arXiv:submit/4845389 [econ.EM] 16 Apr 2023

Penalized Likelihood Inference with Survey Data*

Joann Jasiak† Purevdorj Tuvaandorj‡

April 16, 2023

Abstract

This paper extends three Lasso inferential methods, Debiased Lasso, $C(\alpha)$ and Selective Inference to a survey environment. We establish the asymptotic validity of the inference procedures in generalized linear models with survey weights and/or heteroskedasticity. Moreover, we generalize the methods to inference on nonlinear parameter functions e.g. the average marginal effect in survey logit models. We illustrate the effectiveness of the approach in simulated data and Canadian Internet Use Survey 2020 data.

Keywords: Survey data, Survey weights, Lasso, Logit, Average Marginal Effect, Post-Selection Inference

1 Introduction

Survey data are widely used in many disciplines of social sciences. The statistical methodology for survey samples has been well-developed and culminated in a large body of literature (see e.g. Cameron and Trivedi, 2009; Wooldridge, 2010; Fuller, 2011; Thompson, 2012).

*The authors acknowledge financial support of the Catalyzing Interdisciplinary Research Clusters Initiative, York University. This research is part of the project "Digital Currencies", approved and funded by the office of VPRI and participating faculties of York University. The authors acknowledge access to the data provided by RDC and Statistics Canada (Project 21-MAPA YRK-721).

†York University, jasiakj@yorku.ca.

‡York University, tpujee@yorku.ca.

Despite the rapid development of machine learning/high-dimensional econometrics and the increasing availability of big datasets in recent years, the research on how to adapt/apply these high-dimensional statistical methods to survey data has been lagging. This paper aims to fill this gap in the literature by providing extensions of Lasso inference methods to survey environment. Our hope is to enrich the toolbox of practitioners who want to apply the high-dimensional regression methods to survey data.

Most prediction-oriented methods, including the Lasso, trade off bias and variance, and consequently, deliver a biased estimate that is not suitable for making inference on the model coefficients. Several post-Lasso-selection inference methods that mitigate this shortcoming have been proposed in the literature. Among others, Zhang and Zhang (2014) and Javanmard and Montanari (2014) propose a debiased Lasso (DB) method which is based on one-step iteration of the initial Lasso estimator. Belloni et al. (2016) propose double selection and $C(\alpha)$-type methods in a generalized linear model (GLM) that satisfies sparsity assumptions. The latter is based on an estimating equation orthogonalized against the nuisance parameter "score" function.

Lee et al. (2016) propose a selective inference (SI) method for the parameters in a linear model selected by the Lasso. The method is extended to a homoskedastic GLM by Taylor and Tibshirani (2018). In SI, the target parameters are determined from the data as opposed to being fixed before the the selection events. This feature makes the post-selection method conceptually different from the $C(\alpha)$ and DB methods, where the target parameters are the population parameters.

This paper presents two rather straightforward results. We first extend the $C(\alpha)$, DB and SI methods to a GLM estimated by the Lasso to accommodate survey weights and/or heteroskedasticity. The survey framework we adopt is similar to that of Wooldridge (2001). Accounting for survey weights naturally leads to conditional heteroskedasticity which, in turn, brings about an extra challenge because the active and inactive constraints of the Karush-Kuhn-Tucker condition for the Lasso problem are no longer asymptotically independent, and conditioning only on the active constraints as considered by Taylor and Tibshirani (2018) for a homoskedastic GLM may lead to invalid inference.

Second, we establish the asymptotic validity of the above three methods for inference

on nonlinear parameter functions such as the average marginal effects (AMEs) in a survey logit model.

There exist very few studies on the application of Lasso methods to survey data. McConville et al. (2017) consider a survey-weighted linear Lasso regression and develop a finite population asymptotic theory for Lasso estimators with a fixed number of regressors. In contrast, we consider a survey-weighted GLM and establish the asymptotic validity inference procedures under the usual infinite population framework, see e.g. Wooldridge (2001, 2010) and Cameron and Trivedi (2009) for the latter. Additionally, we allow for a growing number of covariates in the survey extensions of the debiased Lasso and $C(\alpha)$ methods.

The paper is organized as follows. Section 2 lays out the model framework. We propose extensions of the selective inference, debiased Lasso and $C(\alpha)$/orthogonalization methods in Section 3. Section 4 applies the proposed methods to inference on AMEs in a survey logit model. Section 5 provides a simulation evidence on the properties of the proposed methods and Section 6 presents an empirical application to Canadian Internet Use Survey 2020 data. We conclude in Section 7.

