Modified Poisson Regression Analysis of Grouped and Right-censored Counts*

Dr. Qiang Fu
Department of Sociology,
The University of British Columbia,
V6T 1Z1, Vancouver, BC, Canada
qiang.fu@ubc.ca

Tian-Yi Zhou
H. Milton Stewart School of Industrial and Systems Engineering,
Georgia Institute of Technology,
755 Ferst Dr NW, Atlanta, GA, United States
tzhou306@gatech.edu

Dr. Xin Guo
Department of Applied Mathematics,
The Hong Kong Polytechnic University, Hong Kong, and
School of Mathematics and Physics, University of Queensland,
Brisbane, Queensland, 4072, Australia
xin.guo@uq.edu.au

February 20, 2021

*The work described in this paper is supported partially by the Research Grants Council of Hong Kong [Project No. PolyU 15305018]. Please direct all correspondence to Xin Guo, Department of Applied Mathematics, The Hong Kong Polytechnic University, Hong Kong (Email: xin.guo@uq.edu.au).

Abstract

Grouped and right-censored (GRC) counts are widely used in criminology, demography, epidemiology, marketing, sociology, psychology, and other related disciplines to study behavioral and event frequencies, especially when sensitive research topics or individuals with possibly lower cognitive capacities are at stake. Yet, the co-existence of grouping and right-censoring poses major difficulties in regression analysis. To implement generalized linear regression of GRC counts, we derive modified Poisson estimators and their asymptotic properties, develop a hybrid line search algorithm for parameter inference, demonstrate the finite-sample performance of these estimators via simulation, and evaluate its empirical applicability based on survey data of drug use in America. This method has a clear methodological advantage over the ordered logistic model for analyzing GRC counts.

Key words: Regression Analysis, Grouped and Right-censored Counts, Modified Poisson Estimators, Zero Inflation, Hybrid Line Search, Fisher Information

1 Introduction

In survey research, response categories combining both grouped counts (e.g., a response category of “3-4 times” rather than separate categories of “3 times” and “4 times”) and right-censored counts (e.g., an upper end category as “5 or more times”) constitute a useful instrument to solicit information on sensitive themes (e.g., substance use) and/or from populations with lower cognitive capacities (e.g., the elderly and adolescents) (Blair and Burton, 1987; Schaeffer and Dykema, 2011; Schaeffer and Presser, 2003). Consequently, these grouped and right-censored (GRC) counts have so far been widely used in criminology, demography, epidemiology, marketing, sociology, psychology and other related disciplines to study behavioral and event frequencies (Ackard et al., 2002; Akers et al., 1989; Bachman et al., 1990; Hagan et al., 2005; Marsden, 2003). For example, since the year 1975, GRC counts have been repeatedly used in an ongoing national representative study of adolescents, the Monitoring the Future (MTF, or the National High School Senior Survey) project, to track annual changes in substance use and juvenile delinquency based on hundreds of middle and high schools in the United States (Johnston et al., 2017). Besides the MTF project, the only other repeated nationally-representative survey which collects information about risky behaviors among American youth, the Youth Risk Behavior Survey (YRBS), included GRC count responses in multiple survey questions (Kann et al., 2018). As the largest longitudinal survey of adolescents ever conducted, the National Longitudinal Study of Adolescent to Adult Health (Add Health) also uses GRC counts to collect information on a variety of juvenile delinquent behaviors (Harris, 2013).

Two reasons may account for the wide use of GRC counts in surveys (Coughlin, 1990; Groves et al., 2009; Fu et al., 2020; Schaeffer and Presser, 2003). First, respondents are less likely to report exact frequencies when sensitive topics are at stake. Because a further demand for an exact enumeration of sensitive behaviors (e.g., substance use, sex intercourse, and juvenile delinquency) would on the contrary result in excessive missing values, the use of GRC counts as response categories actually allow survey investigators to gain valuable albeit incomplete information. Second, even when respondents are willing to cooperate, recall bias will make their accurate enumeration unreliable and thus discourage a direct estimation of frequencies/counts in questionnaire designs. It is not surprising to find that major social and epidemiological surveys targeting children, adolescents, the elderly, or other individuals with possibly lower cognitive capacities widely adopt GRC count responses for questions on behavioral frequencies (Bauman et al., 2013; Johnston et al., 2017; Kann et al., 2018; Voorrips et al., 1991).

Regression models for right-censored (count) data have long been developed (e.g., Brännäs (1992); Cameron and Trivedi (2013); Gross and Lai (1996); Li and Ma (2010); Sinha et al. (1994) and the regression analysis of right-censored counts can be readily implemented by existing software packages (Raciborski, 2011). However, when right-censored counts are also grouped, the development of regression models and computation tools have received remarkably little attention. Despite large volumes of GRC count data collected across different disciplines and research settings, this type of count data has not been adequately exploited in empirical research. Due to the absence of statistical methods and computing tools, one conventional view shared by scholars across different disciplines is that these GRC counts can only be treated as categories and subsequently be analyzed by (ordered/multinomial) logistic regression models, if not by descriptive methods. Obviously, when counts are being treated as (ordinal) categories, logistic regression models fail to utilize the rich information embedded in their data structure.

Existing literature suggests that Poisson-based regression models are most suitable for analyzing count data, whereas logistic regression models are mainly designed for analyzing categorical data (e.g., Cameron and Trivedi (2013); Hall (2000); Lambert (1992)). Despite a clear methodological advantage over logistic regression models, a Poisson approach to modeling GRC counts are thwarted by several challenges. First, existing Poisson likelihood estimators should be modified to take both grouping and right-censoring into account. Moreover, the modified estimator(s) should be applicable to overdispersed GRC counts given the fact that observed counts are often overdispersed (Hall, 2000; Lambert, 1992; Young et al., 2017). Second, to implement Poisson-based regression models of GRC counts with covariates, any attempt should start with designing data generating processes for GRC counts and deriving asymptotic properties in a framework of generalized linear models. Third, because software packages for regression analysis of GRC counts have yet to be developed, different algorithms and search strategies need to be assessed to compute regression estimates. This study attempts to address all these issues.

2 Modeling Grouped and Right-censored Counts in Surveys

In this study, we implement modified Poisson regression models to analyze GRC counts. To illustrate the conceptual and methodological differences between these models and other existing methods, we review several prior approaches to count or censored data. While the use of GRC counts in survey research leads to grouped counts, misreporting of the exact count can also produce heaping at common counts (e.g., 5, 10, 15) and a mixture of exact and coarsened counts (Wang and Heitjan, 2008; Cummings et al., 2015; Zinn and Würbach, 2016). According to survey methodologists (Schaeffer and Presser, 2003), heaped counts (also known as digital preference) are a product of "satisficing" strategies adopted by respondents who try to conserve energy and provide seemingly good enough answers to survey questions. Heaped counts then correspond to measurement errors when exact enumeration is required in surveys (Cummings et al., 2015). In contrast, grouped counts correspond to design errors when GRC counts are used as survey instruments. Because GRC counts do not require exact enumeration, they may provide a plausible solution to digital preference by pushing errors in the measurement stage back to errors in the design stage. Although measurement errors are largely beyond the control of researchers, optimum experimental designs can be used to reduce design errors (Atkinson et al., 2007).

Two general approaches to heaped counts have been implemented (Cummings et al., 2015; Zinn and Würbach, 2016). The interval-regression approach treats that the heaping multiples as interval-censored counts, while the rescaled-mixture approach assumes that one group of respondents report k times of the requested counts over a $1/k$ interval of the reference period (Wang and Heitjan, 2008; Cummings et al., 2015). The use of these approaches to heaped counts relies on the specification of heaping multiples and understanding of misreporting patterns. Our models presented here do not require such information at the measurement stage.

Other regression models have been implemented to deal with censored count data (Cameron and Trivedi, 2013; Raciborski, 2011). Yet, these existing tools including the -cpoisson- and -rcpoisson- commands in Stata cannot consider overdispersion of censored counts. Moreover, only one censored interval, either is right-censoring, or left-censoring, or interval-censoring, can be considered by these tools. For regression models which can consider an outcome with multiple censored intervals, they have been implemented in the context of a continuous rather than count outcome (Royston, 2007). Again, the overdispersion of censored counts cannot be addressed by these models. This research aims to provide flexible tools for modeling grouped and right-censored counts with or without overdispersion.

Two issues related to the modeling of zeros (e.g., “never”) in GRC counts should further be noted. First, the zero count is often contained in a separate group and does not combine with other frequency counts (e.g., once or twice) in the design of GRC response categories. The rationale is that the prevalence of an event/behavior being studied can still be estimated even though the expected frequency cannot. Second, the modeling of zeros is actually also emphasized in the analysis of overdispersed counts. In general, two regression approaches to overdispersed counts exist (Hall, 2000; Lambert, 1992; Young et al., 2017). The zero-inflated Poisson approach uses a binomial distribution to model excessive zeros, while the negative binomial approach uses a multiplicative effect (i.e., the shape parameter) to denote unobserved heterogeneity. Due to space limit, this study adopts the zero-inflated Poisson approach to overdispersion. But the negative-binomial approach to GRC counts can also be developed based on the analytical procedures described here.

3 Methods

3.1 Modified Poisson Estimators

Consider the random variable Y from a Poisson distribution Pois(μ) with mean μ,

$$\text{Prob}(Y = k) = e^{-\mu}\frac{\mu^k}{k!}, \quad k = 0, 1, \cdots,$$

where the mean and the variance of Y are both μ. In Poisson regression, the expected frequency μ is determined by a linear combination of predictors $X_0, \cdots, X_d \in \mathbb{R}$ through an invertible link function $g_\mu : (0, \infty) \to \mathbb{R}$,

$$\mu = g_\mu^{-1}(\beta_0^* X_0 + \cdots + \beta_d^* X_d), \tag{1}$$

where $\beta_0^*, \cdots, \beta_d^* \in \mathbb{R}$ are unknown coefficients to be inferred from data. X_0 is set to be 1 for models with an intercept. We next consider a Poisson-multinomial approach to GRC counts (Fu et al. (2018)). First, we divide all the non-negative integers into $N \geq 1$ groups, where N corresponds to the total number of groups predetermined by a survey investigator. For each $1 \leq i \leq N$, the i-th group consists of one, or a successive sequence of integers,

$$\text{Group}_i = \{k \in \mathbb{N} : l_i \leq k < l_{i+1}\},$$

where $\mathbb{N} = \{0, 1, 2, \cdots\}$ is the totality of all non-negative integers, and we use $0 = l_1 < l_2 < \cdots < l_{N+1} = \infty$ to mark the boundaries of these N groups. If we use $\mathcal{G} = \{l_i\}_{i=1}^{N+1}$ to denote a grouped and right-censored grouping scheme, and combine the probability masses of $\text{Pois}(\mu)$ to form a multinomial distribution $M(\theta_1, \cdots, \theta_N)$ with one trial, the observation of the counts takes the form $Y_{\mathcal{G}} \sim M(\theta_1, \cdots, \theta_N)$ instead of $Y \sim \text{Pois}(\mu)$, such that for $1 \leq j \leq N$,

$$\text{Prob}(Y_{\mathcal{G}} = j) = \theta^{\mathcal{G}}(j, \mu = g_\mu^{-1}(\boldsymbol{\beta}^{*T}\boldsymbol{X})) = \sum_{k=l_j}^{l_{j+1}-1} e^{-\mu}\frac{\mu^k}{k!} \tag{2}$$

where $\boldsymbol{X} = (X_0, \cdots, X_d)^T$ and $\boldsymbol{\beta}^* = (\beta_0^*, \cdots, \beta_d^*)^T$. Considering a sample $\left\{(\boldsymbol{X}^i, Y_{\mathcal{G}}^i)\right\}_{i=1}^n$, the log-likelihood function of the model above is

$$\ell_n^{\text{Pois}}(\boldsymbol{\beta}) = \sum_{i=1}^n \log \theta^{\mathcal{G}}(Y_{\mathcal{G}}^i, g_\mu^{-1}(\boldsymbol{\beta}^T\boldsymbol{X}^i)). \tag{3}$$

