Statistica Sinica Preprint No: SS-2025-0194	
Title	Post-Selection Inference in Generalized Linear Models via Parametric Programming
Manuscript ID	SS-2025-0194
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202025.0194
Complete List of Authors	Qinyan Shen, Karl Gregory and Xianzheng Huang
Corresponding Authors	Karl Gregory
E-mails	gregorkb@stat.sc.edu
Notice: Accepted author version.	

Post-selection inference in generalized linear models via parametric programming

Qinyan Shen, Karl Gregory, Xianzheng Huang

University of South Carolina

Abstract: We propose a unified framework to draw inferences for regression coefficients in a generalized linear model (GLM) following Lasso-based variable selection. We adapt to non-Gaussian GLMs a recently developed parametric programming strategy for post-selection inference in the linear model with a Gaussian response by drawing parallels between maximum likelihood estimation in GLMs and least squares estimation in linear models. We then conduct post-selection inference based on a linearized model for pseudo response and covariate data strategically created based on the raw data. Using synthetic data generated from regression models for three different types of non-Gaussian responses in simulation experiments, we demonstrate that the proposed method effectively corrects the naive inference that ignores variable selection while achieving greater efficiency than a polyhedral-based post-selection adjustment.

Key words and phrases: beta regression, Lasso, logistic regression, Poisson regression, selection event

1. Introduction

Traditional statistical inference assumes all hypotheses of interest are formulated prior to the observation of data. In regression contexts, practitioners often explore their data in order to select from a set of available variables a subset to use as covariates and, after fitting a model with these covariates, wish to make inferences on their effects. Such a strategy permits the data to dictate which hypotheses are ultimately tested, wherein lies the danger of incurring higher Type I error rates than intended. This danger has motivated the development of many post-selection inference methods to account for model selection. We may categorize these as *data splitting methods* (Wasserman and Roeder, 2009; Meinshausen et al., 2009; Rinaldo et al., 2019; Rasines and Young, 2023), *simultaneous inference methods* (Berk et al., 2013; Zhang and Cheng, 2017; Bachoc et al., 2019, 2020), and *conditioning methods* (Lee et al., 2016; Tibshirani et al., 2016; Taylor and Tibshirani, 2018; Kuchibhotla et al., 2020; Pirenne and Claeskens, 2024; Neufeld et al., 2022).

We here pursue a conditioning method, whereby we make inferences on a parameter in the selected model based on the conditional sampling distribution of a relevant statistic given the selection event. The seminal work Lee et al. (2016) studied the sampling distribution of a linear contrast of Gaussian responses in

a linear model given selection via L_1-penalization of the least-squares criterion, showing that the selection event can be characterized as the union of many polyhedra in the support of the response data. To obtain a more tractable sampling distribution of the statistic, these authors further condition on the signs of the selected regression coefficients. This additional conditioning costs efficiency in statistical inference, leading to wider-than-necessary confidence intervals that, although guaranteeing a nominal coverage probability, could have infinite expected width (Kivaranovic and Leeb, 2021). Some recent developments in this vein extend beyond linear models and Lasso regularization, where one seeks a useful conditional sampling distribution of the target statistic after intersecting the selection event with additional characteristics of the selected model so that the conditional event is (approximately) polyhedral in the support of the response data (Panigrahi and Taylor, 2023; Zhao et al., 2022; Shen et al., 2024; Taylor and Tibshirani, 2018). We refer to these methods as polyhedral methods.

Le Duy and Takeuchi (2021) made improvements to the work of Lee et al. (2016) by introducing a parametric programming (PP) approach to find the sampling distribution of a linear contrast of the response data conditional only on the selection event, avoiding the efficiency loss incurred by the sign-conditioning of the polyhedral method. Pirenne and Claeskens (2024) extended the PP approach to inference following model selection via adaptive Lasso, adaptive elastic net,

and group Lasso. In our work we adapt the PP approach to generalized linear models (GLMs) for non-Gaussian responses.

Our proposed strategy consists of two steps. Section 2 provides the development of the first step, in which we "linearize" the regression problem specified by a GLM for non-Gaussian data. Section 3 elaborates on the second step, where we apply the PP method to the linearized regression model. Section 4 describes in detail the implementation of the proposed method in three non-Gaussian regression settings. Section 5 presents simulation studies comparing the proposed method with the naive method (the method which ignores model selection), the polyhedral method. Section 6 presents three case studies in which we make inferences on covariate effects following variable selection using different types of non-Gaussian data arising from real-life applications. Section 7 outlines key takeaways and suggestions for future research.

2. Pre-selection inference in GLMs

Suppose we observe $(\mathbf{x}_1, Y_1), \ldots, (\mathbf{x}_n, Y_n)$, where $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^p$ are fixed covariate vectors and $Y_1, \ldots, Y_n$ are independent responses such that

$$Y_i \sim f(y; \eta_i, \phi) = \exp\left\{\frac{y\eta_i - b(\eta_i)}{a(\phi)} - c(y, \phi)\right\}, \tag{2.1}$$

where $\eta_i \equiv \beta_0 + \mathbf{x}_i^\top \boldsymbol{\beta}$, for $i = 1, \ldots, n$, where $a(\cdot)$, $b(\cdot)$ and $c(\cdot, \cdot)$ are known functions, ϕ is a dispersion parameter, and β_0 and $\boldsymbol{\beta} \equiv (\beta_1, \ldots, \beta_p)^\top$ are pa-

rameters with unknown values. This is the canonical generalized linear model (GLM); see McCullagh and Nelder (1989). Note that the mean and variance of Y_i are $b'(\eta_i)$ and $a(\phi)b''(\eta_i)$, respectively, where $b'(\cdot)$ and $b''(\cdot)$ are the first two derivatives of $b(\cdot)$. Throughout, let $M_0 = \{j \in \{1, \ldots, p\} : \beta_j \neq 0\}$ be the set of indices corresponding to nonzero regression coefficients. We will consider making inferences following the selection of a model $M \subset \{1, \ldots, p\}$.

2.1 Maximum likelihood estimation in a submodel

To focus on making inferences on the regression coefficients in the GLM specified by (2.1), we will for most of the paper assume ϕ is known. The maximum likelihood estimator (MLE), which we denote by $(\hat{\beta}_0, \hat{\boldsymbol{\beta}})$, of the tuple $(\beta_0, \boldsymbol{\beta})$ can be obtained by maximizing the log-likelihood

$$\ell_\phi(t_0, \mathbf{t}) \equiv \frac{1}{a(\phi)} \sum_{i=1}^{n} (Y_i \eta_i(t_0, \mathbf{t}) - b(\eta_i(t_0, \mathbf{t}))) - \sum_{i=1}^{n} c(Y_i, \phi)$$

over $(t_0, \mathbf{t}) \in \mathbb{R} \times \mathbb{R}^p$, where $\eta_i(t_0, \mathbf{t}) \equiv t_0 + \mathbf{x}_i^\top \mathbf{t}$. Instead of including all covariates, one may consider selecting a model $M \subset \{1, \ldots, p\}$ indexing which covariates to include in the construction of the linear predictors $\eta_1, \ldots, \eta_n$.

The MLE in model M, denoted by $(\hat{\beta}_{0,M}^{\text{mle}}, \hat{\boldsymbol{\beta}}_M^{\text{mle}})$, maximizes the function

$$\ell_{\phi,M}(t_0, \mathbf{t}) \equiv \frac{1}{a(\phi)} \sum_{i=1}^{n} (Y_i \eta_{i,M}(t_0, \mathbf{t}) - b(\eta_{i,M}(t_0, \mathbf{t}))) - \sum_{i=1}^{n} c(Y_i, \phi)$$

over $(t_0, \mathbf{t}) \in \mathbb{R} \times \mathbb{R}^{|M|}$, where $\eta_{i,M}(t_0, \mathbf{t}) \equiv t_0 + (\mathbf{x}_i)_M^\top \mathbf{t}$, $|M|$ is the cardinality of

M, and $(\mathbf{x}_i)_M$ is the vector constructed from the entries of $\mathbf{x}_i$ with indices in M. Under regularity conditions (White, 1982), as $n \to \infty$, $(\hat{\beta}^{\text{mle}}_{0,M}, \hat{\boldsymbol{\beta}}^{\text{mle}}_M)$ converges in probability to the limiting value of

$$(\beta^*_{0,M}, \boldsymbol{\beta}^*_M) \equiv \underset{(t_0,\mathbf{t})\in\mathbb{R}\times\mathbb{R}^{|M|}}{\arg\max} \; E\frac{1}{n}\{\ell_\phi(\beta_0, \boldsymbol{\beta}) - \ell_{\phi,M}(t_0, \mathbf{t})\},$$

where the maximand is proportional to the Kullback-Leibler divergence of the true model from model M. One may therefore regard $(\beta^*_{0,M}, \boldsymbol{\beta}^*_M)$ as the target of estimation when model M is considered. Denoting by $\boldsymbol{\beta}_M$ the vector containing the entries of $\boldsymbol{\beta}$ with indices in M, we will have $(\beta^*_{0,M}, \boldsymbol{\beta}^*_M) = (\beta_0, \boldsymbol{\beta}_M)$ if M contains all the indices in which $\boldsymbol{\beta}$ is nonzero, that is if $M \supset M_0$.

If model M results from variable selection based on the observed data, valid inferences for $\boldsymbol{\beta}^*_M$ using the same data should account for the selection event. In this study, we focus on Lasso-based variable selection. For example, one may select a model by finding the L_1-penalized MLE, defined as

$$(\hat{\beta}^{\text{mle}}_{0,\lambda}, \hat{\boldsymbol{\beta}}^{\text{mle}}_\lambda) \equiv \underset{(t_0,\mathbf{t})\in\mathbb{R}\times\mathbb{R}^p}{\arg\min} \Big\{ -\frac{1}{n}\ell_\phi(t_0, \mathbf{t}) + \lambda\|\mathbf{t}\|_1 \Big\}, \tag{2.2}$$

where $\|\mathbf{t}\|_1$ is the L_1-norm of $\mathbf{t}$ and $\lambda > 0$ is a tuning parameter governing the sparsity of $\hat{\boldsymbol{\beta}}^{\text{mle}}_\lambda$. A selected model is then $\hat{M}^{\text{mle}}_\lambda \equiv \{j : (\hat{\boldsymbol{\beta}}^{\text{mle}}_\lambda)_j \neq 0\}$. To make inferences on $(\boldsymbol{\beta}^*_M)_j$ based $(\hat{\boldsymbol{\beta}}^{\text{mle}}_M)_j$, where for a generic vector $\mathbf{v}$ we denote by $(\mathbf{v})_i$ entry i of $\mathbf{v}$, it is necessary to obtain the conditional sampling distribution of $(\hat{\boldsymbol{\beta}}^{\text{mle}}_M)_j$ given the selection event $\{\hat{M}^{\text{mle}}_\lambda = M\}$. However, this conditional

distribution does not appear to be analytically tractable in the GLM setting. For this reason we pursue a two-step strategy whereby we first "linearize" the GLM and then apply a post-selection inference method developed for the linear model.