Notations and terminology Let $1(\cdot)$ denote the indicator function, and $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$ denote the smallest and the largest eigenvalue of a symmetric matrix A, respectively. For a $k \times 1$ vector $a = (a_1, \ldots, a_k)'$, we define $\|a\|_0 \equiv \text{supp}(a)$ (the number of nonzero components of the vector a) and $\|a\|_1 \equiv \sum_{i=1}^k |a_j|$. For a real matrix $A = (a_{ij})$, let $\|A\|_\infty \equiv \max_{i,j} |a_{ij}|$, and $\|A\| = \sqrt{\text{tr}(A'A)}$ and $\|A\|_2 = \sqrt{\lambda_{\max}(A'A)}$ denote its Frobenius and spectral norms, respectively. The sub-Gaussian norm of a random variable X is defined as

$$\|X\|_{\psi_2} \equiv \sup_{m \geq 1} m^{-1/2} (\mathrm{E}[|X|^m])^{1/m}. \tag{1.1}$$

A random variable X is called sub-Gaussian if $\|X\|_{\psi_2} \leq C < \infty$ for a constant $C > 0$. A random vector $X \in \mathbb{R}^p$ is called sub-Gaussian if the one-dimensional marginals $X'b$ are sub-gaussian random variables for all $b \in \mathbb{R}^p$. The sub-Gaussian norm for the random vector is defined as $\|X\|_{\psi_2} \equiv \sup_{\|b\|=1} \|X'b\|_{\psi_2}$. The sub-exponential norm of a random

variable $X \in \mathbb{R}$ is

$$\|X\|_{\psi_1} \equiv \inf\{t > 0 : \mathrm{E}[\exp(|X|/t)] \leq 2\}. \tag{1.2}$$

Moreover, let $\mathbf{1}_m = (1, \ldots, 1)'$ and $0_m = (0, \ldots, 0)'$ denote the $m \times 1$ vector of ones and zeros, respectively, and e_{jm} denote the $m \times 1$ unit vector whose j-th element is 1 and the remaining elements are 0.

Let $F(x; \mu, \sigma^2, a, b)$ denote the CDF of a $N(\mu, \sigma^2)$ random variable truncated on the interval $[a, b]$, that is,

$$F(x; \mu, \sigma^2, a, b) \equiv \frac{\Phi((x-\mu)/\sigma) - \Phi((a-\mu)/\sigma)}{\Phi((b-\mu)/\sigma) - \Phi((a-\mu)/\sigma)},$$

where $\Phi(\cdot)$ is the CDF of a $N(0, 1)$ random variable. Also, let $\Lambda(z) \equiv \exp(z)/(1 + \exp(z))$ denote the CDF of logistic distribution.

We abbreviate central limit theorem and continuous mapping theorem as CLT and CMT, respectively.

2 Model

We consider a GLM that specifies the conditional density of a scalar outcome variable y_i given a $(p+1) \times 1$ vector of covariates x_i which includes a constant as

$$f(y_i|x_i, \theta_0) = \exp(y_i x_i' \theta_0 - a(x_i' \theta_0)) c(y_i), \quad i = 1, \ldots, n,$$

where θ_0 is the true value of the parameter vector $\theta \in \mathbb{R}^{p+1}$, and $a(\cdot)$ and $c(\cdot)$ are known functions. To each vector of observations $(y_i, x_i')', i = 1, \ldots, n$, there corresponds a positive, bounded survey weight denoted as $w_i, i = 1, \ldots, n$.[1] Let $g(y, x'\theta) \equiv -\log f(y, x'\theta)$ and define the weighted log-likelihood function as follows:

$$L(\theta) \equiv -n^{-1} \sum_{i=1}^{n} w_i g(y_i, x_i' \theta). \tag{2.1}$$

[1] In our framework, $\{(y_i, x_i', w_i)'\}_{i=1}^n$ actually forms a triangular array $\{\{(y_{ni}, x_{ni}', w_{ni})'\}_{i=1}^n$. We drop the index n for notational simplicity.

As is well known, the weighted likelihood framework is commonly used in survey data analysis (Manski and Lerman, 1977; Cameron and Trivedi, 2009; Wooldridge, 2010), and accommodates, among others, the following stratification schemes.