The equi-dispersion assumption (that the mean equals the variance) under the Poisson distribution is violated when an empirical distribution has excessive zeros. The zero-inflated Poisson (ZIP) distribution is proposed to address excessive zeros (Hall, 2000; Lambert, 1992). For a count variable $Y \sim \text{ZIP}(\mu, p)$ with $\mu > 0$ and $0 < p < 1$, we have

$$\text{Prob}(Y = k) = \begin{cases} p + (1-p)e^{-\mu}, & k = 0, \\ (1-p)e^{-\mu}\frac{\mu^k}{k!}, & k \geq 1, \end{cases}$$

where $(1-p)$ is the proportion of population subject to $\text{Pois}(\mu)$. $Y_{\mathcal{G}}$ is again modeled by a multinomial distribution $M(\theta_1, \cdots, \theta_N)$ where each θ_i is a combined $\text{ZIP}(\mu, p)$ probability of the corresponding group. In addition to $\boldsymbol{X}$, we assume another set of predictors $\mathbf{U} = (U_0, \cdots, U_{d'})^T \in \mathbb{R}^{d'+1}$, their corresponding coefficients $\boldsymbol{\gamma}^* = (\gamma_0^*, \cdots, \gamma_{d'}^*)^T$, and an invertible link function $g_p : (0,1) \to \mathbb{R}$, to model the parameter $p = g_p^{-1}(\boldsymbol{\gamma}^{*T}\boldsymbol{U})$. Note that $\boldsymbol{U}$ and $\boldsymbol{X}$ may share some common predictors, and they may even be the same vector. Now the data set takes the form $\{(\boldsymbol{X}^i, \boldsymbol{U}^i, Y_{\mathcal{G}}^i)\}_{i=1}^n$, and we write the log-likelihood function as

$$\ell_n^{\text{ZIP}}(\boldsymbol{\beta}, \boldsymbol{\gamma}) = \sum_{i=1}^n \log \theta^{\mathcal{G}}(Y_{\mathcal{G}}^i, g_\mu^{-1}(\boldsymbol{\beta}^T\boldsymbol{X}^i), g_p^{-1}(\boldsymbol{\gamma}^T\boldsymbol{U}^i)). \tag{4}$$

Different link functions can be adopted as g_μ and g_p, including the log link function, $g_{\textsf{log}} : (0, \infty) \to \mathbb{R}$,

$$g_{\textsf{log}}(\lambda) = \log \lambda, \quad g_{\textsf{log}}^{-1}(t) = e^t, \tag{5}$$

and the logit link $g_{\textsf{logit}} : (0,1) \to \mathbb{R}$,

$$g_{\textsf{logit}}(\lambda) = \log\left(\frac{\lambda}{1-\lambda}\right), \quad g_{\textsf{logit}}^{-1}(t) = (1+e^{-t})^{-1}.$$

For any C^2 multivariate function[1] f, we have ∇f and $\text{Hess}(f)$ as its gradient vector and Hessian matrix, respectively. Next we give the consistency and asymptotic efficiency of the MLE's defined by (3) and (4), respectively. Our discussion below draws on the assumption of stochastic regressors (Fahrmeir and Kaufmann, 1985; van der Vaart, 1998). Let $\rho^{\text{GL,P}}$ be a Borel probability measure on $\mathbb{R}^{d+1} \times \{1, \ldots, N\}$ with marginal distribution $\rho_X^{\text{GL,P}}$ on $\mathbb{R}^{d+1}$. Assume that for a.s. $\boldsymbol{X} \sim \rho_X^{\text{GL,P}}$, the conditional distribution $\rho^{\text{GL,P}}(\cdot|\boldsymbol{X})$ is defined by (2) with $\mu = g_\mu^{-1}(\boldsymbol{\beta}^{*T}\boldsymbol{X})$. Let $\mathbf{0}$ denote a zero vector. The following theorem demonstrates the asymptotic existence, consistency, and asymptotic normality of MLE based on model (3).

Theorem 3.1. *Assume that*

1. *The marginal distribution $\rho_X^{\text{GL,P}}$ is supported on a compact set $\mathcal{X} \subset \mathbb{R}^{d+1}$, and g_μ^{-1} is C^2 with $(g_\mu^{-1})' > 0$ everywhere;*
2. *$\int_{\mathbb{R}^{d+1}} \langle \boldsymbol{u}, \boldsymbol{x}\rangle^2 \, d\rho_X^{\text{GL,P}}(\boldsymbol{x}) > 0$ for any $\boldsymbol{u} \in \mathbb{R}^{d+1} \backslash \{\mathbf{0}\}$;*
3. *$N \geq 2$.*

There then exists a sequence $\hat{\boldsymbol{\beta}}_n$ of random vectors and a random integer n_1, such that as the sample size $n \to \infty$,

(i). $\text{Prob}\left(\nabla \ell_n^{\text{Pois}}(\hat{\boldsymbol{\beta}}_n) = \mathbf{0} \text{ for all } n \geq n_1\right) = 1$ *(asymptotic existence);*

(ii). $\hat{\boldsymbol{\beta}}_n \xrightarrow{a.s.} \boldsymbol{\beta}^*$ *(strong consistency);*

(iii). The Fisher information matrix $\mathbb{F}(\boldsymbol{\beta}) := -\mathbb{E}[\text{Hess}(\ell_1^{\text{Pois}}(\boldsymbol{\beta}))]$ exists, and it is strictly positive definite for any $\boldsymbol{\beta}$. Moreover, $\sqrt{n}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}^*) \xrightarrow{Law} \mathcal{N}(\mathbf{0}, (\mathbb{F}(\boldsymbol{\beta}^*))^{-1})$ (asymptotic normality).

In essence, the assumption 2 in Theorem 3.1 requires that the distribution $\rho_X^{\text{GL,P}}$ is not concentrated on any non-trivial subspace. For the ZIP case, we define $\rho^{\text{GL,ZIP}}$ as its joint probability distribution on $\mathbb{R}^{d+1} \times \mathbb{R}^{d'+1} \times \{1, \ldots, N\}$, and define $\rho_X^{\text{GL,ZIP}}$ ($\rho_U^{\text{GL,ZIP}}$), and $\rho^{\text{GL,ZIP}}(\cdot|\boldsymbol{X}, \boldsymbol{U})$ as its marginal and conditional distributions, respectively. The following theorem demonstrates the asymptotic existence, consistency, and asymptotic normality of MLE based on model (4).

Theorem 3.2. *Assume that*

1. *The marginal distributions $\rho_X^{\text{GL,ZIP}}$ and $\rho_U^{\text{GL,ZIP}}$ are compactly supported. Both the inverses of the link functions g_μ^{-1} and g_p^{-1} are C^2 with $(g_\mu^{-1})' > 0$ and $(g_p^{-1})' > 0$ everywhere;*
2. *For both $\rho = \rho_X^{\text{GL,ZIP}}$ and $\rho = \rho_U^{\text{GL,ZIP}}$, $\int \langle \boldsymbol{u}, \boldsymbol{x}\rangle^2 \, d\rho(\boldsymbol{x}) > 0$ for any $\boldsymbol{u} \neq \mathbf{0}$;*

[1] Here in this paper, a function is called C^2 if its first and second (partial) derivatives exist and are continuous.

3. $N \geq 3$.

Then, there exists a sequence $(\hat{\boldsymbol{\beta}}_n, \hat{\boldsymbol{\gamma}}_n)$ *of random vectors and a random integer* n_1*, such that as the sample size* $n \to \infty$*,*

(i). $\operatorname{Prob}\left(\nabla \ell_n^{\mathrm{ZIP}}(\hat{\boldsymbol{\beta}}_n, \hat{\boldsymbol{\gamma}}_n) = \mathbf{0} \text{ for all } n \geq n_1\right) = 1$ *(asymptotic existence);*

(ii). $\hat{\boldsymbol{\beta}}_n \xrightarrow{a.s.} \boldsymbol{\beta}^*$*, and* $\hat{\boldsymbol{\gamma}}_n \xrightarrow{a.s.} \boldsymbol{\gamma}^*$ *(strong consistency);*

(iii). The Fisher information matrix $\mathbb{F}(\boldsymbol{\beta}, \boldsymbol{\gamma}) := -\mathbb{E}[\operatorname{Hess}(\ell_1^{\mathrm{ZIP}}(\boldsymbol{\beta}, \boldsymbol{\gamma}))]$ *exists, and it is strictly positive definite for any* $\boldsymbol{\beta}$ *and* $\boldsymbol{\gamma}$*. Moreover,*

$$\sqrt{n}\left(\begin{pmatrix} \hat{\boldsymbol{\beta}}_n \\ \hat{\boldsymbol{\gamma}}_n \end{pmatrix} - \begin{pmatrix} \boldsymbol{\beta}^* \\ \boldsymbol{\gamma}^* \end{pmatrix}\right) \xrightarrow{Law} \mathcal{N}(\mathbf{0}, (\mathbb{F}(\boldsymbol{\beta}^*, \boldsymbol{\gamma}^*))^{-1}),$$

(asymptotic normality).

In Appendix A, we give a rigorous proof of Theorems 3.1 and 3.2 by developing a more general framework of generalized linear models for GRC counts.

3.2 Computing Modified Poisson Estimates

We define the log-likelihood functions for the Poisson and ZIP cases by (3) and (4), respectively. The maximizers of these log-likelihood functions are modified Poisson and ZIP estimates respectively. To compute these modified Poisson estimates, we develop a hybrid line search algorithm, which consists of both first-order gradient method and second-order iteratively-reweighted-least-squares (IRLS, which is derived from the Newton-Raphson method) method. When GRC counts are at stake, we find that the use of this hybrid line search algorithm offers an effective way to compute estimates and avoid non-convergence. While Algorithm 1 presents step-by-step details of this algorithm, its logic can be further illustrated as follows.