2.2 Model linearization

In linear regression with a Gaussian response, the exact conditional distribution of $(\hat{\boldsymbol{\beta}}^{\text{mle}}_M)_j | \{\hat{M}^{\text{mle}}_\lambda = M\}$ has been found (Lee et al., 2016; Le Duy and Takeuchi, 2021). This motivates our strategy of linearizing the GLM and then performing model selection and post-selection inference in the linearized model. Our linearization step is inspired by the Newton-Raphson (NR) update leading to the MLEs of $(\beta_0, \boldsymbol{\beta})$, which we describe next.

For each $(t_0, \mathbf{t}) \in \mathbb{R} \times \mathbb{R}^p$ define $\mathbf{z}(t_0, \mathbf{t})$ as the vector with entries

$$z_i(t_0, \mathbf{t}) \equiv \sqrt{b''(\eta_i(t_0, \mathbf{t}))}\eta_i(t_0, \mathbf{t}) + \frac{Y_i - b'(\eta_i(t_0, \mathbf{t}))}{\sqrt{b''(\eta_i(t_0, \mathbf{t}))}},$$

the vector $\mathbf{u}_0(t_0, \mathbf{t})$ with entries $\sqrt{b''(\eta_i(t_0, \mathbf{t}))}$, and the matrix $\mathbf{U}(t_0, \mathbf{t})$ with rows $\sqrt{b''(\eta_i(t_0, \mathbf{t}))}\mathbf{x}_i^\top$ for $i = 1, \ldots, n$. Then, with initial value $(t_0^{(0)}, \mathbf{t}^{(0)})$, the iteratively reweighted least squares formulation of the NR update is

$$(t_0^{(k)}, \mathbf{t}^{(k)}) \leftarrow \underset{(t_0, \mathbf{t}) \in \mathbb{R} \times \mathbb{R}^p}{\arg\min} \|\mathbf{z}^{(k-1)} - (\mathbf{u}_0^{(k-1)} t_0 + \mathbf{U}^{(k-1)}\mathbf{t})\|^2, \tag{2.3}$$

where $\mathbf{z}^{(k-1)} \equiv \mathbf{z}(t_0^{(k-1)}, \mathbf{t}^{(k-1)})$, $\mathbf{u}_0^{(k-1)} \equiv \mathbf{u}_0(t_0^{(k-1)}, \mathbf{t}^{(k-1)})$, and $\mathbf{U}^{(k-1)} \equiv \mathbf{U}(t_0^{(k-1)}, \mathbf{t}^{(k-1)})$; see Davison (2003).

Upon convergence of $(t_0^{(k)}, \mathbf{t}^{(k)})$ to $(\hat{\beta}_0, \hat{\boldsymbol{\beta}})$, define $\hat{\mathbf{z}} \equiv \mathbf{z}(\hat{\beta}_0, \hat{\boldsymbol{\beta}})$, $\hat{\mathbf{u}}_0 \equiv \mathbf{u}(\hat{\beta}_0, \hat{\boldsymbol{\beta}})$, and $\hat{\mathbf{U}} \equiv \mathbf{U}(\hat{\beta}_0, \hat{\boldsymbol{\beta}})$. To focus on the parameters in $\boldsymbol{\beta}$, we now define "centered" versions of the response vector $\hat{\mathbf{z}}$ and the design matrix $\hat{\mathbf{U}}$. First define $\mathbf{P}_0(t_0, \mathbf{t}) \equiv \|\mathbf{u}_0(t_0, \mathbf{t})\|^{-2}\mathbf{u}_0(t_0, \mathbf{t})\mathbf{u}_0(t_0, \mathbf{t})^\top$ as well as $\mathbf{z}_0(t_0, \mathbf{t}) = (\mathbf{I} - \mathbf{P}_0(t_0, \mathbf{t}))\mathbf{z}(t_0, \mathbf{t})$ and $\mathbf{U}_0(t_0, \mathbf{t}) \equiv (\mathbf{I} - \mathbf{P}_0(t_0, \mathbf{t}))\mathbf{U}(t_0, \mathbf{t})$. Then set $\hat{\mathbf{P}}_0 \equiv \mathbf{P}_0(\hat{\beta}_0, \hat{\boldsymbol{\beta}})$ as well as $\hat{\mathbf{z}}_0 \equiv (\mathbf{I} - \hat{\mathbf{P}}_0)\hat{\mathbf{z}}$ and $\hat{\mathbf{U}}_0 \equiv (\mathbf{I} - \hat{\mathbf{P}}_0)\hat{\mathbf{U}}$, noting that $\hat{\mathbf{P}}_0$ is the orthogonal projection onto the space spanned by the "intercept" vector $\hat{\mathbf{u}}_0$.

Then we have $\hat{\boldsymbol{\beta}} = \arg\min_{\mathbf{t} \in \mathbb{R}^p} \|\hat{\mathbf{z}}_0 - \hat{\mathbf{U}}_0\mathbf{t}\|^2$, so that we may regard $\hat{\boldsymbol{\beta}}$ as the least squares estimator in linear regression with response vector $\hat{\mathbf{z}}_0$ and design matrix $\hat{\mathbf{U}}_0$. From here our strategy will be to treat the response vector $\hat{\mathbf{z}}_0$ as though it arose from a Gaussian linear model with design matrix $\hat{\mathbf{U}}_0$ and to apply a post-selection inference method developed for Gaussian linear regression.

Since our strategy is to treat $\hat{\mathbf{z}}_0$ and $\hat{\mathbf{U}}_0$ as Gaussian linear model data, we propose to select a model via L_1-penalization of the Gaussian log-likelihood with $\hat{\mathbf{z}}_0$ and $\hat{\mathbf{U}}_0$ plugged in. That is, we propose computing the sparse estimator

$$\hat{\boldsymbol{\beta}}_\lambda \equiv \underset{\mathbf{t} \in \mathbb{R}^p}{\arg\min} \frac{1}{2n}\|\hat{\mathbf{z}}_0 - \hat{\mathbf{U}}_0\mathbf{t}\|^2 + \lambda\|\mathbf{t}\|_1, \tag{2.4}$$

for some $\lambda > 0$ and selecting the model $\hat{M}_\lambda \equiv \{j : (\hat{\boldsymbol{\beta}}_\lambda)_j \neq 0\}$. Note that the model selected in this way may be distinct from the model $\hat{M}_\lambda^{\text{mle}}$ selected via L_1-penalization of the original GLM log-likelihood; however, in Section 2.3 we argue that these models should be reliably similar as $n \to \infty$. Then, on the event

that the model $\hat{M}_\lambda = M$ is chosen, we consider the conditional distribution of the vector $\hat{\boldsymbol{\beta}}_M \equiv \arg\min_{\mathbf{t}\in\mathbb{R}^p} \|\hat{\mathbf{z}}_0 - \hat{\mathbf{U}}_{0,M}\mathbf{t}\|^2$, the entries of which may be expressed as contrasts in the vector $\hat{\mathbf{z}}_0$ of the form

$$(\hat{\boldsymbol{\beta}}_M)_j = \mathbf{e}_j^T(\hat{\mathbf{U}}_{0,M}^\top\hat{\mathbf{U}}_{0,M})^{-1}\hat{\mathbf{U}}_{0,M}^\top\hat{\mathbf{z}}_0, \tag{2.5}$$

for $j = 1, \ldots, |M|$, where $\mathbf{e}_j$ is the $|M| \times 1$ vector with entry j equal to one and remaining entries equal to zero.

In order to study the conditional distributions of contrasts of the form in (2.5), we next introduce idealized counterparts to $\hat{\mathbf{z}}_0$ and $\hat{\mathbf{U}}_0$, which we denote by $\mathbf{z}_0$ and $\mathbf{U}_0$, which one would observe if one knew the true values of the parameters β_0 and $\boldsymbol{\beta}$.

2.3 The idealized linear model

The response vector $\hat{\mathbf{z}}$, the vector $\hat{\mathbf{u}}_0$, and design matrix $\hat{\mathbf{U}}$ can be viewed as approximations to unobservable, idealized counterparts $\mathbf{z}$, $\mathbf{u}_0$, and $\mathbf{U}$, respectively, which we define as $\mathbf{z} \equiv \mathbf{z}(\beta_0, \boldsymbol{\beta})$, $\mathbf{u}_0 \equiv \mathbf{u}_0(\beta_0, \boldsymbol{\beta})$, and $\mathbf{U} \equiv \mathbf{U}(\beta_0, \boldsymbol{\beta})$. Moreover, letting $\mathbf{z}_0 \equiv \mathbf{z}_0(\beta_0, \boldsymbol{\beta})$ and $\mathbf{U}_0 \equiv \mathbf{U}_0(\beta_0, \boldsymbol{\beta})$, we obtain idealized counterparts to the "centered" response vector $\hat{\mathbf{z}}_0$ and design matrix $\hat{\mathbf{U}}_0$.

If one could observe $\mathbf{z}_0$ and $\mathbf{U}_0$, one could base inferences on the idealized linear estimator of $\boldsymbol{\beta}$ given by $\tilde{\boldsymbol{\beta}} \equiv \arg\min_{\mathbf{t}\in\mathbb{R}^p} \|\mathbf{z}_0 - \mathbf{U}_0\mathbf{t}\|^2$. Likewise, one

could compute the idealized sparse estimator

$$\tilde{\boldsymbol{\beta}}_\lambda \equiv \underset{\mathbf{t}\in\mathbb{R}^p}{\arg\min}\, \frac{1}{2n}\|\mathbf{z}_0 - \mathbf{U}_0\mathbf{t}\|^2 + \lambda\|\mathbf{t}\|_1, \tag{2.6}$$

and the corresponding idealized selected model $\tilde{M}_\lambda \equiv \{j : (\tilde{\boldsymbol{\beta}}_\lambda)_j \neq 0\}$. Furthermore, on the event $\tilde{M}_\lambda = M$, one could make conditional inferences by considering the conditional distributions of the entries of the idealized estimator $\tilde{\boldsymbol{\beta}}_M \equiv \underset{\mathbf{t}\in\mathbb{R}^p}{\arg\min}\|\mathbf{z}_0 - \mathbf{U}_{0,M}\mathbf{t}\|^2$, which could be expressed as contrasts in the vector $\mathbf{z}_0$ with form

$$(\tilde{\boldsymbol{\beta}}_M)_j \equiv \mathbf{e}_j^\top(\mathbf{U}_{0,M}^\top\mathbf{U}_{0,M})^{-1}\mathbf{U}_{0,M}^\top\mathbf{z}_0. \tag{2.7}$$

Note that the contrast in (2.7) is an idealized version of the contrast in (2.5). Defining for each $M \subset \{1,\ldots,p\}$ and $j = 1,\ldots,|M|$ the vector

$$\mathbf{c}_{M,j}(t_0,\mathbf{t})^\top = \mathbf{e}_j^\top(\mathbf{U}_{0,M}(t_0,\mathbf{t})^\top\mathbf{U}_{0,M}(t_0,\mathbf{t}))^{-1}\mathbf{U}_{0,M}(t_0,\mathbf{t})^\top,$$

where $\mathbf{U}_{0,M}(t_0,\mathbf{t})$ is the matrix formed with the columns of $\mathbf{U}_0(t_0,\mathbf{t})$ having indices in M, we set $\hat{\mathbf{c}}_{M,j} \equiv \mathbf{c}_{M,j}(\hat{\beta}_0,\hat{\boldsymbol{\beta}})$, which has idealized counterpart $\mathbf{c}_{M,j} \equiv \mathbf{c}_{M,j}(\beta_0,\boldsymbol{\beta})$. This allows us to write (2.5) and (2.7) as $(\hat{\boldsymbol{\beta}}_M)_j = \hat{\mathbf{c}}_{M,j}^\top\hat{\mathbf{z}}_0$ and $(\tilde{\boldsymbol{\beta}}_M)_j = \mathbf{c}_{M,j}^\top\mathbf{z}_0$, respectively.