Example 1 (Standard stratified sampling)**.** Let $\mathcal{Z}$ be the population for $z = (y, x')'$ which is assumed to be infinite (or contain a large number of units). $\mathcal{Z}$ is stratified into J, nonempty, mutually exclusive and exhaustive strata such that $\mathcal{Z} = \bigcup_{j=1}^{J} \mathcal{Z}_j$. n_j observations $\{z_{ij}\}_{i=1}^{n_j} = \{(y_{ij}, x'_{ij})'\}_{i=1}^{n_j}$ are sampled randomly from each stratum $\mathcal{Z}_j, j = 1, \ldots, J$. The strata sample sizes, n_js, are non-random, and the population frequencies $q_j = P[z \in \mathcal{Z}_j] > 0, j = 1, \ldots, J$, are assumed to be known. The weights on the observations from the j-th stratum are given by $w_{n_0+\cdots+n_{j-1}+1} = \cdots = w_{n_0+\cdots+n_{j-1}+n_j} = q_j/(n_j/n), j = 1, \ldots, J$, with $n_0 = 0$ and $n = \sum_{j=1}^{J} n_j$. Let us re-label the observations as $z_{ij} = z_{n_0+\cdots+n_{j-1}+i}, i = 1, \ldots, n_j, j = 1, \ldots, J$. The corresponding likelihood function is then

$$L(\theta) = -\sum_{j=1}^{J} q_j \left(n_j^{-1} \sum_{i=1}^{n_j} g(y_{ij}, x'_{ij}\theta) \right) = -n^{-1} \sum_{i=1}^{n} w_i g(y_i, x'_i\theta). \tag{2.2}$$

Example 2 (Exogenous stratification)**.** Let $\mathcal{Z} = \mathcal{Y} \times \mathcal{X}$, where $\mathcal{Y}$ and $\mathcal{X}$ are the sample spaces for y and x. The population is stratified into J strata according to a deterministic function of x_i: $\mathcal{X} = \cup_{j=1}^{J} \mathcal{X}_j$, where $\mathcal{X}_j, j = 1 \ldots, J$, are mutually exclusive. The population frequencies $q_j = P[z \in \mathcal{Z}_j] = P[x \in \mathcal{X}_j] > 0, j = 1, \ldots, J$, are assumed to be known. Given $n = \sum_{j=1}^{J} n_j$ observations $\{z_{ij}\}_{i=1,\ldots,n_j,\, j=1,\ldots,J} = \{(y_{ij}, x'_{ij})'\}_{i=1,\ldots,n_j,\, j=1,\ldots,J}$, where $\{z_{ij}\}_{i=1}^{n_j} = \{(y_{ij}, x'_{ij})'\}_{i=1}^{n_j}$ sampled randomly from each stratum $\mathcal{Z}_j, j = 1, \ldots, J$, the likelihood function can be formulated as in (2.2).

Wooldridge (2001) established the asymptotic properties of M-estimator under the above two sampling schemes. We use the same sampling schemes to establish the asymptotic validity of the Lasso-based inference methods described below.

The score function, the sample information and negative Hessian matrices correspond-

ing to (2.1) are defined as

$$S(\theta) \equiv \frac{\partial L(\theta)}{\partial \theta} = -n^{-1} \sum_{i=1}^{n} w_i x_i \dot{g}(y_i, x_i'\theta), \quad \dot{g}(y,t) \equiv \frac{\partial g(y,t)}{\partial t}, \tag{2.3}$$

$$\hat{I}(\theta) \equiv n^{-1} \sum_{i=1}^{n} w_i^2 x_i x_i' \dot{g}(y_i, x_i'\theta)^2, \tag{2.4}$$

$$\hat{H}(\theta) \equiv -\frac{\partial^2 L(\theta)}{\partial \theta \partial \theta'} = n^{-1} \sum_{i=1}^{n} w_i x_i x_i' \ddot{g}(y_i, x_i'\theta), \quad \ddot{g}(y,t) \equiv \frac{\partial^2 g(y,t)}{\partial t^2}. \tag{2.5}$$

Moreover, we define $H(\theta_0) \equiv \mathrm{E}[\hat{H}(\theta_0)]$ and $I(\theta_0) \equiv \mathrm{E}[\hat{I}(\theta_0)]$.