For the Poisson case (the ZIP case can be analyzed similarly), we denote $\eta^i = \boldsymbol{\beta}^T \boldsymbol{X}^i$ and $\xi^i = g_\mu^{-1}(\eta^i)$ in (3), and write the partial derivative of the log-likelihood function with respect to a regression coefficient β_r, or a coordinate of the gradient, as:

$$\frac{\partial \ell_n^{\mathrm{Pois}}(\boldsymbol{\beta})}{\partial \beta_r} = \frac{\partial}{\partial \beta_r} \sum_{i=1}^n \log \theta^{\mathcal{G}}(Y_{\mathcal{G}}^i, \xi^i) = \sum_{i=1}^n \frac{1}{\theta^{\mathcal{G}}(Y_{\mathcal{G}}^i, \xi^i)} \cdot \frac{\partial \theta^{\mathcal{G}}(Y_{\mathcal{G}}^i, \xi^i)}{\partial \xi^i} \cdot \left(g_\mu^{-1}\right)'(\eta^i) \cdot X_r^i. \tag{6}$$

When the log link is used and counts are not grouped or right-censored, there is a simple analytical expression of (6), which is $\sum_{i=1}^n (Y^i - \xi^i) X_r^i$. However, this simple analytical expression no longer exists when counts are grouped and right-censored. Instead, the computation of (6) with GRC counts involves the summation of many (exponential) terms, which might be prone to overflow errors. Especially, overflow errors tend to take place if the initial choices of regression coefficients are distant from modified Poisson estimates in early iteration steps. We therefore evaluate the gradient using a finite-difference numerical method. Here, a numerical first-order instead of second-order method is implemented because a numerical approximation to the Hessian matrix $\operatorname{Hess}(\ell_n^{\mathrm{Pois}})$ is more time-consuming. We expect that this numerical first-order method mainly determines the search directions in early iteration steps.

Yet, the first-order method may converge slowly in final iteration steps due to the common zig-zagging behavior of a first-order method (Luenberger and Ye, 2016, Section 8.2). Instead, the second-order IRLS method should be used to diversify search directions and facilitate a quick convergence. Rather than using a numerical approximation, we draw on an analytical expression for $\mathrm{Hess}(\ell_n^{\mathrm{Pois}})$ to implement the IRLS for two reasons. First, when the search process is close to modified Poisson estimates in its final stage, overflow errors are no longer a serious concern. Second, as compared with a numerical approximation to $\mathrm{Hess}(\ell_n^{\mathrm{Pois}})$, an analytical method still provides a more precise evaluation of $\mathrm{Hess}(\ell_n^{\mathrm{Pois}})$ and subsequently yields more effective modified Poisson likelihood estimates.

The idea of our hybrid algorithm is further illustrated in Figure 1, where solid contour lines represent different levels of log-likelihood function ℓ_n^{Pois}. The dashed ellipse gives the contour of the second-order evaluation of ℓ_n^{Pois}, of which the location of the maximum is marked by the star. If we start from point A, the two arrows show the search directions of the gradient and IRLS methods, respectively. In line with our discussion above, we experimented with this hybrid search and found that, the gradient method helps to increase the log-likelihood in early iteration steps, while the IRLS method tends to cause overflow errors. However, the gradient method slows down at the final stage of the search process but the IRLS provides very quick convergence towards a maximizer. A critical issue is that we do not know exactly when the IRLS method starts to outperform the gradient method in providing a meaningful search direction. In practice, the answer to this question varies with model specification, variable selection, and datasets. Meanwhile, due to the complex structure of ℓ_n^{Pois} (e.g., its n summands with exponential functions), the gradient $\nabla\ell_n^{\mathrm{Pois}}$ can change drastically as $\boldsymbol{\beta}$ changes, which makes it difficult to determine an appropriate step size for gradient ascent. To address these practical concerns, we implement line searches in each iteration. By calculating and comparing the log-likelihood of multiple locations along the directions informed by both the gradient and the IRLS methods, our hybrid line search algorithm eventually converges to modified Poisson estimates.

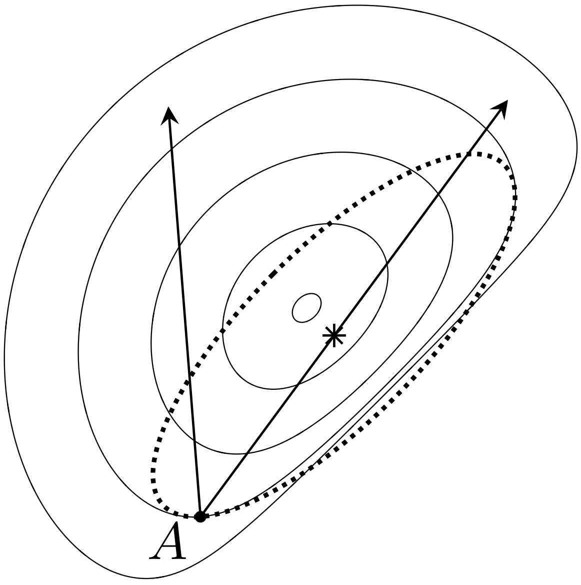

Figure 1: Graphical model representation of the hybrid line search algorithm

Algorithm 1: A hybrid line search algorithm for MLE

```
Input: the log-likelihood function $\ell_n$.
$m \leftarrow 0$, $\boldsymbol{\beta}^{(0)} \leftarrow \mathbf{0}$, $\mathsf{maxIter} \leftarrow 100$
Set $b > 1 > a > 0$.
$R \leftarrow \{ab^i : 1 \leq i \leq 15\}$, $P_{\mathsf{rel}} \leftarrow 10^{-8}$
repeat
    $m \leftarrow m + 1$
    $\boldsymbol{n}_1^* \leftarrow \nabla \ell_n(\boldsymbol{\beta}^{(m-1)})$ /* $\boldsymbol{n}_1^*$ calculated using a numerical method */
    $\boldsymbol{n}_1 \leftarrow \boldsymbol{n}_1^* / \|\boldsymbol{n}_1^*\|$, or $\mathbf{0}$ if $\boldsymbol{n}_1^*$ overflows.
    $\boldsymbol{n}_2^* \leftarrow -\mathrm{Hess}(\ell_n)(\boldsymbol{\beta}^{(m-1)})^{-1} \nabla \ell_n(\boldsymbol{\beta}^{(m-1)})$
    /* $\boldsymbol{n}_2^*$ calculated using the analytical expression */
    $\boldsymbol{n}_2 \leftarrow \boldsymbol{n}_2^* / \|\boldsymbol{n}_2^*\|$, or $\mathbf{0}$ if $\boldsymbol{n}_2^*$ overflows.
    $\mathcal{D} \leftarrow \{\mathbf{0}\} \cup \{r\boldsymbol{n}_1 : r \in R\} \cup \{r\boldsymbol{n}_2 : r \in R\}$
    $\boldsymbol{v} = \arg\max\{\ell_n(\boldsymbol{\beta}^{(m-1)} + \boldsymbol{u}) : \boldsymbol{u} \in \mathcal{D}\}$
    $\boldsymbol{\beta}^{(m)} \leftarrow \boldsymbol{v}$
until $m > \mathsf{maxIter}$, or $\|\boldsymbol{v} - \boldsymbol{\beta}^{(m-1)}\| \leq P_{\mathsf{rel}} \cdot \|\boldsymbol{\beta}^{(m-1)}\|$;
Output: $\boldsymbol{\beta}^{(m)}$
```

4 Simulation and Empirical Results

To assess finite-sample properties of the modified (zero-inflated) Poisson estimators, we experiment with artificial data using different sample sizes. For the Poisson parameter μ in both modified Poisson and ZIP models, we use a log link function with an intercept $\beta_0 = -1$. Two regressors X_1 and X_2 are independently drawn from the standard normal distribution, with their coefficients $\beta_1 = 1$ and $\beta_2 = 2$, respectively. For the ZIP case, we use a logit link function for its zero-inflation parameter p. This link function has an intercept $\gamma_0 = 1$ and one additional regressor γ_1 drawn independently from the standard normal distribution. The coefficient is set as $\gamma_1 = -1$. Two sets of results from simulation studies are presented in Table 1 (for the Poisson case) and Table 2 (for the ZIP case). For both sets of results, the finite-sample properties of the proposed methods are assessed based on sample sizes of $n = 100$, 400, 1200, 3600, and 10800, respectively. For each set of results with a specific sample size n, 1000 replications were carried out to calculate the averaged bias of estimates (**BIAS** in the tables), standard error (approximated by the standard deviation of estimates from replications and denoted by **SE** in the tables), the average of estimated standard errors from replications (**SEE** in the tables), and the empirical coverage probability of confidence intervals with a confidence level of 95% (**CP 95%** in the tables).

Results from Table 1 and Table 2 clearly suggest the validity of modified Poisson estimators. For the Poisson case (see Table 1), the bias in estimating coefficients appears to be acceptable even with a small sample size ($n = 100$) and it diminishes when the sample size gets larger. This observation also holds for the estimation of standard errors: with a small sample size ($n = 100$), the average of estimated standard errors is similar to the actual standard error, and their difference becomes negligible when sample size n increases. The empirical coverage probability stays very close to 95% and this pattern does not vary with sample size.

To illustrate the methodological advantage of modified Poisson estimators over conventional ways of modeling GRC counts, we also use ordered logistic regression models (proportional odds models with the parallel regression assumption) to estimate the partial effects of

X_1 and X_2 on the outcome effects and the results from simulations with 1,000 replications are also presented in Table 1. Here, the data generating model remains the same as that for simulation results of the modified Poisson model. Because ordered logistic regression models incorrectly treat counts as categories and cannot consider the design of grouping schemes, the bias in estimating β_1 and β_2 is quite large and does not appear to diminish with a larger sample size. While both **SE** and **SEE** tend to decrease with a larger sample size, the bias is so large that the 95% confidence intervals fail to contain the true values of β_1 and β_2 across all scenarios with $n \geq 100$. Again, these results suggest that neither should GRC counts be treated in conventional models as (ordinal) categories, nor would logistic models provide a satisfactory way to analyze GRC counts. It should be noted that, because ordered logistic regression uses several thresholds to model intercepts related to different categories in the GRC counts, a direct comparison between the two methods in the estimation of β_0 cannot be conducted.