Our first result gives conditions under which, prior to model selection, one can make inferences based on the observable $\hat{\mathbf{z}}_0$ and $\hat{\mathbf{U}}_0$ which are asymptotically equivalent to those based on their idealized counterparts $\mathbf{z}_0$ and $\mathbf{U}_0$. To state our

result, define for any $(t_0, \mathbf{t}) \in \mathbb{R} \times \mathbb{R}^p$ the function

$$g_{M,j,n}(t_0, \mathbf{t}; \boldsymbol{\beta}) \equiv \frac{\mathbf{c}_{M,j}(t_0, \mathbf{t})^\top (\mathbf{z}(t_0, \mathbf{t}) - \mathbf{U}(t_0, \mathbf{t})\boldsymbol{\beta})}{\sqrt{\mathbf{c}_{M,j}(t_0, \mathbf{t})^\top \mathbf{c}_{M,j}(t_0, \mathbf{t})}}.$$

Then the function $g_{M,j,n}(t_0, \mathbf{t}; \boldsymbol{\beta})$ may be used to construct some useful pivotal quantities. Specifically we define

$$\hat{g}_{M,j,n}(\boldsymbol{\beta}) \equiv g_{M,j,n}(\hat{\beta}_0, \hat{\boldsymbol{\beta}}; \boldsymbol{\beta}) = \frac{\hat{\mathbf{c}}_{M,j}^\top (\hat{\mathbf{z}}_0 - \hat{\mathbf{U}}_0 \boldsymbol{\beta})}{\sqrt{\hat{\mathbf{c}}_{M,j}^\top \hat{\mathbf{c}}_{M,j}}}$$

as a feasible pivotal quantity and

$$g_{M,j,n}(\boldsymbol{\beta}) \equiv g_{M,j,n}(\beta_0, \boldsymbol{\beta}; \boldsymbol{\beta}) = \frac{\mathbf{c}_{M,j}^\top (\mathbf{z}_0 - \mathbf{U}_0 \boldsymbol{\beta})}{\sqrt{\mathbf{c}_{M,j}^\top \mathbf{c}_{M,j}}}$$

as its idealized counterpart. After stating an assumption, we can present our first main result.

Assumption 1. *Given $M \subset \{1, \ldots, p\}$ and an index $j = 1, \ldots, |M|$, suppose (i) $\|\mathbf{c}_{M,j}\|_\infty / \|\mathbf{c}_{M,j}\| \to 0$ as $n \to \infty$ and (ii) for some $n_0 \geq 1$ and $\delta, C \in [0, \infty)$,*

$$\mathbb{E}|g_{M,j,n}(t_0, \mathbf{t}; \boldsymbol{\beta}) - g_{M,j,n}(\beta_0, \boldsymbol{\beta}; \boldsymbol{\beta})| \leq C \|(t_0, \mathbf{t}^\top)^\top - (\beta_0, \boldsymbol{\beta}^\top)^\top\|$$

for all $n > n_0$ for all $(t_0, \mathbf{t})$ such that $\|(t_0, \mathbf{t}^\top)^\top - (\beta_0, \boldsymbol{\beta}^\top)^\top\| \leq \delta$.

Assumption 1(i) is mild and holds if the maximum leverage in the linear model with design matrix $\mathbf{U}_M$ converges to zero; see Huber (2011). Assumption 1(ii) is a smoothness condition on the function $g_n(t_0, \mathbf{t}; \boldsymbol{\beta})$ in the neighborhood

of the tuple $(\beta_0, \boldsymbol{\beta})$. Namely, it requires that the expected change in the (random) function $g_{M,j,n}(t_0, \mathbf{t}; \boldsymbol{\beta})$ as $(t_0, \mathbf{t})$ moves away from $(\beta_0, \boldsymbol{\beta})$ is bounded above by some constant times the distance between $(t_0, \mathbf{t})$ and $(\beta_0, \boldsymbol{\beta})$. Both assumptions describe conditions on the sequence of design vectors $\{\mathbf{x}_n\}_{n\geq 1}$.

Theorem 1. *Under Assumption 1 we have (i)* $g_{M,j,n}(\boldsymbol{\beta}) \xrightarrow{d} \mathcal{N}(0, a(\phi))$ *and (ii)* $|\hat{g}_{M,j,n}(\boldsymbol{\beta}) - g_{M,j,n}(\boldsymbol{\beta})| \xrightarrow{p} 0$ *as* $n \to \infty$.

The following corollary shows how Theorem 1 would enable inference on an entry of $\boldsymbol{\beta}_M$ based on the observable $\hat{\mathbf{z}}_0$ and $\hat{\mathbf{U}}_0$.

Corollary 1. *Under Assumption 1, if* $M \supset M_0$ *then, as* $n \to \infty$, *we have*

$$\frac{(\hat{\boldsymbol{\beta}}_M)_j - (\boldsymbol{\beta}_M)_j}{\sqrt{\mathbf{e}_j^T(\hat{\mathbf{U}}_{0,M}^\top \hat{\mathbf{U}}_{0,M})^{-1}\mathbf{e}_j}} \xrightarrow{d} \mathcal{N}(0, a(\phi)).$$

Note that if $M \not\supset M_0$, then the estimation targets in the linearized model become the entries of the vector

$$\tilde{\boldsymbol{\beta}}_M^* \equiv (\mathbf{U}_{0,M}^\top \mathbf{U}_{0,M})^{-1}\mathbf{U}_{0,M}^\top \mathbf{U}_0 \boldsymbol{\beta}. \tag{2.8}$$

Note that Theorem 1 and Corollary 1 give asymptotic distributions which are not yet conditioned on the selection of a model. In order to establish asymptotic equivalence of conditional inferences after model selection based on the observable linear model data $\hat{\mathbf{z}}_0$ and $\hat{\mathbf{U}}_0$ and those based on the idealized linear model data $\mathbf{z}_0$ and $\mathbf{U}_0$, we must investigate whether the selected model $\hat{M}_\lambda$ based

on the sparse estimator $\hat{\boldsymbol{\beta}}_\lambda$ in (2.4) and the selected model $\tilde{M}_\lambda$ based on $\tilde{\boldsymbol{\beta}}_\lambda$ in (2.6) will agree with high probability. If so, one may assume in every step of the analysis that one has observed the idealized response $\mathbf{z}_0$ and design matrix $\mathbf{U}_0$ instead of their observable counterparts $\hat{\mathbf{z}}_0$ and $\hat{\mathbf{U}}_0$. Proofs of Theorem 1 and Corollary 1 are given in the Supplementary Material.

Defining the vector of correlations $\tilde{\mathbf{r}}_\lambda \equiv \tilde{\mathbf{U}}_0^\top(\tilde{\mathbf{z}}_0 - \tilde{\mathbf{U}}_0\tilde{\boldsymbol{\beta}}_\lambda)$, the KKT conditions give that $|(\tilde{\mathbf{r}}_\lambda)_j| = \lambda$ for all $j \in \tilde{M}_\lambda$ and $|(\tilde{\mathbf{r}}_\lambda)_j| \leq \lambda$ for all $j \notin \tilde{M}_\lambda$. Likewise defining $\hat{\mathbf{r}}_\lambda \equiv \hat{\mathbf{U}}_0^\top(\hat{\mathbf{z}}_0 - \hat{\mathbf{U}}_0\hat{\boldsymbol{\beta}}_\lambda)$, we have $|(\hat{\mathbf{r}}_\lambda)_j| = \lambda$ for all $j \in \hat{M}_\lambda$ and $|(\hat{\mathbf{r}}_\lambda)_j| \leq \lambda$ for all $j \notin \hat{M}_\lambda$. If one assumes for the idealized pseudo-data that (i) $|(\tilde{\mathbf{r}}_\lambda)_j| \leq \lambda(1-\rho)$ for all $j \notin \tilde{M}_\lambda$ for some $\rho \in (0,1)$, a condition called strict dual feasibility (Wainwright, 2009), and (ii) $\min_{j\in\tilde{M}_\lambda} |(\tilde{\boldsymbol{\beta}}_\lambda)_j| \geq c > 0$, a so-called beta-min condition (Zhao and Yu, 2006), hold with probability tending to one as $n \to \infty$, then $P(\hat{M}_\lambda = \tilde{M}_\lambda) \to 1$ provided

$$\|\hat{\mathbf{r}}_\lambda - \tilde{\mathbf{r}}_\lambda\| \xrightarrow{p} 0 \quad \text{and} \quad \|\hat{\boldsymbol{\beta}}_\lambda - \tilde{\boldsymbol{\beta}}_\lambda\| \xrightarrow{p} 0 \tag{2.9}$$

as $n \to \infty$. This is due to the fact that under (i), the first convergence in (2.9) implies that for all indices j for which $|(\tilde{\mathbf{r}}_\lambda)_j| < \lambda$ (and therefore $(\tilde{\boldsymbol{\beta}}_\lambda)_j = 0$), we will also have $|(\hat{\mathbf{r}}_\lambda)_j| < \lambda$ (and therefore $(\hat{\boldsymbol{\beta}}_\lambda)_j = 0$). So we will have $(\hat{\boldsymbol{\beta}}_\lambda)_j = 0$ for all j such that $(\tilde{\boldsymbol{\beta}}_\lambda)_j = 0$. Under the beta-min condition (ii), the second convergence in (2.9) implies that for all j such that $(\tilde{\boldsymbol{\beta}}_\lambda)_j \neq 0$, we will have $(\hat{\boldsymbol{\beta}}_\lambda)_j \neq 0$. The convergences in (2.9) can be established under

mild smoothness and convexity conditions on the objective function defined by $q_n(\mathbf{s}; t_0, \mathbf{t}) \equiv (2n)^{-1}\|\mathbf{z}_0(t_0, \mathbf{t}) - \mathbf{U}_0(t_0, \mathbf{t})\mathbf{s}\|^2 + \lambda\|\mathbf{s}\|_1$, for which we can write $\tilde{\boldsymbol{\beta}}_\lambda = \arg\min_{\mathbf{s}} q_n(\mathbf{s}; \beta_0, \boldsymbol{\beta})$ and $\hat{\boldsymbol{\beta}}_\lambda = \arg\min_{\mathbf{s}} q_n(\mathbf{s}; \hat{\beta}_0, \hat{\boldsymbol{\beta}})$. Under such conditions, $\|(\hat{\beta}_0, \hat{\boldsymbol{\beta}}^\top)^\top - (\beta_0, \boldsymbol{\beta}^\top)^\top\| \xrightarrow{p} 0$ will imply the convergences in (2.9).