Let us partition $x_i = (1, \tilde{x}_i')' \in \mathbb{R}^{p+1}$ and $\theta = (\alpha, \beta')' \in \mathbb{R}^{p+1}$, where $\tilde{x}_i = (\tilde{x}_{i1}, \ldots, \tilde{x}_{ip})' \in \mathbb{R}^p$, $\alpha \in \mathbb{R}$ and $\beta \in \mathbb{R}^p$ so that $x_i'\theta = \alpha + \tilde{x}_i'\beta$.

In this paper, the variable selection, estimation and inference are performed using a survey-weighted Lasso where the negative of the weighted log-likelihood function (2.1) is minimized subject to ℓ_1 penalty on the slope parameters:

$$\min_{\theta=(\alpha,\beta')' \in \mathbb{R}^{p+1}} \left(-L(\theta) + \lambda \|\beta\|_1\right), \tag{2.6}$$

where $\lambda \geq 0$ is a tuning parameter. Note here that, as it is standard in the Lasso literature, only the "slope" parameters in $\beta = (\beta_1, \ldots, \beta_p)'$ are penalized. The j-th elements of θ and θ_0 are denoted as $\theta_{(j)}$ and $\theta_{0(j)}$, respectively.

Hereafter, $M \subseteq \{1, \ldots, p+1\}$ denotes the subset of regressors that includes the constant term and non-constant regressors with a vector of (non-zero) Lasso estimates $\hat{\beta}_M \in \mathbb{R}^{|M|-1}$ and $\hat{\mathtt{s}}_M \equiv \mathrm{sign}(\hat{\beta}_M) \in \{-1, 1\}^{|M|-1}$. Also, let $\beta_M \in \mathbb{R}^{|M|-1}$ be the subvector of β corresponding to M, $\theta_M = (\alpha, \beta_M')'$ and $\hat{\theta}_M = (\hat{\alpha}, \hat{\beta}_M')'$.

The Lasso solution in (2.6), with λ fixed, returns a random subset of regressors $\hat{M} \subset \{1, \ldots, p+1\}$. Since the intercept α is not penalized, $\hat{M}$ always includes the constant term. The target parameter vector in the selective inference considered in Section 3.1 is $\theta_{M0} = (\alpha_0, \beta_{M0}')'$, the true value of θ_M in the selected model $\hat{M} = M$. In contrast, the DB and $C(\alpha)$ in Sections 3.2–3.3 target the entire vector θ_0.

Let $\mathtt{s}_M \equiv \mathrm{sign}(\beta_{M0}) \in \{-1, 1\}^{|M|-1}$, and denote by $m_0 \equiv \|\theta_0\|_0$ the number of nonzero elements of θ_0. Moreover, let $\theta_{-M} \in \mathbb{R}^{p+1-|M|}$ be the subvector of parameters other than θ_M and $L_M(\theta_M)$ be the weighted log-likelihood function for the selected model with parameters

θ_M. It is clear that $L_M(\theta_M)$ can be obtained by evaluating $L(\theta)$ at θ^* whose non-zero elements are θ_M and remaining $p+1-|M|$ elements are 0. We partition (2.3)-(2.5) as follows:

$$
\begin{aligned}
S(\theta) &= [S_M(\theta)', S_{-M}(\theta)']', \ S_M(\theta) \in \mathbb{R}^{|M|}, \ S_{-M}(\theta) \in \mathbb{R}^{p+1-|M|}, \\
\hat{H}(\theta) &= \begin{bmatrix} \hat{H}_M(\theta_M) & \hat{H}_{M(-M)}(\theta_M) \\ \hat{H}_{-MM}(\theta_M) & \hat{H}_{-M}(\theta_M) \end{bmatrix}, \\
\hat{I}(\theta) &= \begin{bmatrix} \hat{I}_M(\theta_M) & \hat{I}_{M(-M)}(\theta_M) \\ \hat{I}_{-MM}(\theta_M) & \hat{I}_{-M}(\theta_M) \end{bmatrix},
\end{aligned}
$$

where $\hat{H}_M(\theta_M) \in \mathbb{R}^{|M|\times|M|}$ and $S_M(\theta_M) \in \mathbb{R}^{|M|}$ denote the negative Hessian matrix and score functions corresponding to θ_M, respectively.