For the modified zero-inflated Poisson estimators (see Table 2), possibly related to more parameters to be inferred, the averaged bias of regression estimates tends to be large with a small sample size ($n = 100$). However, the bias substantially reduces when the sample size is moderate ($n = 400$) or large. Similarly, the average of estimated standard errors does not provide a very accurate approximation of its corresponding standard error of the estimates when the sample size is small ($n = 100$). However, the average of estimated standard errors gets (very) close to the true standard error if the sample size increases. The empirical coverage probability somewhat fluctuates around 95% with a small ($n = 100$) or moderate ($n = 400$) sample size; yet it gets much closer to 95% when the sample size gets larger.

We also show an empirical application of modified Poisson estimators. Administered by the University of Michigan, the MTF (Monitoring the Future) project tracks annual changes in drug use and juvenile delinquency in the United States since the year 1975 (Johnston et al., 2017). As the largest repeated cross-sectional survey of its kind, the MTF studies students from hundreds of American middle and high schools each year. Data used in the current study are retrieved from the 2015 wave of the MTF project, which was released by the survey team for public use.

The existing literature suggests that adolescent marijuana use is strongly associated with their socio-demographic background, which guides our selection of regressors. A notable age increase in marijuana use has been documented from early to late adolescence, where male adolescents exhibit higher prevalence rates of marijuana use than females do (Chen and Jacobson, 2012; Finn, 2006). While mixed patterns have been observed on racial disparities in adolescent marijuana use (Miech et al., 2019), black adolescents often have lower rates of marijuana use than their white counterparts (Chen and Jacobson, 2012; Keyes et al., 2011). With regard to the impacts of family background, adolescents from single-parent families or with less-educated parents generally report higher levels of substance use (Barrett and Turner, 2006; Cambron et al., 2018). Finally, compared with adolescents living in rural areas, adolescents living in metropolitan areas tend to have higher levels of marijuana use (Martino et al., 2008).

The outcome variable and covariates are described as follows. One's lifetime frequency of marijuana use is measured by the survey's grouping scheme [never, 1-2 times, 3-5 times, 6-9 times, 10-19 times, 20-39 times, 40+ times]. For comparison, we consider the same set of eight variables (including intercepts). **Grade 10** and **grade 12** are dummy variables denoting the grade of respondents, with 8^{th} graders as reference. The demographic

Table 1: Modified Poisson estimators for GRC counts: Regression results based on 1,000 replications

n	Coef	Generalized Linear Models				Ordered Logistic Regression			
		BIAS	SE	SEE	CP 95%	BIAS	SE	SEE	CP 95%
100	β_0	-0.032	0.221	0.217	96.1%	–	–	–	–
	β_1	0.017	0.130	0.129	95.6%	1.090	0.456	0.389	14.8%
	β_2	0.025	0.189	0.183	95.5%	2.163	0.773	0.633	2.2%
400	β_0	-0.011	0.105	0.106	95.8%	–	–	–	–
	β_1	0.005	0.061	0.062	95.9%	1.002	0.199	0.183	0.0%
	β_2	0.011	0.087	0.089	96.0%	2.001	0.335	0.297	0.0%
1200	β_0	-0.005	0.059	0.061	96.5%	–	–	–	–
	β_1	0.000	0.034	0.035	95.2%	0.984	0.113	0.104	0.0%
	β_2	0.004	0.050	0.051	95.2%	1.977	0.190	0.170	0.0%
3600	β_0	0.000	0.034	0.035	95.7%	–	–	–	–
	β_1	0.000	0.019	0.020	96.0%	0.977	0.063	0.060	0.0%
	β_2	-0.000	0.029	0.029	95.0%	1.954	0.107	0.097	0.0%
10800	β_0	0.000	0.020	0.020	95.8%	–	–	–	–
	β_1	-0.000	0.011	0.012	95.2%	0.976	0.037	0.035	0.0%
	β_2	-0.001	0.017	0.017	95.7%	1.949	0.063	0.056	0.0%

Note: n: sample size; Coef: regression coefficient; BIAS: averaged bias of estimates; SE: standard error of the estimates; SEE: average of estimated standard errors; CP 95%: empirical coverage probability of 95% confidence intervals.

background of respondents is further indicated by **male** (versus *female*) and **black** (versus *non-African American*). **Intact Family** means both parents were present (coded as one) and the variable is coded as zero if otherwise. **Parental Education** is coded as one if either of the parents has completed college education and is coded as zero if otherwise. Finally, **metropolitan areas** indicates if a school was located in a metropolitan region (coded as one) or not (coded as zero).

As suggested by Table 3, male, more senior students and students from metropolitan areas were significantly associated with higher frequencies of marijuana use, while students from intact families or with college-educated parents were significantly associated with lower frequencies. Consistent with existing literature (Pacek et al., 2012; Finn, 2006), black adolescents reported significantly lower frequencies as compared with their counterparts. Results from the binomial part of the ZIP case showed that students from intact families or with college-educated parents were less exposed to marijuana use, while male, more senior students, and students living in metropolitan areas are more exposed to marijuana use. As expected, measures of goodness of fit, such as Akaike information criterion (**AIC**) and Bayesian information criterion (**BIC**), favor ZIP models over Poisson models. All models are weighted by survey weights.

Table 2: Modified zero-inflated Poisson estimators for GRC counts: Regression results based on 1,000 replications

n	Coef	BIAS	SE	SEE	CP 95%
100	β_0	−0.223	0.590	0.518	94.0%
	β_1	0.099	0.330	0.274	92.8%
	β_2	0.200	0.482	0.410	93.9%
	γ_0	−0.071	0.499	0.472	96.2%
	γ_1	−0.227	0.687	0.556	96.5%
400	β_0	−0.043	0.251	0.242	93.6%
	β_1	0.022	0.127	0.125	94.9%
	β_2	0.042	0.192	0.188	94.9%
	γ_0	−0.023	0.225	0.220	95.0%
	γ_1	−0.035	0.237	0.235	95.6%
1200	β_0	−0.006	0.143	0.139	94.5%
	β_1	0.002	0.071	0.071	94.8%
	β_2	0.008	0.110	0.107	94.9%
	γ_0	−0.003	0.127	0.126	94.9%
	γ_1	−0.014	0.133	0.133	94.4%
3600	β_0	−0.005	0.079	0.080	95.3%
	β_1	0.002	0.041	0.041	94.7%
	β_2	0.003	0.062	0.061	95.0%
	γ_0	−0.001	0.072	0.073	94.7%
	γ_1	−0.004	0.074	0.076	95.6%
10800	β_0	−0.000	0.049	0.046	93.9%
	β_1	0.000	0.024	0.024	94.7%
	β_2	0.001	0.037	0.035	93.5%
	γ_0	0.001	0.042	0.042	94.5%
	γ_1	−0.000	0.045	0.044	94.9%

Note: n: sample size; Coef: regression coefficient; BIAS: averaged bias of estimates; SE: standard error of the estimates; SEE: average of estimated standard errors; CP 95%: empirical coverage probability of 95% confidence intervals.

5 Discussion

Grouped and right-censored counts have been widely employed by survey investigators across different disciplines yet the conventional wisdom for modeling GRC data fails to take into account their latent count structure. By proposing a general framework of generalized linear models, we provide a valid tool, modified Poisson estimators, for much more precise estimation. Methodologically speaking, this general framework presented here has a clear advantage over conventional logistic regression analysis of GRC counts. This is because the former takes into account the design of grouping schemes and allows a direct assessment of counts, frequencies and rates, whereas the latter cannot consider the design of grouping schemes and requires GRC counts be collapsed into categories.

Modified Poisson estimators provide a flexible framework for analyzing GRC data regardless of whether the true data generating process is Poisson, zero-inflated Poisson, or other Poisson-based distributions. This framework has been developed, implemented and assessed in the present study as follows: we define/derive modified Poisson estimators and their asymptotic properties, develop a hybrid line search algorithm for parameter inference, demonstrate finite-sample performance of these estimators via simulation, and evaluate its empirical relevance based on survey data of marijuana use in America. Proof and findings from the current study evidently corroborate the validity and applicability of the generalized linear models for GRC counts.

Further efforts on generalized linear modeling of GRC counts are warranted. First and foremost, the relation between the choice of grouping schemes and the applicability of modified Poisson estimators needs to be further elucidated. A better understanding of the relation relies on future inquires into the design and optimality of GRC counts, and the accumulation of empirical applications in different research fields. Second, it is useful to extend the current focus on modified Poisson/ZIP regression models to other related parametric models of count data, such as negative binomial models, hurdle models, or zero-inflated negative binomial models (see Guo et al. (2020) for a discussion on negative-binomial models). Third, it is useful to incorporate other computing methods/algorithms in the modified Poisson estimators so that they can be applied to a wider range of research settings, such as random-effect analysis.

Appendix A A Proof of Asymptotic Properties

To prove Theorems 3.1 and 3.2, we now develop a general framework of generalized linear models for GRC counts, and these two theorems can be treated as corollaries. Let Y be a random variable, of which the probability density/mass function $f(y|\boldsymbol{\xi} = (\xi_1, \cdots, \xi_r))$ is parameterized by r continuous parameters. Here, the Poisson case only has μ to be estimated ($r = 1$) and the ZIP case has both μ and p to be estimated ($r = 2$). Assume that for each i, ξ_i is defined on an open interval $\mathcal{I}_i$, mapped to $\mathbb{R}$ by a homeomorphic link function g_i such that g_i^{-1} is C^2, and $(g_i^{-1})' > 0$ everywhere. We assume that $f(y|\boldsymbol{\xi})$ is a C^2 function of $\boldsymbol{\xi}$ on the (possibly unbounded) open "brick" $\mathcal{I}_1 \times \cdots \times \mathcal{I}_r$. For each parameter ξ_i, denote a vector of stochastic regressors $\boldsymbol{X}_i = (X_{i,0}, \cdots, X_{i,d_i})^T \in \mathbb{R}^{d_i+1}$, and a vector of their corresponding coefficients $\boldsymbol{\beta}_i = (\beta_{i,0}, \cdots, \beta_{i,d_i})^T \in \mathbb{R}^{d_i+1}$. Define

$$\xi_i = g_i^{-1}(\boldsymbol{\beta}_i^T \boldsymbol{X}_i).$$

We combine the regressors into one vector $\boldsymbol{X} = (\boldsymbol{X}_1^T, \cdots, \boldsymbol{X}_r^T)^T$, and use a pair of two numbers (s, s') to index the vector $\boldsymbol{X}$, where $1 \le s \le r$ and $0 \le s' \le d_s$. Similarly, we write $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^T, \cdots, \boldsymbol{\beta}_r^T)^T$ and $\boldsymbol{\xi} = \boldsymbol{\xi}(\boldsymbol{\beta}, \boldsymbol{X}) = (\xi_1 = g_1^{-1}(\boldsymbol{\beta}_1^T \boldsymbol{X}_1), \cdots, \xi_r)$. In particular, $X_{i,0}$ is fixed to be 1 for models with an intercept. Again, we use $\mathcal{G} = \{l_i\}_{i=1}^{N+1}$, $M(\theta_1, \cdots, \theta_N)$, and $Y_{\mathcal{G}} \in \{1, \cdots, N\}$ to denote a predetermined grouping scheme, its associated 1-trial multinomial distribution, and the (observed) occurrence of a group among the N possible groups, respectively.