We next describe post-selection inference based on treating $\hat{\mathbf{z}}_0$ and $\hat{\mathbf{U}}_0$ as though they were equal to $\mathbf{z}_0$ and $\mathbf{U}_0$ and treating $\hat{M}_\lambda$ as though it matched $\tilde{M}_\lambda$.

2.4 Variable selection after linearization

Here we consider whether variable selection in the linear model with pseudo-data $\hat{\mathbf{z}}_0$ and $\hat{\mathbf{U}}_0$ will be perform similarly to variable selection based on L_1-penalization of the original GLM likelihood; that is, we consider how likely it is that selected models $\hat{M}_\lambda$ and $\hat{M}_\lambda^{\text{mle}}$ will match.

We find that if the conditions are met for model selection consistency by L_1-penalization of the GLM likelihood, then the conditions are also met for model selection consistency by L_1-penalization in the linear model with the idealized pseudo-data $\mathbf{z}_0$ and $\mathbf{U}_0$. In particular, Lee et al. (2015) present two main assumptions allowing for model selection consistency of a class of regularized M-estimators. These are a restricted strong convexity (RSC) condition and the so-called irrepresentable condition, the first version of which appeared in Zhao and Yu (2006). To express these conditions in our GLM setting with L_1-

penalization, denote the Hessian of the scaled negative log-likelihood appearing in (2.2) by $\mathcal{Q}(t_0, \mathbf{t}) \equiv (na(\phi))^{-1} \sum_{i=1}^{n} b''(\eta_i(t_0, \mathbf{t}))\tilde{\mathbf{x}}_i\tilde{\mathbf{x}}_i^\top$, where $\tilde{\mathbf{x}}_i \equiv [1\ \mathbf{x}_i^\top]^\top$, $i = 1, \ldots, n$, and set $\mathcal{Q}_\phi \equiv \mathcal{Q}(\beta_0, \boldsymbol{\beta})$. Now let $\mathcal{C}_0 \times \mathcal{C} \subset \mathbb{R} \times \mathbb{R}^p$ be a known convex set containing $(\beta_0, \boldsymbol{\beta})$ and set $\mathcal{M}_0 \equiv \text{span}\{\mathbf{e}_j, j \in M_0\}$, where $\{\mathbf{e}_j, j = 1, \ldots, p\}$ are elementary basis vectors in $\mathbb{R}^p$. Furthermore let $\tilde{\boldsymbol{\delta}}$ represent a vector $\tilde{\boldsymbol{\delta}} \equiv (\delta_0, \boldsymbol{\delta}^\top)^\top$ for $(\delta, \boldsymbol{\delta}) \in \mathbb{R} \times \mathbb{R}^p$. Then the RSC condition in our GLM setting becomes $\tilde{\boldsymbol{\delta}}^T \mathcal{Q}_\phi(t_0, \mathbf{t})\tilde{\boldsymbol{\delta}} \geq \kappa\|\tilde{\boldsymbol{\delta}}\|$ for all $(\delta_0, \boldsymbol{\delta}), (t_0, \mathbf{t}) \in \mathcal{C}_0 \times (\mathcal{C} \cap \mathcal{M}_0)$ for some $\kappa > 0$ (cf. Assumption 3.1 of Lee et al. (2015)). In addition, the irrepresentable condition becomes $\|\mathcal{Q}_{\phi, M_0^c, M_0} \mathcal{Q}_{\phi, M_0, M_0}^{-1} \text{sign}(\boldsymbol{\beta}_{M_0})\|_\infty \leq 1 - \xi$ for some $\xi \in (0, 1)$, where $\mathcal{Q}_{,\phi,\mathcal{A},\mathcal{B}}$ denotes the matrix constructed from $\mathcal{Q}_\phi$ by keeping rows with indices in $\mathcal{A}$ and columns with indices in $\mathcal{B}$.

Now, in the idealized linear model the Hessian of the loss function $\|\mathbf{z}_0 - (\mathbf{u}_0 t_0 + \mathbf{U}\mathbf{t})\|^2$ is exactly $\mathcal{Q}_\phi$. Therefore if the RSC condition holds, then the same condition holds when $\mathcal{Q}_\phi(t_0, \mathbf{t})$ is replaced by $\mathcal{Q}_\phi = \mathcal{Q}_\phi(\beta_0, \boldsymbol{\beta})$, since $(\beta_0, \boldsymbol{\beta})$ belongs to the set $\mathcal{C}_0 \times (\mathcal{C} \cap \mathcal{M}_0)$. Therefore, if the RSC is satisfied for the GLM, it will also be satisfied in the idealized linear model. Moreover, the irrepresentable condition for the GLM is identical to its counterpart in the linear model with the idealized pseudo-data, as it is formulated in terms of Hessian evaluated at the true parameter values.

Therefore, if these conditions are met for consistent variable selection via

L_1-penalization of the GLM log-likelihood, they will also be met for consistent variable selection via L_1-penalization of the least-squares criterion in the idealized pseudo-data. By the discussion at the end of Section 2.3, model selections based on $\hat{\mathbf{z}}_0$ and $\hat{\mathbf{U}}_0$ will reliably match those based on the idealized $\mathbf{z}_0$ and $\mathbf{U}_0$, so that, by extension, model selection via L_1-penalization of the least-squares criterion in $\hat{\mathbf{z}}_0$ and $\hat{\mathbf{U}}_0$, giving $\hat{M}_\lambda$, will be reliable whenever model selection via L_1-penalization of the GLM log-likelihood, giving $\hat{M}_\lambda^{\text{mle}}$, is reliable. In support of these findings, we present an empirical comparison of selected models $\hat{M}_\lambda$ and $\hat{M}_\lambda^{\text{mle}}$ on simulated data sets in the Supplementary Material.

3. Post-selection inference based on the linearized model

The (centered) idealized response vector $\mathbf{z}_0$ may be written as

$$\mathbf{z}_0 = \mathbf{U}_0\boldsymbol{\beta} + \boldsymbol{\xi}_0, \tag{3.1}$$

where the term $\boldsymbol{\xi}_0$ is defined as the centered version $\boldsymbol{\xi}_0 \equiv (\mathbf{I} - \mathbf{P}_0)\boldsymbol{\xi}$ of the vector $\boldsymbol{\xi} \equiv (\xi_1, \ldots, \xi_n)^\top$, where $\xi_i \equiv (Y_i - b'(\eta_i))/\sqrt{b''(\eta_i)}$, $i = 1, \ldots, n$ are independent random variables with mean zero and variance $a(\phi)$, so that the covariance matrix of $\boldsymbol{\xi}_0$ is $(\mathbf{I} - \mathbf{P}_0)a(\phi)$.

Here we apply the parametric programming (PP) approach in Le Duy and Takeuchi (2021) for making post-selection inferences based on observing the idealized data $\mathbf{z}_0$ and $\mathbf{U}_0$ in (3.1), treating this as a Gaussian linear model.

3.1 Post-selection sampling distribution of idealized LSE

Given a model $M \subset \{1, \ldots, p\}$ the estimator $\tilde{\boldsymbol{\beta}}_M$ has entry j given by $(\tilde{\boldsymbol{\beta}}_M)_j = \mathbf{c}_{M,j}^\top \mathbf{z}_0$, and its estimation target is $(\tilde{\boldsymbol{\beta}}^*_M)_j$ with $\tilde{\boldsymbol{\beta}}^*_M$ defined in (2.8). Recall that $\tilde{M}_\lambda$ is the model selected via L_1-penalization in (2.6) of the least-squares criterion in the idealized pseudo-data $\mathbf{z}_0$ and $\mathbf{U}_0$. Given the event $\tilde{M}_\lambda = M$, we consider making conditional inferences on $(\tilde{\boldsymbol{\beta}}^*_M)_j$ based on the conditional distribution of $(\tilde{\boldsymbol{\beta}}_M)_j | \{\tilde{M}_\lambda = M\}$.

This is reminiscent of the problem raised in Section 2.1 of finding the distribution of $(\hat{\boldsymbol{\beta}}^{\text{mle}}_M)_j | \{\hat{M}^{\text{mle}}_\lambda = M\}$ in order to make inferences on $(\boldsymbol{\beta}^*_M)_j$ in the GLM setting. In essence, we transform the original intractable problem in GLMs to an easier (and solved) problem in linear models. The analogy of these two problems is justified by the asymptotic equivalence of the pivotal quantities considered in Theorem 1 and, moreover, as we discuss in Section 2.4, the fact that selected models $\hat{M}^{\text{mle}}_\lambda$ and $\hat{M}_\lambda$ will be reliably similar under standard conditions.