With the partitioning above, Lee et al. (2016) and Taylor and Tibshirani (2018) show that the event $\{\hat{M} = M, \hat{\mathtt{s}}_{\hat{M}} = \mathtt{s}_M\}$ holds if and only if there exist random vectors $\hat{\theta}_M \in \mathbb{R}^{|M|}$ and $\mathtt{u} \in \mathbb{R}^{p+1-|M|}$ in the Karush-Kuhn-Tucker condition for the problem (2.6) such that

$$
\frac{\partial L_M(\hat{\theta}_M)}{\partial \theta_M} - (0, \lambda\, \mathtt{s}_M')' = S_M(\hat{\theta}_M) - (0, \lambda\, \mathtt{s}_M')' = 0, \quad \mathtt{s}_M = \mathrm{sign}(\hat{\beta}_M) \in \{-1, 1\}^{|M|-1}, \tag{2.7}
$$

$$
\frac{\partial L_M(\hat{\theta}_M)}{\partial \theta_{-M}} - \lambda \mathtt{u} = S_{-M}(\hat{\theta}_M) - \lambda \mathtt{u} = 0, \quad \mathtt{u} \in \mathbb{R}^{p+1-|M|}, \quad \|\mathtt{u}\|_\infty < 1. \tag{2.8}
$$

3 Post-Lasso selection inference

We establish the asymptotic validity of the three inference methods under the following assumptions imposed directly on the loss function $g(y,t)$ which are similar to the assumptions employed in van de Geer et al. (2014) and Xia et al. (2021).

Assumption 1 (Asymptotic validity)**.**

(a) $\{(y_i, x_i')'\}_{i=1}^n$ *are independent with* $\max_{1\le i\le n} a_i < C_u < \infty$ *a.s. where*

$$
a_i \in \{\|x_i\|_{\psi_2}, \|x_i\|_\infty, \|X\theta_0\|_\infty\}.
$$

Moreover, w_i is non-random with $0 < C_l < w_i < C_u$ for all n, i.

(b) *For $A \in \{H(\theta_0), I(\theta_0), \mathrm{E}[n^{-1}X'X]\}$, there exist positive constants λ_l and λ_u such that $0 < \lambda_l \leq \lambda_{\min}(A) \leq \lambda_{\max}(A) \leq \lambda_u < \infty$.*

(c) *The function $g(y,t) \equiv a(t) - yt - \log c(y)$ is convex in $t \in \mathbb{R}$ for all y, and twice differentiable with $\dot{g}(y,t) \equiv \partial g(y,t)/\partial t$ and $\ddot{g}(y,t) \equiv \partial^2 g(y,t)/\partial t^2$ for all (y,t). There exist a positive definite matrix H and $\eta > 0$ such that $\lambda_{\min}(H) > \lambda_l > 0$ and*

$$n^{-1}\sum_{i=1}^{n} \mathrm{E}[w_i(g(y_i, x_i'\theta) - g(y_i, x_i'\theta_0))] \geq \|H^{1/2}(\theta - \theta_0)\|^2 \tag{3.1}$$

for all $\|X(\theta - \theta_0)\|_\infty < \eta$. Furthermore, $\ddot{g}(y,t)$ is Lipschitz with some constant $L_0 > 0$:

$$\max_{t_0 \in \{x_i'\theta_0\}} \sup_{\max(|t-t_0|, |\tilde{t}-t_0|) \leq \eta} \sup_{y \in \mathcal{Y}} \frac{|\ddot{g}(y,t) - \ddot{g}(y,\tilde{t})|}{|t - \tilde{t}|} \leq L_0, \tag{3.2}$$

and

$$\max_{t_0 \in \{x_i'\theta_0\}} \sup_{y \in \mathcal{Y}} |\dot{g}(y, t_0)| \leq C_u, \tag{3.3}$$

$$\max_{t_0 \in \{x_i'\theta_0\}} \sup_{|t-t_0| \leq \eta} \sup_{y \in \mathcal{Y}} |\ddot{g}(y,t)| \leq C_u. \tag{3.4}$$

The boundedness of the variables stated in Assumption 1(a) is employed frequently in the literature, see Negahban et al. (2012), van de Geer et al. (2014) and Xia et al. (2021). To deal with survey samples, we relax the i.i.d. assumption used in these papers, and although the proofs of validity of the inference procedures considered below are quite standard, much of the effort of the proof goes into verifying that the same results that hold in an i.i.d. setup carries over to independent non-identically distributed (i.n.i.d.) samples.