We let $\mathbb{I} = \mathbb{I}(\mathcal{G}, \boldsymbol{\xi}) = (I_{s,t}^{\mathcal{G}})_{r \times r}$ be the Fisher information matrix of a random integer $1 \le Y_{\mathcal{G}} \le N$ with respect to parameters $\boldsymbol{\xi}$ (without considering regressors). That is,

$$I_{s,t}^{\mathcal{G}} = I_{s,t}^{\mathcal{G}}(\boldsymbol{\xi}) = \mathbb{E}\left[\frac{\partial}{\partial \xi_s} \log \theta^{\mathcal{G}}(Y_{\mathcal{G}}, \boldsymbol{\xi}) \frac{\partial}{\partial \xi_t} \log \theta^{\mathcal{G}}(Y_{\mathcal{G}}, \boldsymbol{\xi})\right], \tag{7}$$

where for $1 \le k \le N$,

$$\theta^{\mathcal{G}}(k, \boldsymbol{\xi}) = \sum_{j=l_k}^{l_{k+1}-1} f(j|\boldsymbol{\xi}) \tag{8}$$

gives the probability mass function of the N groups. Since $Y_{\mathcal{G}}$ takes values in the finite set $\{1, \cdots, N\}$ and the expectation (7) is a finite sum, $I_{s,t}^{\mathcal{G}}$ is well defined for all $\boldsymbol{\xi} \in \mathcal{I}_1 \times \cdots \times \mathcal{I}_r$.

Let $\{(\boldsymbol{X}^i, Y_{\mathcal{G}}^i)\}_{i=1}^n$ be a sample drawn independently from a distribution ρ on $\mathbb{R}^{d_1+1} \times \cdots \times \mathbb{R}^{d_r+1} \times \{1, \cdots, N\}$ with the true coefficient vector given by $\boldsymbol{\beta}^* = ((\boldsymbol{\beta}_1^*)^T, \cdots, (\boldsymbol{\beta}_r^*)^T)^T$. We write the log-likelihood function as

$$\ell_n(\boldsymbol{\beta}) = \sum_{i=1}^n \log \theta^{\mathcal{G}}(Y_{\mathcal{G}}^i, \boldsymbol{\xi}(\boldsymbol{\beta}, \boldsymbol{X}^i)).$$

To show large sample properties of estimators based on the above framework of generalized linear models for GRC counts, we write $d = d_1 + \cdots + d_r$, use $\|\cdot\|$ to denote the Euclidean norm, and have the following theorem.

Theorem A.1. *Assume that*

1. *The marginal distribution ρ_X of ρ on $\mathbb{R}^{d_1+1} \times \cdots \times \mathbb{R}^{d_r+1}$ is supported on a compact set $\mathcal{X}$. $\theta^{\mathcal{G}}(Y_{\mathcal{G}}, \boldsymbol{\xi})$ is a C^2 function of $\boldsymbol{\xi}$. For each $1 \le j \le s$, g_j^{-1} is C^2 with $(g_j^{-1})' > 0$ everywhere;*

2. *For any $1 \leq j \leq r$ and $\boldsymbol{u} \in \mathbb{R}^{d_j+1} \backslash \{\mathbf{0}\}$, $\int \langle \boldsymbol{u}, \boldsymbol{x} \rangle^2 \, d\rho_j(\boldsymbol{x}) > 0$, where ρ_j is the marginal distribution of ρ on $\mathbb{R}^{d_j+1}$;*

3. *The matrix $\mathbb{I}(\mathcal{G}, \boldsymbol{\xi})$ is continuous on $\boldsymbol{\xi} \in \mathcal{I}_1 \times \cdots \times \mathcal{I}_r$, and is strictly positive definite everywhere.*

Then, there exists a sequence $\hat{\boldsymbol{\beta}}_n$ of random vectors and a random integer n_1, such that as the sample size $n \to \infty$,

(i). $\text{Prob}\left(\nabla \ell_n(\hat{\boldsymbol{\beta}}_n) = \mathbf{0} \textit{ for all } n \geq n_1\right) = 1$ *(asymptotic existence);*

(ii). $\hat{\boldsymbol{\beta}}_n \xrightarrow{a.s.} \boldsymbol{\beta}^*$ *(strong consistency);*

(iii). The Fisher Information matrix $\mathbb{F}(\boldsymbol{\beta}) := -\mathbb{E}[\text{Hess}(\ell_1)(\boldsymbol{\beta})]$ exists, and it is strictly positive definite for any $\boldsymbol{\beta}$. Moreover, $\sqrt{n}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}^*) \xrightarrow{Law} \mathcal{N}(\mathbf{0}, (\mathbb{F}(\boldsymbol{\beta}^*))^{-1})$ (asymptotic normality).

Proof of Theorem A.1. For any $\boldsymbol{\beta} \in \mathbb{R}^{d+r}$, we write $\Delta\boldsymbol{\beta} = \boldsymbol{\beta} - \boldsymbol{\beta}^*$ and consider the Taylor expansion of the average log-likelihood,

$$\frac{1}{n}\left[\ell_n(\boldsymbol{\beta}) - \ell_n(\boldsymbol{\beta}^*)\right] = \frac{1}{n}\Delta\boldsymbol{\beta}^T \nabla \ell_n(\boldsymbol{\beta}^*) + \frac{1}{2n}\Delta\boldsymbol{\beta}^T \mathbf{H}_n(\tilde{\boldsymbol{\beta}})\Delta\boldsymbol{\beta}, \tag{9}$$

where $\mathbf{H}_n = \mathbf{H}_n^{\mathcal{G}} = \text{Hess}(\ell_n)$ and $\tilde{\boldsymbol{\beta}}$ is a vector between $\boldsymbol{\beta}$ and $\boldsymbol{\beta}^*$. Recall that the coefficient vector $\boldsymbol{\beta} \in \mathbb{R}^{d+r}$ has the structure $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^T, \cdots, \boldsymbol{\beta}_r^T)^T$. For each $1 \leq s \leq r$, $\boldsymbol{\beta}_s = (\beta_{s,0}, \cdots, \beta_{s,d_s})^T \in \mathbb{R}^{d_s+1}$ is a GLM coefficient vector associated with the model parameter ξ_s. For any fixed $\boldsymbol{\beta} \in \mathbb{R}^{d+r}$, $1 \leq s \leq r$, and $0 \leq s' \leq d_s$, the (s, s')-coordinate of $\nabla \ell_n(\boldsymbol{\beta})$ is

$$\left[\nabla \ell_n(\boldsymbol{\beta})\right]_{s,s'} = \sum_{i=1}^{n} \frac{\partial}{\partial \xi_s} \log \theta^{\mathcal{G}}(Y_{\mathcal{G}}^i, \boldsymbol{\xi}(\boldsymbol{\beta}, \boldsymbol{X}^i))(g_s^{-1})'(\boldsymbol{\beta}_s^T \boldsymbol{X}_s^i) X_{s,s'}^i. \tag{10}$$

Given the boundedness of $\mathcal{X}$ and the C^2 smoothness of g_s^{-1} and $\theta^{\mathcal{G}}$, every summand on the right-hand side of (10) is bounded. Since $Y_{\mathcal{G}}^i$ only takes values from the finite set $\{1, \ldots, N\}$, the conditional expectation $\mathbb{E}[\cdot|\boldsymbol{X}^i]$ of the i-th summand is just a sum of N terms. For notational simplicity we write $\theta_i^{\mathcal{G}} = \theta^{\mathcal{G}}(Y_{\mathcal{G}}^i, \boldsymbol{\xi}(\boldsymbol{\beta}, \boldsymbol{X}^i))$. In particular, when $\boldsymbol{\beta} = \boldsymbol{\beta}^*$, for any $1 \leq i \leq n$,

$$\mathbb{E}\left[\frac{\partial}{\partial \xi_s} \log \theta^{\mathcal{G}}(Y_{\mathcal{G}}^i, \boldsymbol{\xi}(\boldsymbol{\beta}^*, \boldsymbol{X}^i)) \middle| \boldsymbol{X}^i\right] = \frac{\partial}{\partial \xi_s} \sum_{j=1}^{N} \theta^{\mathcal{G}}(j, \boldsymbol{\xi}(\boldsymbol{\beta}^*, \boldsymbol{X}^i)) = \frac{\partial}{\partial \xi_s} 1 = 0.$$

So when $\boldsymbol{\beta} = \boldsymbol{\beta}^*$, the n summands in (10) are bounded i.i.d. random variables with zero means. According to the strong law of large numbers,

$$\frac{1}{n}\nabla \ell_n(\boldsymbol{\beta}^*) \xrightarrow{a.s.} \mathbf{0}, \quad \text{as } n \to \infty. \tag{11}$$

For any $1 \leq s, t \leq r$ with $0 \leq s' \leq d_s$ and $0 \leq t' \leq d_t$, recall that

$$[\mathbf{H}_n(\boldsymbol{\beta})]_{s,s',t,t'} = \sum_{i=1}^{n} \frac{\partial^2}{\partial \beta_{s,s'} \partial \beta_{t,t'}} \log \theta^{\mathcal{G}}(Y_{\mathcal{G}}^i, \boldsymbol{\xi}(\boldsymbol{\beta}, \boldsymbol{X}^i)). \tag{12}$$

Given the compactness of $\mathcal{X}$, and the assumption that $\theta^{\mathcal{G}}$ and $\{g_i^{-1}\}_{i=1}^r$ are all C^2, the summands in (12) are i.i.d. and bounded for any fixed $\boldsymbol{\beta} \in \mathbb{R}^{d+r}$. Thus they have expectations. For any $1 \leq i \leq n$,

$$\begin{aligned}
&\mathbb{E}\left[\frac{\partial^2}{\partial \beta_{s,s'} \partial \beta_{t,t'}} \log \theta^{\mathcal{G}}(Y_{\mathcal{G}}^i, \boldsymbol{\xi}(\boldsymbol{\beta}, \boldsymbol{X}^i)) \middle| \{\boldsymbol{X}^j\}_{j=1}^n \right] \\
=& \frac{\partial^2}{\partial \beta_{s,s'} \partial \beta_{t,t'}} \sum_{j=1}^{N} \theta^{\mathcal{G}}(j, \boldsymbol{\xi}(\boldsymbol{\beta}, \boldsymbol{X}^i)) - \mathbb{E}\left[\frac{\partial \log \theta_i^{\mathcal{G}}}{\partial \beta_{s,s'}} \frac{\partial \log \theta_i^{\mathcal{G}}}{\partial \beta_{t,t'}} \middle| \boldsymbol{X}^i \right] \\
=& 0 - \mathbb{E}\left[\frac{\partial \log \theta_i^{\mathcal{G}}}{\partial \xi_s} \frac{\partial \log \theta_i^{\mathcal{G}}}{\partial \xi_t} \middle| \boldsymbol{X}^i \right] \frac{\partial \xi_s}{\partial \beta_{s,s'}} \frac{\partial \xi_t}{\partial \beta_{t,t'}} \\
=& - I_{s,t}^{\mathcal{G}}(\boldsymbol{\xi}(\boldsymbol{\beta}, \boldsymbol{X}^i))(g_s^{-1})'(\boldsymbol{\beta}_s^T \boldsymbol{X}_s^i)(g_t^{-1})'(\boldsymbol{\beta}_t^T \boldsymbol{X}_t^i) X_{s,s'}^i X_{t,t'}^i.
\end{aligned} \tag{13}$$

Again, given the compactness of $\mathcal{X}$, the mean of the above conditional expectation exists and is exactly $[-\mathbb{F}(\boldsymbol{\beta})]_{s,s',t,t'}$. The continuity of $\mathbb{I}(\mathcal{G}, \boldsymbol{\xi})$ and $\{(g_s^{-1})'\}_{s=1}^r$, together with the dominated convergence theorem, guarantees that $\mathbb{F}(\boldsymbol{\beta})$ is continuous on $\mathbb{R}^{d+r}$.