Treating $\boldsymbol{\xi}_0$ as multivariate Gaussian, the estimator $(\tilde{\boldsymbol{\beta}}_M)_j$, prior to conditioning on the selection of model M, has the $\mathcal{N}((\tilde{\boldsymbol{\beta}}^*_M)_j, a(\phi)\|\mathbf{c}_{M,j}\|^2)$ distribution, where this approximately holds when $\boldsymbol{\xi}_0$ is non-Gaussian, by Theorem 1(i). To account for variable selection, inference should be based on the *conditional* distribution of $(\tilde{\boldsymbol{\beta}}_M)_j$ given the selection event $\{\tilde{M}_\lambda = M\} = \{\mathbf{z}_0 \in \mathbb{R}^n :$

$\tilde{M}_\lambda(\mathbf{z}_0) = M\}$, where we use $\tilde{M}_\lambda(\mathbf{z}_0)$ to indicate the dependence of $\tilde{M}_\lambda$ on $\mathbf{z}_0$

Given any contrast $\mathbf{c}^\top\mathbf{z}_0$ of interest, the main idea of the PP approach of Le Duy and Takeuchi (2021) is to "parameterize" all the relevant response vectors $\mathbf{z}_0$ in $\mathbb{R}^n$ by linking them to a "parameter" τ in $\mathbb{R}$ so that the selection event can be identified with a subset of parameter values in $\mathbb{R}$ rather than with a subset of response vectors in $\mathbb{R}^n$. To achieve this parameterization, we define another event $\{\mathbf{z}_0 \in \mathbb{R}^n : \hat{\mathbf{q}}(\mathbf{z}_0) = \mathbf{q}\}$, where $\hat{\mathbf{q}}(\mathbf{z}_0) \equiv (\mathbf{I} - \mathbf{P_c})\mathbf{z}_0$, where $\mathbf{P_c} \equiv \|\mathbf{c}\|^{-2}\mathbf{c}\mathbf{c}^\top$, and $\mathbf{q}$ is in the column space of $(\mathbf{I} - \mathbf{P_c})$ as a realization of $\hat{\mathbf{q}}(\mathbf{z}_0)$ in a given application where M is the realization of $\tilde{M}_\lambda(\mathbf{z}_0)$. By construction, $\hat{\mathbf{q}}(\mathbf{z}_0)$ and $\mathbf{c}^\top\mathbf{z}_0$ are uncorrelated, and since we are treating $\boldsymbol{\xi}_0$ as multivariate Gaussian, they are independent. Thus, using "$\overset{d}{=}$" to refer to "equivalent in distribution," we have

$$
\begin{aligned}
&\mathbf{c}^\top\mathbf{z}_0|\{\mathbf{z}_0 \in \mathbb{R}^n : \tilde{M}_\lambda(\mathbf{z}_0) = M\} \\
&\overset{d}{=} \mathbf{c}^\top\mathbf{z}_0|\{\mathbf{z}_0 \in \mathbb{R}^n : \tilde{M}_\lambda(\mathbf{z}_0) = M,\ \hat{\mathbf{q}}(\mathbf{z}_0) = \mathbf{q}\} \\
&\overset{d}{=} \mathbf{c}^\top\mathbf{z}_0|\{\mathbf{z}_0 \in \mathbb{R}^n : \tilde{M}_\lambda(\mathbf{z}_0) = M,\ (\mathbf{I} - \mathbf{P_c})\mathbf{z}_0 = \mathbf{q}\} \\
&\overset{d}{=} \mathbf{c}^\top\mathbf{z}_0|\{\mathbf{z}_0 \in \mathbb{R}^n : \tilde{M}_\lambda(\mathbf{z}_0) = M,\ \mathbf{z}_0 = \mathbf{q} + \mathbf{P_c}\mathbf{z}_0\} \\
&\overset{d}{=} \mathbf{c}^\top\mathbf{z}_0|\{\mathbf{z}_0 \in \mathbb{R}^n : \tilde{M}_\lambda(\mathbf{q} + \mathbf{P_c}\mathbf{z}_0) = M\} \\
&\overset{d}{=} \mathbf{c}^\top\mathbf{z}_0|\{\mathbf{z}_0 \in \mathbb{R}^n : \tilde{M}_\lambda(\mathbf{q} + \|\mathbf{c}\|^{-2}\mathbf{c}\mathbf{c}^\top\mathbf{z}_0) = M\} \\
&\overset{d}{=} \mathbf{c}^\top\mathbf{z}_0|\{\tau \in \mathbb{R} : \tilde{M}_\lambda(\mathbf{z}_0(\tau)) = M\}, \qquad (3.2)
\end{aligned}
$$

with $\mathbf{z}_0(\tau) \equiv \mathbf{q} + \tau\|\mathbf{c}\|^{-2}\mathbf{c}$, indexed by τ, so that it moves across the support of $\mathbf{z}_0$ as τ moves across $\mathbb{R}$. Similarly, we "parameterize" the models selected via the minimization in (2.6) be writing

$$\tilde{\boldsymbol{\beta}}_\lambda(\tau) \equiv \underset{\mathbf{t}\in\mathbb{R}^p}{\arg\min}\, \frac{1}{2n}\|\mathbf{z}_0(\tau) - \mathbf{U}_0\mathbf{t}\|^2 + \lambda\|\mathbf{t}\|_1, \tag{3.3}$$

so that $\tilde{M}_\lambda(\mathbf{z}_0(\tau)) \equiv \{j : (\tilde{\boldsymbol{\beta}}_\lambda(\tau))_j \neq 0\}$ as τ varies in $\mathbb{R}$. This translates the search of $\mathbf{z}_0$ in $\mathbb{R}^n$ for the selection event $\{\tilde{M}_\lambda(\mathbf{z}_0) = M\}$ to the search of τ in $\mathbb{R}$ that satisfies $\tilde{M}_\lambda(\mathbf{z}_0(\tau)) = M$ for a given model M. Le Duy and Takeuchi (2021) provided an efficient algorithm to identify the values of τ in $\mathcal{T}_M \equiv \{\tau \in \mathbb{R} : \tilde{M}_\lambda(\mathbf{z}_0(\tau)) = M\}$ as a union of disjoint intervals in $\mathbb{R}$.

With $\mathbf{c} = \mathbf{c}_{M,j}$ and $\mathcal{T}_{M,j}$ as the support of the conditional distribution of $(\tilde{\boldsymbol{\beta}}_M)_j = \mathbf{c}_{M,j}^\top\mathbf{z}_0$, by (3.2) we have

$$(\tilde{\boldsymbol{\beta}}_M)_j|\{\tilde{M}_\lambda(\mathbf{z}_0) = M\} \sim \mathcal{N}_{\mathcal{T}_{M,j}}((\tilde{\boldsymbol{\beta}}^*_M)_j,\, a(\phi)\|\mathbf{c}_{M,j}\|^2),$$

where $\mathcal{N}_{\mathcal{T}_{M,j}}(a, b)$ represents the normal distribution with mean a and variance b truncated to have support on $\mathcal{T}_{M,j}$. Denoting by $F_{\mathcal{T}_{M,j}}(\cdot; a, b)$ the cumulative distribution function (CDF) of this distribution, we can pass $(\tilde{\boldsymbol{\beta}}_M)_j$ through its own CDF to construct a pivotal quantity with which a post-selection $(1-\alpha)100\%$ confidence interval for $(\boldsymbol{\beta}^*_M)_j$ may be constructed as

$$\text{CI}_{M,j} \equiv \left\{\mu \in \mathbb{R} : \alpha/2 \leq F_{\mathcal{T}_{M,j}}((\tilde{\boldsymbol{\beta}}_M)_j; \mu, a(\phi)\|\mathbf{c}_{M,j}\|^2) \leq 1 - \alpha/2\right\} \tag{3.4}$$

for $\alpha \in [0, 1/2]$. Similarly, a post-selection p-value for testing $(\tilde{\boldsymbol{\beta}}^*_M)_j = 0$ versus $(\tilde{\boldsymbol{\beta}}^*_M)_j \neq 0$ for $j \in M$ based on $(\tilde{\boldsymbol{\beta}}_M)_j$ can be defined as

$$2 \times \min \Big\{ F_{\mathcal{T}_M}((\tilde{\boldsymbol{\beta}}_M)_j; 0, a(\phi)\|\mathbf{c}_{M,j}\|^2),\ 1 - F_{\mathcal{T}_M}((\tilde{\boldsymbol{\beta}}_M)_j; 0, a(\phi)\|\mathbf{c}_{M,j}\|^2) \Big\}. \tag{3.5}$$

We propose carrying out these steps, substituting for $\mathbf{z}_0$ and $\mathbf{U}_0$ the observable counterparts $\hat{\mathbf{z}}_0$ and $\hat{\mathbf{U}}_0$ after selecting the model $\hat{M}_\lambda$ based on the estimator $\hat{\boldsymbol{\beta}}_\lambda$ in (2.4).

3.2 Accounting for penalty parameter selection

Instead of being pre-specified as assumed in Section 3.1, we may allow the penalty parameter λ to be selected based on the data $\mathbf{z}_0$ and $\mathbf{U}_0$, as is usually done in practice. This change demands a revision of the selection event in (3.2) to acknowledge the additional selection of λ. This extra selection can again be "parameterized" via τ when λ is chosen based on one partition of $(\mathbf{z}_0, \mathbf{U}_0)$ into a training set, $(\mathbf{z}_0^{\text{train}}, \mathbf{U}_0^{\text{train}})$, and a validation set, $(\mathbf{z}_0^{\text{val}}, \mathbf{U}_0^{\text{val}})$.

More specifically, the selection of λ can be formulated as an optimization indexed by τ, $\tilde{\lambda}(\mathbf{z}_0(\tau)) \equiv \arg\min_{\lambda \in \Lambda} \|\mathbf{z}_0^{\text{val}}(\tau) - \mathbf{U}_0^{\text{val}} \tilde{\boldsymbol{\beta}}_\lambda^{\text{train}}(\mathbf{z}(\tau))\|^2$, where Λ is the set of candidate values for λ, and, for each $\lambda \in \Lambda$, mimicking (3.3), $\tilde{\boldsymbol{\beta}}_\lambda^{\text{train}}(\mathbf{z}_0(\tau)) \equiv \arg\min_{\mathbf{t} \in \mathbb{R}^p} (2n)^{-1} \|\mathbf{z}_0^{\text{train}}(\tau) - \mathbf{U}_0^{\text{train}} \mathbf{t}\|^2 + \lambda \|\mathbf{t}\|_1$. Define $\mathcal{T}_\lambda \equiv \{\tau \in \mathbb{R} : \tilde{\lambda}(\mathbf{z}_0(\tau)) = \lambda\}$, viewing λ as a realization of $\tilde{\lambda}(\mathbf{z}_0(\tau))$ correspond-

ing to M as the realization of $\tilde{M}_\lambda(\mathbf{z}_0(\tau))$. Then the complete selection event is $\mathcal{T}_\lambda \cap \mathcal{T}_{M,j} \equiv \{\tau \in \mathbb{R} : \tilde{\lambda}(\mathbf{z}_0(\tau)) = \lambda,\ \tilde{M}_\lambda(\mathbf{z}_0(\tau)) = M\}$, and thus $(\tilde{\boldsymbol{\beta}}_M)_j | \{\tilde{M}_\lambda = M\} \sim \mathcal{N}_{\mathcal{T}_\lambda \cap \mathcal{T}_{M,j}}((\tilde{\boldsymbol{\beta}}^*_M)_j,\ a(\phi)\|\mathbf{c}_{M,j}\|^2)$. Le Duy and Takeuchi (2021) developed an algorithm for identifying $\mathcal{T}_\lambda$. It is now straightforward to obtain a $(1-\alpha)100\%$ confidence interval for $(\tilde{\boldsymbol{\beta}}^*_M)_j$ and to compute the p-value for testing the significance of this covariate effect based on $(\tilde{\boldsymbol{\beta}}_M)_j$: one simply changes the support of the post-selection sampling distribution of $(\tilde{\boldsymbol{\beta}}_M)_j$ from $\mathcal{T}_{M,j}$ to $\mathcal{T}_\lambda \cap \mathcal{T}_{M,j}$ in the distribution function in (3.4) and (3.5).

