 (2009)

Ruppert, D., Wand, M.P., Carroll, R.J.: Semiparametric Regression. Cambridge, New York (2003)

Stasinopoulos, D.M., Rigby, R.A.: Generalized additive models for location scale and shape (GAMLSS) in R. Journal of Statistical Software **23**, 1–46 (2007)

Wade, A.M., Ades, A.E.: Age-related reference ranges: Significance tests for models and confidence intervals for centiles. Statistics in Medicine **13**, 2359–2367 (1994)