The weight w_i is deterministic but we require that it is bounded from above and below away from 0, so it rules out strata that become asymptotically degenerate. Moreover, the weights do not need to sum to 1. In the `R` package `glmnet`, the weights are rescaled to sum to n.

Assumption 1(b) is a mild condition that ensures nonsingularity of the Hessian and information matrices in the case of slowly diverging number of covariates considered below.

Assumption 1(c) is standard and requires the convexity and boundedness of the first two derivatives and Lipschitz continuity of the second derivative of $g(y,t)$ with respect to t uniformly in a neighborhood of $x_i'\theta_0$ (see van de Geer et al. (2014) and Xia et al. (2021)).

The condition (3.1) is essentially the the Quadratic Margin Condition needed for the consistency of the Lasso (Bühlmann and van de Geer, 2011) and see also Negahban et al. (2012) for a related (stochastic) Restricted Strong Convexity condition. A sufficient condition for (3.1) is that $\ddot{g}(y, x'\theta)$ is bounded away from zero locally around $x_i'\theta_0$ for all $i = 1, \ldots, n$.

3.1 Selective inference

In this section, we extend the selective inference argument of Taylor and Tibshirani (2018) for a homoskedastic GLM to a GLM with survey weights and/or heteroskedasticity. In the SI, the target parameters are the coefficients selected by the Lasso. As a result, they are random before the selection, but not so conditional on the Lassso selection events. This feature distinguishes the selective inference method from the $C(\alpha)$ and debiased Lasso inference where the target parameters are the population parameters. See Lee et al. (2016) for further discussions about the difference between the SI and other inference methods.

As in Taylor and Tibshirani (2018), we fix $\lambda > 0$ and consider the following one-step estimator

$$\tilde{\theta}_M \equiv \hat{\theta}_M + \hat{H}_M(\hat{\theta}_M)^{-1} S_M(\hat{\theta}_M), \tag{3.5}$$

where $S_M(\hat{\theta}_M) = (0, \lambda\, \mathtt{s}_M')'$, from which we obtain the one-step estimator of β_{M0}:

$$\tilde{\beta}_M = \hat{\beta}_M + [0_{|M|-1}, I_{|M|-1}]\hat{H}_M(\hat{\theta}_M)^{-1} S_M(\hat{\theta}_M). \tag{3.6}$$

The SI is based on the asymptotic distribution of $\tilde{\beta}_M$ conditional on the selection event $\hat{M} = M$ and $\hat{\mathtt{s}}_{\hat{M}} = \mathtt{s}_M$.[2] From (2.7) and (3.6), it follows that

$$\mathtt{s}_M = \text{sign}\left(\tilde{\beta}_M - [0_{|M|-1}, I_{|M|-1}]\hat{H}_M(\hat{\theta}_M)^{-1}(0, \lambda\, \mathtt{s}_M')'\right),$$

[2] Lee et al. (2016) also propose a test statistic which is conditional on $\hat{M} = M$ only by taking the union of the events characterized by polyhedral constraints over all possible combinations of the signs of the selected coefficients.

hence

$$\operatorname{diag}(\mathtt{s}_M)\left(\tilde{\beta}_M - [0_{|M|-1}, I_{|M|-1}]\hat{H}_M(\hat{\theta}_M)^{-1}(0, \lambda\, \mathtt{s}_M')'\right) \geq 0. \tag{3.7}$$

As argued by Taylor and Tibshirani (2018) (see Equation (21) therein), in a homoskedastic GLM, the random quantities appearing in the active and inactive constraints (2.7) and (2.8) are asymptotically independent (after suitable normalizations). However, this no longer holds in our setup because the covariance matrix of the limiting Gaussian random variables is not block-diagonal in the presence of survey weights and heteroskedasticity. This entails conditioning not only on the active constraints, but also on the inactive constraints. In light of this, we next derive an affine constraint corresponding to (2.8). Let

$$\tilde{S}_{-M}(\hat{\theta}_M) \equiv S_{-M}(\hat{\theta}_M) - \hat{H}_{-MM}(\hat{\theta}_M)\hat{H}_M(\hat{\theta}_M)^{-1}S_M(\hat{\theta}_M). \tag{3.8}$$