4. Implementation of the proposed method

Here we describe the construction of $\hat{\mathbf{z}}_0$ and $\hat{\mathbf{U}}_0$ in three non-Gaussian models. The first two are GLMs, while the third, though not a GLM, admits a construction of $\hat{\mathbf{z}}_0$ and $\hat{\mathbf{U}}_0$ analogous to that in GLMs. We note that if the MLE is undefined due to, for example, a complete or quasi-complete separation in binary response data (Albert and Anderson, 1984), or if $p > n$, we prescribe replacing the MLE with a slightly regularized estimator, such as an L_1-penalized estimator with weak penalization. Theorem 1(ii) will hold as long as $\hat{\mathbf{z}}_0$ and $\hat{\mathbf{U}}_0$ are constructed with a consistent estimator of $(\beta_0, \boldsymbol{\beta})$.

4.1 Logistic and Poisson models

For a binary response, the logistic regression model is most widely used. Here $a(\phi) = 1$ and $b(\eta) = \log(1 + e^{\eta})$ for all $\eta \in \mathbb{R}$ in (2.1), yielding $b'(\eta) = e^{\eta}/(1+e^{\eta})$ and $b''(\eta) = b'(\eta)(1-b'(\eta))$ as the key quantities needed to evaluate the $\hat{\mathbf{z}}_0$ and $\hat{\mathbf{U}}_0$. For responses which are counts, the Poisson regression model is often used. Here $a(\phi) = 1$ and $b(\eta) = b'(\eta) = b''(\eta) = e^{\eta}$ for all $\eta \in \mathbb{R}$. If a GLM involves an unknown ϕ, we prescribe substituting for ϕ the MLE $\hat{\phi}$. This is also our strategy for dealing with a nuisance parameter irrelevant to the linear predictor η in the third regression model described next.

4.2 Beta regression

For continuous responses $Y_1, \ldots, Y_n \in [0, 1]$, such as rates or proportions, observed with $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^p$, Ferrari and Cribari-Neto (2004) considered beta regression under which $Y_i \sim \text{Beta}(\mu_i\phi, \nu_i\phi)$ with $\mu_i \equiv 1/(1 + e^{-\eta_i})$ and $\nu_i \equiv 1 - \mu_i$ and η_i as before, for $i = 1, \ldots, n$, where $\phi > 0$ is a precision parameter. In this model Y_i has mean μ_i and variance $\mu_i(1 - \mu_i)/(1 + \phi)$ for $i = 1, \ldots, n$. The NR update is not the same in this setting as in Section 2.2 due to the fact that the Hessian of the log-likelihood depends on the responses (see Appendix A in Ferrari and Cribari-Neto, 2004). As an alternative to NR, we adopt Fisher scoring (Davison, 2003, Section 4.4.1), whereby the Hessian is replaced by its

expected value. If ϕ is known, we can describe the Fisher scoring update by setting $\mu_i(t_0, \mathbf{t}) \equiv 1/(1 + e^{-\eta_i(t_0,\mathbf{t})})$ and $\nu_i(t_0, \mathbf{t}) \equiv 1 - \mu_i(t_0, \mathbf{t})$ with $\eta_i(t_0, \mathbf{t}) \equiv t_0 + \mathbf{x}_i^\top \mathbf{t}$ as well as $\tilde{Y}_i \equiv \log(Y_i/(1 - Y_i))$ and $\tilde{\mu}_i(t_0, \mathbf{t}) \equiv \psi(\mu_i(t_0, \mathbf{t})\phi) - \psi(\nu_i(t_0, \mathbf{t})\phi)$ for $i = 1, \ldots, n$, where $\psi(\cdot)$ is the digamma function. From here we define weights

$$w_i(t_0, \mathbf{t}) \equiv \sqrt{\phi}\mu_i(t_0, \mathbf{t})\nu_i(t_0, \mathbf{t})\sqrt{\psi'(\mu_i(t_0, \mathbf{t})\phi) + \psi'(\nu_i(t_0, \mathbf{t})\phi)}$$

for $i = 1, \ldots, n$ and the vector $\mathbf{z}(t_0, \mathbf{t})$ having entries

$$z_i(t_0, \mathbf{t}) = w_i(t_0, \mathbf{t})\eta_i(t_0, \mathbf{t}) + \mu_i(t_0, \mathbf{t})\nu_i(t_0, \mathbf{t})(\tilde{Y}_i - \tilde{\mu}(t_0, \mathbf{t}))/w_i(t_0, \mathbf{t}),$$

the vector $\mathbf{u}_0(t_0, \mathbf{t})$ having entries $w_i(t_0, \mathbf{t})$ and the matrix $\mathbf{U}(t_0, \mathbf{t})$ having rows $w_i(t_0, \mathbf{t})\mathbf{x}_i^\top$ for $i = 1, \ldots, n$. Then, for a fixed value of ϕ, an initial value $(t_0^{(0)}, \mathbf{t}^{(0)})$ can be updated as in (2.3).

To estimate unknown ϕ, one maximizes the log-likelihood evaluated at the current $(t_0^{(k)}, \mathbf{t}^{(k)})$ with respect to ϕ (the maximizer can be obtained in closed form); then one updates $(t_0^{(k)}, \mathbf{t}^{(k)})$ via (2.3) with ϕ fixed at its current estimate. Iterating until convergence yields the MLE $(\hat{\beta}_0, \hat{\boldsymbol{\beta}}, \hat{\phi})$ of $(\beta_0, \boldsymbol{\beta}, \phi)$. As before, we set $\hat{\mathbf{z}} \equiv \mathbf{z}(\hat{\beta}_0, \hat{\boldsymbol{\beta}})$, $\hat{\mathbf{u}}_0 \equiv \mathbf{u}_0(\hat{\beta}_0, \hat{\boldsymbol{\beta}})$, and $\hat{\mathbf{U}} \equiv \mathbf{U}(\hat{\beta}_0, \hat{\boldsymbol{\beta}})$ as the pseudo-data, and to disregard the intercept we construct $\hat{\mathbf{z}}_0$ and $\hat{\mathbf{U}}_0$ as in Section 2.2.

The idealized counterparts to $\hat{\mathbf{z}}$, $\hat{\mathbf{u}}_0$, and $\hat{\mathbf{U}}$ are defined, as before, as $\mathbf{z} \equiv \mathbf{z}(\beta_0, \boldsymbol{\beta})$, $\mathbf{u}_0 \equiv \mathbf{u}_0(\beta_0, \boldsymbol{\beta})$, and $\mathbf{U} \equiv \mathbf{U}(\beta_0, \boldsymbol{\beta})$, and the "centered versions"

are also obtained as before. Thus we may write $\mathbf{z}_0 = \mathbf{U}_0\boldsymbol{\beta} + \boldsymbol{\xi}_0$ as in (3.1), where $\boldsymbol{\xi}_0 = (\mathbf{I} - \mathbf{P}_0)\boldsymbol{\xi}$, except that in beta regression the vector $\boldsymbol{\xi}$ has entries $\xi_i \equiv w_i^{-1}\mu_i\nu_i(\tilde{Y}_i - \tilde{\mu}_i)$ with $w_i \equiv w_i(\beta_0, \boldsymbol{\beta})$ and $\tilde{\mu}_i \equiv \tilde{\mu}_i(\beta_0, \boldsymbol{\beta})$ for $i = 1, \ldots, n$. Since $\tilde{Y}_i$ has mean $\tilde{\mu}_i$ and variance $\psi'(\mu_i\phi) + \psi'(\nu_i\phi)$, ξ_i has mean zero and variance ϕ^{-1} for $i = 1, \ldots, n$. The covariance matrix of $\boldsymbol{\xi}_0$ is thus $(\mathbf{I} - \mathbf{P}_0)\phi^{-1}$ in beta regression. Therefore, in Section 3.1 $a(\phi)$ is replaced with $a(\phi) \equiv \phi^{-1}$, where for the value of ϕ we plug in the MLE.

5. Simulation study

Here compare on simulated data sets the performance of the proposed method of applying parametric programming after linearization, which we abbreviate as PPL, with i) standard Wald-type inference in the selected model without any conditioning on the selection event and with ii) polyhedral-based post-selection inference as in Lee et al. (2016) following out linearization step. Comparisons of PPL and data-splitting are provided in the Supplementary Material.

5.1 Parametric programming versus the naive method

Under the logistic, Poisson, and beta regression models in Section 4, we generate response data after drawing covariate vectors $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^p$ having independent $\mathcal{N}(0, 1)$ entries under $n = 500$ and $p = 20$. On each of 1000

simulated data sets we implement our PPL method as well as the naive method to construct confidence intervals for the coefficients in the selected model $\hat{M}_\lambda$. On each data set we perform model selection at each value in a grid of twenty values for λ, equally spaced on logarithmic scale. In each model we set $\beta_0 = -2$ and $M_0 = \{1, 2, 3\}$. For logistic regression we set $\boldsymbol{\beta}_{M_0} = (2, 2, 1)^\top$, for Poisson, $\boldsymbol{\beta}_{M_0} = (1, 1, -1)^\top$, and for beta, $\boldsymbol{\beta}_{M_0} = (1, -1/2, 1/2)^\top$ and consider a set of λ values in the intervals $[2, 12]$, $[8, 56]$, and $[2, 10]$, respectively, where these intervals were chosen to yield wide spreads of model sizes.

For each simulated data set, at each λ, we perform post-selection inference by constructing the confidence interval $\text{CI}_{M,j}$ in (3.4) at $\alpha = 0.05$ for each index j in the selected model M. We then record a realization of the Type I error rate as the proportion of unimportant covariates in M for which a nonzero regression coefficient is inferred, that is, we record the ratio

$$\text{Type I error} \equiv \frac{|\{j \in M : 0 \notin \text{CI}_{M,j} \text{ and } (\boldsymbol{\beta}_M)_j = 0\}|}{|\{j \in M : (\boldsymbol{\beta}_M)_j = 0\}|},$$

where $\text{CI}_{M,j}$ is defined in (3.4). If $|\{j \in M : \beta_j = 0\}| = 0$, we record a zero for the Type I error. For the naive method, we use Wald-type confidence intervals based on $\hat{\boldsymbol{\beta}}_M^{\text{mle}}$ that ignore selection events.

Figure 1 shows the average Type I error achieved by the PPL method across 1000 Monte Carlo replicates for all regression settings at each λ with the average Type I error from the naive method overlaid. The average sizes of the selected

models at each λ are also depicted (by the heights of the bars). As expected, the naive method leads to increasingly inflated Type I error rates as λ increases. In contrast, our PPL method provides reliable inference after model selection, maintaining a Type I error close to the nominal level of 0.05 across all λ values.