For any $\boldsymbol{\beta} \in \mathbb{R}^{d+r}$, let $\boldsymbol{u} = (u_{1,0}, \cdots, u_{1,d_1}, u_{2,0}, \cdots, u_{r,d_r})$ be an arbitrary nonzero real vector. Based on (13), we have

$$\begin{aligned}
\langle u, \mathbb{F}(\boldsymbol{\beta}) u \rangle =& \sum_{s=1}^{r} \sum_{t=1}^{r} \sum_{s'=0}^{d_s} \sum_{t'=0}^{d_t} \mathbb{E}\left[I_{s,t}^{\mathcal{G}}(\boldsymbol{\xi})(g_s^{-1})'(\boldsymbol{\beta}_s^T \boldsymbol{X}_s)(g_t^{-1})'(\boldsymbol{\beta}_t^T \boldsymbol{X}_t) u_{s,s'} u_{t,t'} X_{s,s'} X_{t,t'} \right] \\
=& \mathbb{E} \sum_{s=1}^{r} \sum_{t=1}^{r} I_{s,t}^{\mathcal{G}}(\boldsymbol{\xi})(g_s^{-1})'(\boldsymbol{\beta}_s^T \boldsymbol{X}_s)(g_t^{-1})'(\boldsymbol{\beta}_t^T \boldsymbol{X}_t) \left(\boldsymbol{u}_s^T \boldsymbol{X}_s \right) \left(\boldsymbol{u}_t^T \boldsymbol{X}_t \right) \\
\geq& \mathbb{E} \sigma_{\mathsf{min}}(\mathbb{I}) \sum_{s=1}^{r} \left((g_s^{-1})'(\boldsymbol{\beta}_s^T \boldsymbol{X}_s)(\boldsymbol{u}_s^T \boldsymbol{X}_s) \right)^2 \\
=& \sum_{s=1}^{r} \int_{\mathcal{X}_s} \left(\sigma_{\mathsf{min}}(\mathbb{I})(g_s^{-1})'(\boldsymbol{\beta}_s^T \boldsymbol{x})^2 \right) \langle \boldsymbol{u}_s, \boldsymbol{x} \rangle^2 \, d\rho_s(\boldsymbol{x}),
\end{aligned}$$

where $\mathcal{X}_s$ is the compact support of ρ_s. The continuity and the positiveness of $\sigma_{\mathsf{min}}(\mathbb{I})$ and $(g_s^{-1})'$ implies that there exists some constant $C_s > 0$ such that

$$\sigma_{\mathsf{min}}(\mathbb{I})(g_s^{-1})'(\boldsymbol{\beta}_s^T \boldsymbol{x})^2 \geq C_s, \quad \text{for any } \boldsymbol{x} \text{ in } \mathcal{X}_s.$$

Since among $\boldsymbol{u}_1, \cdots, \boldsymbol{u}_s$, there is at least one non-zero vector, we derive from Assumption 2 that

$$\langle u, \mathbb{F}(\boldsymbol{\beta}) u \rangle > 0.$$

$\mathbb{F}(\boldsymbol{\beta})$ is thus strictly positive definite.

We now write $\mathsf{CM}(B_\varepsilon(\boldsymbol{\beta}^*))$ the space of all the $(d+r)\times(d+r)$ real symmetric matrix-valued continuous functions on $B_\varepsilon(\boldsymbol{\beta}^*)$, which is defined as $B_\varepsilon(\boldsymbol{\beta}^*) := \{\boldsymbol{\beta} \in \mathbb{R}^{d+r} : \|\boldsymbol{\beta} - \boldsymbol{\beta}^*\| \leq \varepsilon\}$ for some $\varepsilon > 0$. $\mathsf{CM}(B_\varepsilon(\boldsymbol{\beta}^*))$ is equipped with the norm

$$\|D\|_C = \max_{\boldsymbol{x}\in B_\varepsilon(\boldsymbol{\beta}^*)} \|D(\boldsymbol{x})\|_F,$$

where $\|\cdot\|_F$ denotes the Frobenius norm of matrices. With polynomials on $\mathbb{R}^{d+r} \supset B_\varepsilon(\boldsymbol{\beta}^*)$ with rational coefficients, it is easy to show that $\mathsf{CM}(B_\varepsilon(\boldsymbol{\beta}^*))$ is a separable Banach space. Again, considering the compactness of $\mathcal{X}$, and the C^2 smoothness of $\theta^{\mathcal{G}}$ and $\{g_i^{-1}\}_{i=1}^r$, the random variable $\|\mathbf{H}_1(\boldsymbol{\beta})\|_F$ is uniformly bounded on $B_\varepsilon(\boldsymbol{\beta}^*)$. Therefore, $\mathbb{E}\left[\|\mathbf{H}_1\|_C\right] < \infty$ and $\mathbb{E}\left[\mathbf{H}_1\right]$ exists. We see that $\frac{1}{n}\mathbf{H}_n$ is the average of n independent copies of $\mathbf{H}_1$, with $\mathbb{E}[\mathbf{H}_1(\boldsymbol{\beta})] = -\mathbb{F}(\boldsymbol{\beta})$ for any $\boldsymbol{\beta} \in B_\varepsilon(\boldsymbol{\beta}^*)$. According to the strong law of large numbers in the separable Banach spaces (see, e.g., Theorem 4.1.1 in (Padgett and Taylor, 1973, page 42)), we have

$$\text{Prob}\left(\lim_{n\to\infty}\left\|\frac{1}{n}\mathbf{H}_n + \mathbb{F}\right\|_C = 0\right) = 1. \tag{14}$$

If we denote $\sigma_{\mathsf{min}}(D)$ the minimum eigenvalue of D for any symmetric matrix D, σ_{min} is a Lipschitz continuous function on the space of symmetric matrices (Weyl's perturbation theorem, see, e.g., (Bhatia, 1997, page 63)). For any $E \in \mathsf{CM}(B_\varepsilon(\boldsymbol{\beta}^*))$, $\sigma_{\mathsf{min}}(E(\boldsymbol{\beta}))$ is then also a continuous function on $B_\varepsilon(\boldsymbol{\beta}^*)$ and the minimum

$$\lambda_{\mathsf{min}}(E) := \min_{\boldsymbol{\beta}\in B_\varepsilon(\boldsymbol{\beta}^*)} \sigma_{\mathsf{min}}(E(\boldsymbol{\beta}))$$

is achievable. For any $D, E \in \mathsf{CM}(B_\varepsilon(\boldsymbol{\beta}^*))$, we assume that $\lambda_{\mathsf{min}}(D) = \sigma_{\mathsf{min}}(D(\boldsymbol{\beta}_D)) \leq \lambda_{\mathsf{min}}(E) = \sigma_{\mathsf{min}}(E(\boldsymbol{\beta}_E))$ and obtain

$$\begin{aligned}|\lambda_{\mathsf{min}}(D) - \lambda_{\mathsf{min}}(E)| =&\sigma_{\mathsf{min}}(E(\boldsymbol{\beta}_E)) - \sigma_{\mathsf{min}}(E(\boldsymbol{\beta}_D)) + \sigma_{\mathsf{min}}(E(\boldsymbol{\beta}_D)) - \sigma_{\mathsf{min}}(D(\boldsymbol{\beta}_D))\\ \leq& 0 + \|E(\boldsymbol{\beta}_D) - D(\boldsymbol{\beta}_D)\|_F \leq \|E - D\|_C\,.\end{aligned}$$

λ_{min} is then (Lipschitz) continuous on $\mathsf{CM}(B_\varepsilon(\boldsymbol{\beta}^*))$. (14) further implies that

$$\lambda_{\mathsf{min}}\left(-\frac{1}{n}\mathbf{H}_n\right) \xrightarrow{a.s.} \lambda_{\mathsf{min}}(\mathbb{F}) > 0, \quad \text{as } n \to \infty.$$

For notational simplicity, we write below $\lambda_{\mathsf{min}} = \lambda_{\mathsf{min}}(\mathbb{F})$ and $\lambda_{\mathsf{min}}^n = \lambda_{\mathsf{min}}(-\frac{1}{n}\mathbf{H}_n)$. Obviously, λ_{min}^n is measurable.