By the fact that $\|\mathtt{u}\|_\infty < 1$, and after some algebra, we can express the inactive constraints in (2.8) as follows:

$$\tilde{S}_{-M}(\hat{\theta}_M) \leq \lambda(\mathbf{1}_{p+1-|M|} - \hat{H}_{-MM}(\hat{\theta}_M)\hat{H}_M(\hat{\theta}_M)^{-1}(0, \mathtt{s}_M')'), \tag{3.9}$$

$$-\tilde{S}_{-M}(\hat{\theta}_M) \leq \lambda(\mathbf{1}_{p+1-|M|} + \hat{H}_{-MM}(\hat{\theta}_M)\hat{H}_M(\hat{\theta}_M)^{-1}(0, \mathtt{s}_M)'), \tag{3.10}$$

where the inequalities hold element-wise. The Lasso selection events in (3.7), (3.9) and (3.10) can be rewritten in a compact form as

$$\{AZ \leq b\}, \tag{3.11}$$

where

$$A \equiv \begin{bmatrix} -\operatorname{diag}(\mathtt{s}_M) & 0_{(|M|-1)\times(p+1-|M|)} \\ 0_{(p+1-|M|)\times(|M|-1)} & I_{p+1-|M|} \\ 0_{(p+1-|M|)\times(|M|-1)} & -I_{p+1-|M|} \end{bmatrix} \in \mathbb{R}^{(2p+1-|M|)\times p}, \quad Z \equiv n^{1/2}\begin{bmatrix} \tilde{\beta}_M \\ \tilde{S}_{-M}(\hat{\theta}_M) \end{bmatrix} \in \mathbb{R}^p,$$

$$b \equiv n^{1/2}\begin{bmatrix} -\operatorname{diag}(\mathtt{s}_M)[0_{|M|-1}, I_{|M|-1}]\hat{H}_M(\hat{\theta}_M)^{-1}(0, \lambda\, \mathtt{s}_M')' \\ \lambda(\mathbf{1}_{p+1-|M|} - \hat{H}_{-MM}(\hat{\theta}_M)\hat{H}_M(\hat{\theta}_M)^{-1}(0, \mathtt{s}_M')') \\ \lambda(\mathbf{1}_{p+1-|M|} + \hat{H}_{-MM}(\hat{\theta}_M)\hat{H}_M(\hat{\theta}_M)^{-1}(0, \mathtt{s}_M')') \end{bmatrix} \in \mathbb{R}^{2p+1-|M|}. \tag{3.12}$$

Assuming that the number of non-constant regressors, p, is fixed, one can establish the asymptotic normality of the one-step estimators before the Lasso selection (see Section A.1):

$$\begin{bmatrix} n^{1/2}(\tilde{\beta}_M - \beta_{M0}) \\ n^{1/2}\tilde{S}_{-M}(\hat{\theta}_M) \end{bmatrix} \xrightarrow{d} N(0, \Sigma), \tag{3.13}$$

where Σ is a $p \times p$ asymptotic covariance matrix. The estimator of Σ is

$$\hat{\Sigma} = \begin{bmatrix} \hat{\Sigma}_{\beta\beta} & \hat{\Sigma}_{\beta s} \\ \hat{\Sigma}'_{\beta s} & \hat{\Sigma}_{ss} \end{bmatrix}, \tag{3.14}$$

where

$$\begin{aligned}
\hat{\Sigma}_{\beta\beta} &\equiv [0_{|M|-1}, I_{|M|-1}]\hat{H}_M(\hat{\theta}_M)^{-1}\hat{I}_M(\hat{\theta}_M)\hat{H}_M(\hat{\theta}_M)^{-1}[0_{|M|-1}, I_{|M|-1}]', \\
\hat{\Sigma}_{\beta s} &\equiv [0_{|M|-1}, I_{|M|-1}]\left[\hat{H}_M(\hat{\theta}_M)^{-1}\hat{I}_{M(-M)}(\hat{\theta}_M) - \hat{H}_M(\hat{\theta}_M)^{-1}\hat{I}_M(\hat{\theta}_M)\hat{H}_M(\hat{\theta}_M)^{-1}\hat{H}_{M(-M)}(\hat{\theta}_M)\right], \\
\hat{\Sigma}_{ss} &\equiv [I_{p+1-|M|}, -\hat{H}_{-MM}(\hat{\theta}_M)\hat{H}_M(\hat{\theta}_M)^{-1}]\begin{bmatrix} \hat{I}_{-M}(\hat{\theta}_M) & \hat{I}_{-MM}(\hat{\theta}_M) \\ \hat{I}_{M(-M)}(\hat{\theta}_M) & \hat{I}_M(\hat{\theta}_M) \end{bmatrix} \\
&\quad [I_{p+1-|M|}, -\hat{H}_{-MM}(\hat{\theta}_M)\hat{H}_M(\hat{\theta}_M)^{-1}]'.
\end{aligned} \tag{3.15}$$