We also compare the average Type I error achieved by the PPL and naive methods under data-based selection of the tuning parameter λ. Here we select a value of λ from among the grid of twenty values as described in Section 3.2 using a single $70\%/30\%$ split of the data. For logistic regression, the average Type I error rates achieved by the PPL and naive methods under $\alpha = 0.05$ were 0.049 and 0.198, respectively; for Poisson regression these were 0.044 and 0.269, respectively; and for beta regression these were 0.048 and 0.349 respectively. Here also we see that the proposed method maintains an average Type I error close to 0.05 across all three regression settings, whereas the naive method leads to drastically inflated Type I errors. This also showcases the versatility of the PPL method when compared with the polyhedral method, which cannot accommodate data-driven penalty parameter selection. Even without involving the additional selection of λ, our method still outperforms the polyhedral method, as we show next.

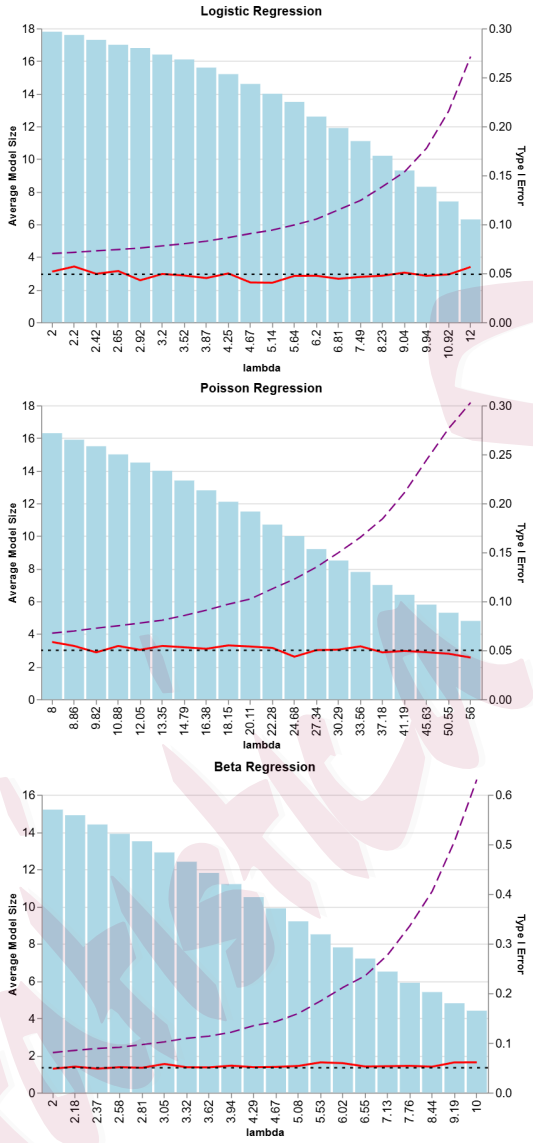

Figure 1: Average Type I error across 1000 Monte Carlo replicates over the grid of λ values for logistic, Poisson, and beta regression achieved by the PPL method (solid lines) and by the naive method (dashed lines). Heights of bars indicate average sizes of the selected models across the λ values.

5.2 Parametric programming versus polyhedral method

Our method of parametric programming after linearization (PPL) prescribes applying the PP approach to a linear model following a linearization step. Alternatively, following the linearization step, one could apply the polyhedral method described in Lee et al. (2016). The advantages of the PP method over the polyhedral method are twofold. First, the PP method avoids the "overconditioning" (on the signs of selected coefficients) that the polyhedral method involves. The minimal conditioning of the PP method enhances the statistical efficiency compared to that of the polyhedral method. Second, the PP method can easily account for data-driven selection of λ.

Focusing on logistic regression, we compare in Table 1 the performances of 95% confidence intervals based on these two strategies. As two valid post-selection inference methods, they both preserve coverage probabilities close to the nominal level of 95%. For the truly non-zero coefficient β_2, the average widths of the CIs obtained from the PP method tend to be wider than those obtained from the polyhedral method for most λ values. But, for β_4 and β_6, which are truly equal to zero, the CIs from the PP method are substantially tighter than those from the polyhedral method. This indicates that the over-conditioning (on signs) required by the polyhedral method is particularly detrimental to statistical

power when inferring a null covariate effect, as the signs of these regression coefficients in the estimated model can be positive or negative with non-vanishing probability even as the sample size grows. In contrast, coefficients that are far away from zero will, with high probability, have the correct signs in the selected model, so conditioning on their signs has less impact on statistical power.

Table 1: Average lower and upper bounds, average width, and empirical coverage probability of 95% confidence intervals from the PPL method and the polyhedral method at certain λ values.

Coefficient	PPL			Polyhedral		
	95% CI	Width	Coverage	95% CI	Width	Coverage
			$\lambda = 2$			
β_2	[0.50, 1.52]	1.02	93.0	[0.54, 1.61]	1.07	93.5
β_4	[−0.49, 0.54]	1.03	95.8	[−0.74, 1.33]	2.07	94.8
β_6	[−0.50, 0.52]	1.02	94.9	[−0.74, 1.35]	2.07	95.1
			$\lambda = 3.5$			
β_2	[1.36, 2.71]	1.35	95.8	[1.57, 2.70]	1.13	90.5
β_4	[−0.42, 0.44]	0.86	94.8	[−0.62, 0.61]	1.23	93.8
β_6	[−0.44, 0.43]	0.87	96.6	[−0.61, 0.57]	1.18	94.9
			$\lambda = 8.2$			
β_2	[1.13, 2.57]	1.44	96.6	[1.56, 2.53]	0.97	93.6
β_4	[−0.41, 0.46]	0.87	94.9	[−0.66, 0.59]	1.25	94.7
β_6	[−0.49, 0.42]	0.91	95.5	[−0.61, 0.59]	1.20	97.6
			$\lambda = 12$			
β_2	[1.20, 2.47]	1.27	96.5	[1.53, 2.42]	0.89	93.5
β_4	[−0.42, 0.47]	0.89	94.6	[−0.56, 0.57]	1.13	93.8
β_6	[−0.44, 0.43]	0.87	93.5	[−0.58, 0.62]	1.20	96.9

6. Illustrations on real data sets

6.1 Logistic regression for binary data

The Spambase data set from the UCI Machine Learning Repository contains descriptive information for $n = 4601$ emails, each classified as spam or not spam (Hopkins et al., 1999). There are $p = 57$ numeric features giving, for example, the frequencies of certain words or symbols, or the lengths of sequences of capital letters. We fit a logistic regression model to predict whether to classify an email as spam. After selecting λ using 70% of the data as training data and 30% as validation data, 25 variables were selected. Our PPL method for post-selection inference found 13 of the variables to be significant in classifying emails as spam or not spam, while naive inference found 19 of the variables to be significant at the 0.05 significance level.

Table 2 summarizes the results, showing confidence intervals from the PPL method and the naive method, along with p-values associated with the 19 variables deemed significant according to the latter. The proposed method produces CIs generally wider than those from the naive method, and with larger p-values than their naive counterparts. All selected features fall in the three types of features frequently reported as strong predictors in existing studies (Prerika et al., 2025): word frequency features, capital letter features, and special character

frequency features. Among the six selected features identified as statistically significant by the naive method but deemed insignificant by the PPL method, three (*word_freq_all*, *word_freq_will*, *word_freq_email*) are not among the top features selected by most existing studies.

Table 2: Inferences on selected covariate effects for the Spambase data set for the 19 covariates found to be significant according to the naive method. Names of covariates found insignificant by PPL are italicized.

	PPL			**Naive**		
Covariate	p-value	CI	Width	p-value	CI	Width
word_freq_all	0.32	$[-0.11, 0.20]$	0.31	0.02	$[0.02, 0.20]$	0.18
word_freq_our	0.00	$[0.22, 0.49]$	0.26	0.00	$[0.23, 0.40]$	0.17
word_freq_over	0.02	$[0.04, 0.32]$	0.28	0.00	$[0.11, 0.31]$	0.19
word_freq_remove	0.00	$[0.54, 0.96]$	0.42	0.00	$[0.84, 1.25]$	0.41
word_freq_internet	0.00	$[0.11, 0.38]$	0.27	0.00	$[0.21, 0.43]$	0.23
word_freq_will	0.18	$[-0.34, 0.09]$	0.43	0.00	$[-0.39, -0.15]$	0.24
word_freq_free	0.00	$[0.53, 0.95]$	0.42	0.00	$[0.40, 0.69]$	0.30
word_freq_business	0.01	$[0.08, 0.43]$	0.34	0.00	$[0.18, 0.41]$	0.24
word_freq_email	0.21	$[-0.06, 0.26]$	0.33	0.01	$[0.03, 0.23]$	0.20
word_freq_you	0.00	$[0.05, 0.32]$	0.27	0.00	$[0.31, 0.52]$	0.21
word_freq_your	0.14	$[-0.12, 0.42]$	0.54	0.00	$[0.27, 0.47]$	0.20
word_freq_000	0.00	$[0.36, 0.82]$	0.47	0.00	$[0.60, 1.07]$	0.47
word_freq_money	0.00	$[0.10, 0.38]$	0.28	0.00	$[0.24, 0.63]$	0.39
word_freq_re	0.05	$[-1.13, -0.00]$	1.02	0.00	$[-1.18, -0.62]$	0.55
word_freq_edu	0.80	$[-1.59, 3.68]$	5.27	0.00	$[-1.66, -0.72]$	0.95
char_freq_!	0.00	$[0.24, 0.53]$	0.29	0.00	$[0.54, 0.90]$	0.36
char_freq_$	0.00	$[0.56, 0.98]$	0.42	0.00	$[0.70, 1.03]$	0.33
capital_run_length_longest	0.05	$[0.00, 0.30]$	0.29	0.00	$[0.35, 0.66]$	0.31
capital_run_length_total	0.32	$[-0.20, 0.36]$	0.56	0.00	$[0.08, 0.29]$	0.21

6.2 Poisson regression for count data

Here, we analyze the Medicaid1986 data from the R package `AER` (Kleiber and Zeileis, 2008) using Poisson regression. These are Medicaid utilization data from the 1986 Medicaid Consumer Survey. Considering only those records under the Aid to Families with Dependent Children program and encoding categorical variables with indicators, the data set contains $n = 485$ observations and $p = 12$ covariates. The response is the number of doctor visits.

After selecting the penalty parameter λ with 70% of the data used for training data and 30% used as validation data, we select 4 covariates. Naive post-selection inference finds all four significantly associated with the number of doctor visits; however, our post-selection inference procedure finds only three to be significant. Table 3 summarizes the results, showing that the naive method claims stronger covariate effects after model selection than our PPL method. For example, the p-value associated with the covariate *school* from the PPL method is above 0.05, while the naive method produces a nearly zero p-value. Previous analyses on the data (Gurmu, 1997) confirmed that health status measures, such as *health1*, are among the most important predictors for the number of doctor visits; covariates such as *school* become less important once these measures are included in the model.