Let $\tau_n = \left\|\frac{1}{n}\nabla\ell_n(\boldsymbol{\beta}^*)\right\|$. Since $\nabla\ell_n(\boldsymbol{\beta}^*)$ is measurable, so is τ_n. We then assume $\boldsymbol{\beta} \in \partial B_\varepsilon(\boldsymbol{\beta}^*)$, where $\partial B_\varepsilon(\boldsymbol{\beta}^*)$ is the boundary of the ball $B_\varepsilon(\boldsymbol{\beta}^*)$, and have $\|\Delta\boldsymbol{\beta}\| = \varepsilon$. Based on (11), we have $\tau_n \overset{a.s.}{\to} 0$. Define

$$m_n = \begin{cases} n, & \text{if } \sup_{k\geq n}\tau_k < \frac{\varepsilon}{4}\lambda_{\mathsf{min}}, \\ \infty, & \text{otherwise.}\end{cases}$$

m_n is clearly measurable. Similarly, define

$$m_n' = \begin{cases} n, & \text{if } \inf_{k\geq n} \lambda^k_{\mathsf{min}} > \frac{1}{2}\lambda_{\mathsf{min}} \\ \infty, & \text{otherwise,} \end{cases}$$

and $n_1 = \max\{\inf_n m_n, \inf_n m_n'\}$. We have that n_1 is measurable and $n_1 < \infty$ almost surely. For any $n \geq n_1$, $\tau_n < \frac{\varepsilon}{4}\lambda_{\mathsf{min}}$ and $\lambda^n_{\mathsf{min}} > \frac{1}{2}\lambda_{\mathsf{min}}$, so

$$\frac{1}{n}\Delta\boldsymbol{\beta}^T\nabla\ell_n(\boldsymbol{\beta}^*) \leq \tau_n\varepsilon < \frac{\varepsilon^2}{4}\lambda_{\mathsf{min}} < \frac{\varepsilon^2}{2}\lambda^n_{\mathsf{min}} \leq -\frac{1}{2n}\Delta\boldsymbol{\beta}^T\mathbf{H}_n(\tilde{\boldsymbol{\beta}})\Delta\boldsymbol{\beta}.$$

According to (9), this implies $\ell_n(\boldsymbol{\beta}^*) > \ell_n(\boldsymbol{\beta})$ for any $\boldsymbol{\beta} \in \partial B_\varepsilon(\boldsymbol{\beta}^*)$. This proves the existence of $\hat{\boldsymbol{\beta}}_n$ in $B_\varepsilon(\boldsymbol{\beta}^*)\backslash\partial B_\varepsilon(\boldsymbol{\beta}^*)$ such that $\nabla\ell_n(\hat{\boldsymbol{\beta}}_n) = 0$.

Here the definition of m_n' also suggests that for $n \geq n_1$, $\lambda^n_{\mathsf{min}} > \frac{1}{2}\lambda_{\mathsf{min}} > 0$, so ℓ_n is strictly concave on $B_\varepsilon(\boldsymbol{\beta}^*)$, and $\hat{\boldsymbol{\beta}}_n$ is thus the unique solution to $\nabla\ell_n = 0$ on $B_\varepsilon(\boldsymbol{\beta}^*)$. The measurability of $\hat{\boldsymbol{\beta}}_n$ can be easily developed when n is sufficiently large. To prove that $\hat{\boldsymbol{\beta}}_n$ is measurable when n is not sufficiently large, we define $\hat{\boldsymbol{\beta}}_n = \boldsymbol{\infty}$ when $n < n_1$. Next, we proceed to prove that each coordinate of $\hat{\boldsymbol{\beta}}_n$ is measurable. To do so, without loss of generality we fix the coordinate index (s, s') when $n < \infty$ and only let t varies. For any $-\infty < t < \infty$, by definition,

$$\{n < n_1\} \cap \left\{[\hat{\boldsymbol{\beta}}_n]_{s,s'} < t\right\} = \emptyset.$$

When $[\boldsymbol{\beta}^*]_{s,s'} + \varepsilon \leq t < \infty$, the set $\left\{[\hat{\boldsymbol{\beta}}_n]_{s,s'} < t\right\} \cap \{n \geq n_1\} = \{n \geq n_1\}$ is measurable. When $-\infty < t \leq [\boldsymbol{\beta}^*]_{s,s'} - \varepsilon$, by definition $\left\{[\hat{\boldsymbol{\beta}}_n]_{s,s'} < t\right\} \cap \{n \geq n_1\} = \emptyset$. When $[\boldsymbol{\beta}^*]_{s,s'} - \varepsilon < t < [\boldsymbol{\beta}^*]_{s,s'} + \varepsilon$, given the uniqueness of $\hat{\boldsymbol{\beta}}_n$,

$$\begin{aligned}\left\{[\hat{\boldsymbol{\beta}}_n]_{s,s'} < t\right\} \cap \{n \geq n_1\} &= \left\{\inf_{\beta_{s,s'} \geq t} \|\nabla\ell_n(\boldsymbol{\beta})\| > 0\right\} \cap \{n \geq n_1\} \\ &= \left\{\inf_{\boldsymbol{\beta} \in B, \beta_{s,s'} \geq t} \|\nabla\ell_n(\boldsymbol{\beta})\| > 0\right\} \cap \{n \geq n_1\},\end{aligned}$$

where $B \subset B_\varepsilon(\boldsymbol{\beta}^*)$ is a countable dense set. So $\left\{[\hat{\boldsymbol{\beta}}_n]_{s,s'} < t\right\} \cap \{n \geq n_1\}$ is also measurable. We have proved that the set $\left\{[\hat{\boldsymbol{\beta}}_n]_{s,s'} < t\right\}$ is measurable. So $\hat{\boldsymbol{\beta}}_n$ is measurable.

To prove (ii), let $\{\varepsilon_k\}_{k=1}^\infty$ (with $\varepsilon_1 = \varepsilon$) be a sequence of positive numbers decreasing to zero. For each $k \geq 2$, we repeat the above argument on the measurability of n_1 to define n_k, with ε substituted by ε_k. For any sufficiently small $\varepsilon_k > 0$, there exists $n_k < \infty$ such that whenever $n \geq n_k$, $\left\|\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}^*\right\| \leq \varepsilon_k$. This completes the proof of (ii).

To prove (iii), we apply the fundamental theorem of calculus to the difference $\nabla\ell_n(\hat{\boldsymbol{\beta}}_n) - \nabla\ell_n(\boldsymbol{\beta}^*)$. When $n \geq n_1$, $\nabla\ell_n(\hat{\boldsymbol{\beta}}_n) = 0$. So

$$-\frac{1}{n}\nabla\ell_n(\boldsymbol{\beta}^*) = \frac{1}{n}\nabla\ell_n(\hat{\boldsymbol{\beta}}) - \frac{1}{n}\nabla\ell_n(\boldsymbol{\beta}^*) = \left[\int_0^1 \frac{1}{n}\mathbf{H}_n(\boldsymbol{\beta}^* + t(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}^*))dt\right](\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}^*). \quad (15)$$

Since $\lambda^n_{\mathsf{min}} > \frac{1}{2}\lambda_{\mathsf{min}}$, for any $0 \le t \le 1$ we know that $-\mathbf{H}_n(\boldsymbol{\beta}^* + t(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}^*))$ is strictly positive definite. The integral in the square bracket in (15) is thus invertible because its maximum eigenvalue is bounded from above by $-\frac{1}{2}\lambda_{\mathsf{min}}$. We have

$$\left[\int_0^1 -\frac{1}{n}\mathbf{H}_n(\boldsymbol{\beta}^* + t(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}^*))\right]^{-1} \frac{1}{\sqrt{n}}\nabla \ell_n(\boldsymbol{\beta}^*) = \sqrt{n}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}^*). \tag{16}$$

By (14) and item (ii), we have almost surely that for any $\varepsilon > 0$ there exists some $n_1' \ge n_1$ such that, whenever $n \ge n_1'$, both of the following inequalities hold

$$\left\| -\frac{1}{n}\mathbf{H}_n(\boldsymbol{\beta}^* + t(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}^*)) - \mathbb{F}(\boldsymbol{\beta}^* + t(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}^*)) \right\|_F \le \frac{\varepsilon}{2}, \quad \text{and}$$
$$\left\| \mathbb{F}(\boldsymbol{\beta}^* + t(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}^*)) - \mathbb{F}(\boldsymbol{\beta}^*) \right\|_F \le \frac{\varepsilon}{2}.$$

Therefore, the absolute value of every entry of the matrix $-\frac{1}{n}\mathbf{H}_n(\boldsymbol{\beta}^* + t(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}^*))$ is bounded uniformly by $\|\mathbb{F}(\boldsymbol{\beta}^*)\|_F + \varepsilon$ for $0 \le t \le 1$. Also,

$$-\frac{1}{n}\mathbf{H}_n(\boldsymbol{\beta}^* + t(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}^*)) \xrightarrow{a.s.} \mathbb{F}(\boldsymbol{\beta}^*), \quad \text{uniformly for } 0 \le t \le 1 \text{ as } n \to \infty.$$

We apply the dominated convergence theorem and the continuity of matrix inversion on the space of strictly positive definite matrices to obtain that

$$\left[\int_0^1 -\frac{1}{n}\mathbf{H}_n(\boldsymbol{\beta}^* + t(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}^*))\right]^{-1} \xrightarrow{a.s.} (\mathbb{F}(\boldsymbol{\beta}^*))^{-1}, \quad \text{as } n \to \infty. \tag{17}$$

From (10) and (13), we see that $\nabla \ell_n(\boldsymbol{\beta}^*)$ is the sum of n i.i.d. random vectors, each of which has mean $\mathbf{0}$ and its covariance matrix $\mathbb{F}(\boldsymbol{\beta}^*)$. So, we apply the central limit theorem to obtain

$$\frac{1}{\sqrt{n}}\nabla \ell_n(\boldsymbol{\beta}^*) \xrightarrow{Law} \mathcal{N}(\mathbf{0}, \mathbb{F}(\boldsymbol{\beta}^*)),$$

which, together with (17) and (16), implies that

$$\sqrt{n}\left(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}^*\right) \xrightarrow{Law} \mathcal{N}(\mathbf{0}, [\mathbb{F}(\boldsymbol{\beta}^*)]^{-1}).$$

The proof is complete.

□

The proof of consistency and asymptotic normality mainly follows Fahrmeir and Kaufmann (1985). The proof of measurability of the MLE estimator follows Serfling (1980). Theorem 3.1 and Theorem 3.2 are direct corollaries of Theorem A.1. In fact, the smoothness of $\theta^{\mathcal{G}}$ on $\boldsymbol{\xi}$ is guaranteed by the analytical forms of Poisson and ZIP probability mass functions, respectively. It is easy to prove the strictly positive definiteness of $\mathbb{I}(\mathcal{G}, \boldsymbol{\xi})$, as well as its continuity on $\boldsymbol{\xi}$ for Poisson model with $N \ge 2$, and for ZIP model with $N \ge 3$, respectively Fu et al. (2020).

References

Ackard, D. M., Croll, J. K. and Kearney-Cooke, A. (2002) Dieting frequency among college females: Association with disordered eating, body image, and related psychological problems. *Journal of Psychosomatic Research*, **52**, 129–136.

Akers, R. L., La Greca, A. J., Cochran, J. and Sellers, C. (1989) Social learning theory and alcohol behavior among the elderly. *Sociological Quarterly*, **30**, 625–638.

Atkinson, A., Donev, A. and Tobias, R. (2007) *Optimum experimental designs, with SAS*. Oxford, UK: Oxford University Press.

Bachman, J. G., Johnston, L. D. and O'Malley, P. M. (1990) Explaining the recent decline in cocaine use among young adults: Further evidence that perceived risks and disapproval lead to reduced drug use. *Journal of Health and Social Behavior*, **31**, 173–184. URL: `http://www.jstor.org/stable/2137171`.

Barrett, A. E. and Turner, R. J. (2006) Family structure and substance use problems in adolescence and early adulthood: examining explanations for the relationship. *Addiction*, **101**, 109–120.

Bauman, S., Toomey, R. B. and Walker, J. L. (2013) Associations among bullying, cyberbullying, and suicide in high school students. *Journal of Adolescence*, **36**, 341 – 350. URL: `http://www.sciencedirect.com/science/article/pii/S0140197112001819`.

Bhatia, R. (1997) *Matrix analysis*, vol. 169 of *Graduate Texts in Mathematics*. Springer-Verlag, New York. URL: `https://doi.org/10.1007/978-1-4612-0653-8`.

Blair, E. and Burton, S. (1987) Cognitive processes used by survey respondents to answer behavioral frequency questions. *Journal of Consumer Research*, **14**, 280–288. URL: `https://doi.org/10.1086/209112`.

Brännäs, K. (1992) Limited dependent Poisson regression. *Journal of the Royal Statistical Society: Series D (The Statistician)*, **41**, 413–423. URL: `https://rss.onlinelibrary.wiley.com/doi/abs/10.2307/2349006`.

Cambron, C., Kosterman, R., Catalano, R. F., Guttmannova, K. and Hawkins, J. D. (2018) Neighborhood, family, and peer factors associated with early adolescent smoking and alcohol use. *Journal of youth and adolescence*, **47**, 369–382.

Cameron, A. C. and Trivedi, P. K. (2013) *Regression analysis of count data*, vol. 53. Cambridge, UK: Cambridge University Press.

Chen, P. and Jacobson, K. C. (2012) Developmental trajectories of substance use from early adolescence to young adulthood: Gender and racial/ethnic differences. *Journal of adolescent health*, **50**, 154–163.

Coughlin, S. S. (1990) Recall bias in epidemiologic studies. *Journal of Clinical Epidemiology*, **43**, 87 – 91. URL: `http://www.sciencedirect.com/science/article/pii/0895435690900603`.

Cummings, T. H., Hardin, J. W., McLain, A. C., Hussey, J. R., Bennett, K. J. and Wingood, G. M. (2015) Modeling heaped count data. *The Stata Journal*, **15**, 457–479.

Fahrmeir, L. and Kaufmann, H. (1985) Consistency and asymptotic normality of the maximum likelihood estimator in generalized linear models. *Ann. Statist.*, **13**, 342–368. URL: https://doi.org/10.1214/aos/1176346597.

Finn, K. V. (2006) Patterns of alcohol and marijuana use at school. *Journal of Research on Adolescence*, **16**, 69–77.

Fu, Q., Guo, X. and Land, K. C. (2018) A poisson-multinomial mixture approach to grouped and right-censored counts. *Communications in Statistics-Theory and Methods*, **47**, 427–447.

— (2020) Optimizing count responses in surveys: A machine-learning approach. *Sociological Methods & Research*, **49**, 637–671.

Gross, S. T. and Lai, T. L. (1996) Nonparametric estimation and regression analysis with left-truncated and right-censored data. *J. Amer. Statist. Assoc.*, **91**, 1166–1180. URL: https://doi.org/10.2307/2291735.

Groves, R. M., Fowler, F. J. J., Couper, M. P., Lepkowski, J. M., Singer, E. and Tourangeau, R. (2009) *Survey Methodology.* Hoboken, NJ: John Wiley & Sons.

Guo, X., Fu, Q., Wang, Y. and Land, K. C. (2020) A numerical method to compute fisher information for a special case of heterogeneous negative binomial regression. *Communications on Pure & Applied Analysis*, **19**, 4179–4189.

Hagan, J., Shedd, C. and Payne, M. R. (2005) Race, ethnicity, and youth perceptions of criminal injustice. *American Sociological Review*, **70**, 381–407. URL: https://doi.org/10.1177/000312240507000302.

Hall, D. B. (2000) Zero-inflated Poisson and binomial regression with random effects: a case study. *Biometrics*, **56**, 1030–1039. URL: https://doi.org/10.1111/j.0006-341X.2000.01030.x.

Harris, K. M. (2013) The add health study: Design and accomplishments. url: http://www.onlinelaege.com/pdf/c121.pdf. Accessed July 17, 2019.

Johnston, L. D., O'Malley, P. M., Miech, R. A., Bachman, J. G. and Schulenberg, J. E. (2017) Monitoring the future national survey results on drug use, 1975–2016: Overview, key findings on adolescent drug use. https://files.eric.ed.gov/fulltext/ED578534.pdf. Accessed July 17, 2019.

Kann, L., McManus, T., Harris, W. A., Shanklin, S. L., Flint, K. H., Queen, B., Lowry, R., Chyen, D., Whittle, L., Thornton, J., Lim, C., Bradford, D., Yamakawa, Y., Leon, M., Brener, N. and Ethier, K. A. (2018) Youth risk behavior surveillance - united states, 2017. *Morbidity and mortality weekly report. Surveillance summaries (Washington, D.C. : 2002)*, **67**, 1–114. URL: https://www.ncbi.nlm.nih.gov/pubmed/29902162. 29902162[pmid].

Keyes, K. M., Schulenberg, J. E., O'Malley, P. M., Johnston, L. D., Bachman, J. G., Li, G. and Hasin, D. (2011) The social norms of birth cohorts and adolescent marijuana use in the united states, 1976–2007. *Addiction*, **106**, 1790–1800.

Lambert, D. (1992) Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics*, **34**, 1–14. URL: https://www.tandfonline.com/doi/abs/10.1080/00401706.1992.10485228.

Li, J. and Ma, S. (2010) Interval-censored data with repeated measurements and a cured subgroup. *J. R. Stat. Soc. Ser. C. Appl. Stat.*, **59**, 693–705. URL: https://doi.org/10.1111/j.1467-9876.2009.00702.x.

Luenberger, D. G. and Ye, Y. (2016) *Linear and nonlinear programming*, vol. 228 of *International Series in Operations Research & Management Science*. Springer, Cham, fourth edn. URL: https://doi.org/10.1007/978-3-319-18842-3.

Marsden, P. V. (2003) Interviewer effects in measuring network size using a single name generator. *Social Networks*, **25**, 1 – 16. URL: http://www.sciencedirect.com/science/article/pii/S0378873302000096.

Martino, S. C., Ellickson, P. L. and McCaffrey, D. F. (2008) Developmental trajectories of substance use from early to late adolescence: A comparison of rural and urban youth. *Journal of studies on alcohol and drugs*, **69**, 430–440.

Miech, R., Terry-McElrath, Y. M., O'Malley, P. M. and Johnston, L. D. (2019) Increasing marijuana use for black adolescents in the united states: A test of competing explanations. *Addictive Behaviors*, **93**, 59–64.

Pacek, L. R., Malcolm, R. J. and Martins, S. S. (2012) Race/ethnicity differences between alcohol, marijuana, and co-occurring alcohol and marijuana use disorders and their association with public health and social problems using a national sample. *The American Journal on Addictions*, **21**, 435–444.

Padgett, W. J. and Taylor, R. L. (1973) *Laws of large numbers for normed linear spaces and certain Fréchet spaces.* Lecture Notes in Mathematics, Vol. 360. Springer-Verlag, Berlin-New York.

Raciborski, R. (2011) Right-censored poisson regression model. *The Stata Journal*, **11**, 95–105.

Royston, P. (2007) Multiple imputation of missing values: further update of ice, with an emphasis on interval censoring. *The Stata Journal*, **7**, 445–464.

Schaeffer, N. C. and Dykema, J. (2011) Questions for surveys: current trends and future directions. *Public opinion quarterly*, **75**, 909–961.

Schaeffer, N. C. and Presser, S. (2003) The science of asking questions. *Annual review of sociology*, **29**, 65–88.

Serfling, R. J. (1980) *Approximation theorems of mathematical statistics.* John Wiley & Sons, Inc., New York. Wiley Series in Probability and Mathematical Statistics.

Sinha, D., Tanner, M. A. and Hall, W. J. (1994) Maximization of the marginal likelihood of grouped survival data. *Biometrika*, **81**, 53–60. URL: https://doi.org/10.1093/biomet/81.1.53.

van der Vaart, A. W. (1998) *Asymptotic statistics*, vol. 3 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge, UK. URL: `https://doi.org/10.1017/CBO9780511802256`.

Voorrips, L., Ravelli, A., Dongelmans, P., Deurenberg, P. and Van Staveren, W. (1991) A physical activity questionnaire for the elderly. *Medicine and science in sports and exercise*, **23**, 974–979. URL: `http://europepmc.org/abstract/MED/1956274`.

Wang, H. and Heitjan, D. F. (2008) Modeling heaping in self-reported cigarette counts. *Statistics in medicine*, **27**, 3789–3804.

Young, D. S., Raim, A. M. and Johnson, N. R. (2017) Zero-inflated modelling for characterizing coverage errors of extracts from the us census bureau's master address file. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, **180**, 73–97.

Zinn, S. and Würbach, A. (2016) A statistical approach to address the problem of heaping in self-reported income data. *Journal of Applied Statistics*, **43**, 682–703.

Table 3: Regression estimates from generalized linear models, the MTF project, 2015

	Poisson Regression Estimates	
	Coefficient	95% Confidence Interval
Intercept	0.749***	$(0.703, 0.796)$
Grade 10	1.278***	$(1.243, 1.313)$
Grade 12	1.760***	$(1.723, 1.797)$
Male	0.325***	$(0.302, 0.348)$
Black	−0.095***	$(-0.130, -0.061)$
Intact Family	−0.472***	$(-0.497, -0.447)$
Parental Education	−0.407***	$(-0.433, -0.382)$
Metropolitan Areas	0.082***	$(0.054, 0.111)$
Log-likelihood	−59166.9417	
AIC	118300	
BIC	118400	

	Zero-inflated Poisson Estimates	
	Coefficient	95% Confidence Interval
Poisson, log link		
Intercept	2.482***	$(2.432, 2.533)$
Grade 10	0.473***	$(0.435, 0.511)$
Grade 12	0.594***	$(0.554, 0.634)$
Male	0.167***	$(0.142, 0.192)$
Black	−0.103***	$(-0.141, -0.066)$
Intact Family	−0.096***	$(-0.123, -0.069)$
Parental Education	−0.176***	$(-0.203, -0.148)$
Metropolitan Areas	−0.009	$(-0.040, 0.023)$
Bernoulli, logit link		
Intercept	1.302***	$(1.117, 1.487)$
Grade 10	−1.063***	$(-1.185, -0.940)$
Grade 12	−1.677***	$(-1.820, -1.534)$
Male	−0.239***	$(-0.339, -0.139)$
Black	−0.046	$(-0.194, 0.103)$
Intact Family	0.629***	$(0.515, 0.742)$
Parental Education	0.424***	$(0.309, 0.540)$
Metropolitan Areas	−0.162*	$(-0.288, -0.037)$
Log-likelihood	−23801.5767	
AIC	47640	
BIC	47750	

Note: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$. The total number of observations is 8,478.