Table 3: Inferences on selected covariate effects for the Medicaid data set. Names of covariates found insignificant by PPL are italicized.

	PPL			Naive		
Covariate	p-value	CI	Width	p-value	CI	Width
children	0.01	$[-0.27, -0.06]$	0.21	0.00	$[-0.27, -0.11]$	0.16
health1	0.00	[0.36, 0.50]	0.14	0.00	[0.36, 0.47]	0.11
access	0.01	[0.06, 1.35]	1.29	0.00	[0.08, 0.22]	0.14
school	0.07	$[-0.02, 0.26]$	0.28	0.00	[0.10, 0.26]	0.16

6.3 Beta regression for proportion data

Lastly, we perform beta regression on the Student Performance data set from the UCI Machine Learning Repository (Hussain, 2018), which examines student achievement at two Portuguese secondary schools. Covariate information includes students' grades and demographic, social, and school-related attributes. After encoding categorical variables and removing outliers, the data set contains $n = 649$ observations and $p = 39$ covariates. The response variable is a student's final grade (ranging from 0 to 20) divided by 20.

Using beta regression, after selecting λ with 70% of the data used as training data and 30% as validation data, 27 covariates were included in the selected model. Our post-selection method identifies 10 variables as significant, while the naive method identifies 13 variables. Table 4 displays the inference results from the two methods. After adjusting for model selection, the PPL method is less aggressive in claiming the strength of covariate effects, with wider interval

estimates and larger p-values than those from the naive method. In particular, the PPL method concludes that the covariates, *goout* and *health*, are statistically insignificant, although they are significant at the 0.05 significance level according to the naive method. The findings of the PPL method are more in line with the consensus in the literature (Bhatia et al., 2025; Kesgin et al., 2025): among the non-grade features, the covariate *absences* and *failures* are top-ranked features, study time and academic support matter (e.g., *studytime*, *schoolsup_yes*, *higher_yes*) also tend to be highly influential on a student's final grade, parental education (*Medu*) is moderately influential, whereas some social and lifestyle variables, such as *goout* and *health*, have much weaker predictive power.

Table 4: Inferences on selected covariate effects for the student performance data set for the 13 covariates found to be significant according to the naive method. Names of covariates found insignificant by PPL are italicized.

	PPL			**Naive**		
Covariate	p-value	CI	Width	p-value	CI	Width
age	0.00	[0.04, 0.13]	0.09	0.00	[0.04, 0.13]	0.08
Medu	0.02	[0.02, 0.23]	0.21	0.01	[0.02, 0.11]	0.10
studytime	0.02	[0.01, 0.10]	0.09	0.01	[0.02, 0.10]	0.08
failures	0.00	[−0.19, −0.10]	0.09	0.00	[−0.19, −0.10]	0.08
goout	0.17	[−0.11, 0.02]	0.14	0.04	[−0.09, −0.00]	0.09
health	0.14	[−0.08, 0.02]	0.10	0.03	[−0.08, −0.00]	0.08
absences	0.00	[−0.14, −0.03]	0.10	0.00	[−0.12, −0.04]	0.08
school_MS	0.03	[−0.12, −0.00]	0.11	0.00	[−0.11, −0.02]	0.08
sex_M	0.02	[−0.11, −0.01]	0.10	0.00	[−0.11, −0.02]	0.09
Fjob_teacher	0.03	[0.01, 0.10]	0.09	0.01	[0.02, 0.10]	0.08
reason_reputation	0.13	[−0.02, 0.10]	0.12	0.02	[0.01, 0.09]	0.08
schoolsup_yes	0.00	[−0.12, −0.04]	0.08	0.00	[−0.12, −0.04]	0.08
higher_yes	0.00	[0.07, 0.14]	0.08	0.00	[0.07, 0.15]	0.08

7. Discussion

We propose parametric programming following a linearization step for performing post-selection inference in generalized linear models, which can be adapted to models outside of GLMs, such as the beta regression model. The proposed method addresses key limitations of the polyhedral method, which involves over-conditioning, leading to wider confidence intervals and compromised statistical power. The proposed method can also adjust for data-based penalty parameter selection, in contrast to the polyhedral method, which as-

sumes a fixed penalty parameter chosen prior to the observation of the data.

Compared to the work by Taylor and Tibshirani (2018), where extension of the polyhedral method in generalized regression models was considered, our work provides more insight into inferences and variable selection based on the pseudo-data in a linear model and those based on the original data in a GLM. The gained insight can potentially lead to further extension of the proposed methodology in several interesting directions. These include post-selection inference in nonparametric regression models, or when response data are partially observed and prone to error as in group testing settings, or when covariates are prone to measurement error. Our current approach is grounded in maximum likelihood estimation. Generalizing it to the broader M-estimation framework is another reachable goal that can broaden its applicability.

Computer code for implementing our proposed method and competing methods considered in the simulation study are available at `https://github.com/kateshen28/InfGLM`.

References

Albert, A. and J. A. Anderson (1984). On the existence of maximum likelihood estimates in logistic regression models. *Biometrika 71*(1), 1–10.

Bachoc, F., H. Leeb, and B. M. Pötscher (2019). Valid confidence intervals for post-model-selection pre-

dictors. *The Annals of Statistics 47*(3), 1475–1504.

Bachoc, F., D. Preinerstorfer, and L. Steinberger (2020). Uniformly valid confidence intervals post-model-selection. *The Annals of Statistics 48*(1), 440–463.

Berk, R., L. Brown, A. Buja, K. Zhang, L. Zhao, et al. (2013). Valid post-selection inference. *The Annals of Statistics 41*(2), 802–837.

Bhatia, R., S. Yadav, R. Sharma, Shubneet, A. R. Yadav, and N. S. Talwandi (2025). A comparative study of feature selection techniques for predicting student academic performance using educational data. In *International Conference on Data Analytics & Management*, pp. 352–362. Springer.

Davison, A. C. (2003). *Statistical models*, Volume 11. Cambridge university press.

Ferrari, S. and F. Cribari-Neto (2004). Beta regression for modelling rates and proportions. *Journal of applied statistics 31*(7), 799–815.

Gurmu, S. (1997). Semi-parametric estimation of hurdle regression models with an application to Medicaid utilization. *Journal of Applied Econometrics 12*(3), 225–242.

Hopkins, M., E. Reeber, G. Forman, and J. Suermondt (1999). Spambase. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C53G6X.

Huber, P. J. (2011). Robust statistics. In *International encyclopedia of statistical science*, pp. 1248–1251. Springer.

Hussain, S. (2018). Student Academics Performance. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C50W30.

Kesgin, K., S. Kiraz, S. Kosunalp, and B. Stoycheva (2025). Beyond performance: Explaining and ensuring fairness in student academic performance prediction with machine learning. *Applied Sciences 15*(15), 8409.

Kivaranovic, D. and H. Leeb (2021). On the length of post-model-selection confidence intervals conditional on polyhedral constraints. *Journal of the American Statistical Association 116*(534), 845–857.

Kleiber, C. and A. Zeileis (2008). *Applied Econometrics with R*. New York: Springer-Verlag.

Kuchibhotla, A. K., L. D. Brown, A. Buja, J. Cai, E. I. George, and L. H. Zhao (2020). Valid post-selection inference in model-free linear regression. *The Annals of Statistics 48*(5), 2953 – 2981.

Le Duy, V. N. and I. Takeuchi (2021). Parametric programming approach for more powerful and general lasso selective inference. *International conference on artificial intelligence and statistics*, 901–909.

Lee, J. D., D. L. Sun, Y. Sun, J. E. Taylor, et al. (2016). Exact post-selection inference, with application to the lasso. *Annals of Statistics 44*(3), 907–927.

Lee, J. D., Y. Sun, and J. E. Taylor (2015). On model selection consistency of regularized M-estimators. *Electronic Journal of Statistics 9*(1), 608 – 642.

McCullagh, P. and J. A. Nelder (1989). *Generalized Linear Models*, Volume 37. CRC Press.

Meinshausen, N., L. Meier, and P. Bühlmann (2009). P-values for high-dimensional regression. *Journal of the American Statistical Association 104*(488), 1671–1681.

Neufeld, A. C., L. L. Gao, and D. M. Witten (2022). Tree-values: selective inference for regression trees. *The Journal of Machine Learning Research 23*(1), 13759–13801.

Panigrahi, S. and J. Taylor (2023). Approximate selective inference via maximum likelihood. *Journal of the American Statistical Association 118*(544), 2810–2820.

Pirenne, S. and G. Claeskens (2024). Parametric programming-based approximate selective inference for adaptive lasso, adaptive elastic net and group lasso. *Journal of Statistical Computation and Simulation 94*(11), 2412–2435.

Prerika, Nitika, and K. Kumar (2025). Email spam detection using artificial neural network with hybrid feature selection. In L. Garg, N. Kesswani, and I. Brigui (Eds.), *AI Technologies for Information Systems and Management Science*, Cham, pp. 108–117. Springer Nature Switzerland.

Rasines, D. G. and G. A. Young (2023). Splitting strategies for post-selection inference. *Biometrika 110*(3), 597–614.

Rinaldo, A., L. Wasserman, M. G'Sell, et al. (2019). Bootstrapping and sample splitting for high-dimensional, assumption-lean inference. *Annals of Statistics 47*(6), 3438–3469.

Shen, Q., K. Gregory, and X. Huang (2024). Post-selection inference in regression models for group testing data. *Biometrics 80*(3), ujae101.

Taylor, J. and R. Tibshirani (2018). Post-selection inference for-penalized likelihood models. *Canadian Journal of Statistics 46*(1), 41–61.

Tibshirani, R. J., J. Taylor, R. Lockhart, and R. Tibshirani (2016). Exact post-selection inference for sequential regression procedures. *Journal of the American Statistical Association 111*(514), 600–620.

REFERENCES

Wainwright, M. J. (2009). Sharp thresholds for high-dimensional and noisy sparsity recovery using l1-constrained quadratic programming (lasso). *IEEE transactions on information theory 55*(5), 2183–2202.

Wasserman, L. and K. Roeder (2009). High dimensional variable selection. *Annals of statistics 37*(5A), 2178–2201.

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica: Journal of the Econometric Society 50*, 1–25.

Zhang, X. and G. Cheng (2017). Simultaneous inference for high-dimensional linear models. *Journal of the American Statistical Association 112*(518), 757–768.

Zhao, P. and B. Yu (2006). On model selection consistency of lasso. *The Journal of Machine Learning Research 7*, 2541–2563.

Zhao, Q., D. S. Small, and A. Ertefaie (2022). Selective inference for effect modification via the lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology 84*(2), 382–413.

University of South Carolina

E-mail: qshen@email.sc.edu

University of South Carolina

E-mail: gregorkb@stat.sc.edu

University of South Carolina

E-mail: huang@stat.sc.edu