From (3.13), we have the distributional approximation for Z

$$Z \overset{a}{\sim} N(\mu, \Sigma), \quad \mu \equiv n^{1/2}[\beta'_{M0}, 0'_{p+1-|M|}]'. \tag{3.16}$$

The latter combined with the affine constraints $\{AZ \leq b\}$ in (3.11), is now amenable to application of Lemma 3.1 below, which summarizes two key results of Lee et al. (2016) (Lemma 5.1 and Theorem 5.2). To describe the lemma, we define the following quantities

for a general $k \times 1$ random vector Z, and $A \in \mathbb{R}^{k\times k}$, $b \in \mathbb{R}^k$ and $\eta \in \mathbb{R}^k$:

$$c = c(\Sigma, \eta) \equiv \Sigma\eta(\eta'\Sigma\eta)^{-1}, \quad r = r(Z, \Sigma, \eta) \equiv (I_k - c\eta')Z, \tag{3.17}$$

$$\mathcal{V}^-(r) \equiv \max_{j:(Ac)_j<0} \frac{b_j - (Ar)_j}{(Ac)_j}, \tag{3.18}$$

$$\mathcal{V}^+(r) \equiv \min_{j:(Ac)_j>0} \frac{b_j - (Ar)_j}{(Ac)_j}, \tag{3.19}$$

$$\mathcal{V}^0(r) \equiv \min_{j:(Ac)_j=0} b_j - (Ar)_j, \tag{3.20}$$

where $(Ac)_j$ denotes the j-th element of Ac. Lee et al. (2016) show the following the result.

Lemma 3.1 (Polyhedral lemma and truncated Gaussian pivot (Lee et al., 2016))**.** *Let $Z \sim N(\mu, \Sigma)$ and $A \in \mathbb{R}^{k\times k}$, $b \in \mathbb{R}^k$ and $\eta \in \mathbb{R}^k$ be fixed quantities. If $c, r, \mathcal{V}^-(r), \mathcal{V}^+(r)$ and $\mathcal{V}^0(r)$ are defined as in (3.17)-(3.20), then*

(a) $\eta' z$ is independent of $\mathcal{V}^-(r)$, $\mathcal{V}^+(r)$ and $\mathcal{V}^0(r)$, and the following events are equivalent:

$$\{AZ \le b\} = \{\mathcal{V}^-(r) \le \eta' Z \le \mathcal{V}^+(r), \mathcal{V}^0(r) \ge 0\}. \tag{3.21}$$

(b) Furthermore,

$$F(\eta' Z; \eta'\mu, \eta'\Sigma\eta, \mathcal{V}^-(r), \mathcal{V}^+(r)) | \{AZ \le b\} \sim U(0, 1). \tag{3.22}$$

In our setup, b defined in (3.12) is random whereas Lemma 3.1 assumes constant b. In addition, we have an approximate normality in (3.16) instead of the exact normality assumed in Lemma 3.1. These lead to an asymptotic version of (3.22), namely, as $n \to \infty$

$$F(e'_{jp}Z, e'_{jp}\mu, e'_{jp}\hat{\Sigma}e_{jp}, \mathcal{V}^-(r), \mathcal{V}^+(r)) | \{AZ \le b\} \xrightarrow{d} U(0, 1). \tag{3.23}$$

(3.23) can be established using the results of Markovic et al. (2017). Although we do not directly use (3.23), it provides the basis of the inference procedures described below.

Suppose we wish to make inference on the j-th element of β_{M0}, $j = 1, \ldots, |M| - 1$ (conditional on the Lasso selection event $\hat{M} = M$ and $\hat{\mathtt{s}}_{\hat{M}} = \mathtt{s}_M$). Let Z, A and b be as in (3.12), μ be as in (3.16), and set $\eta = e_{jp} \in \mathbb{R}^p$, $c = c(Z, \eta)$ and $r = r(Z, \hat{\Sigma}, e_{jp})$ in (3.17),

where $\hat{\Sigma}$ is defined in (3.14). Fix $\zeta \in (0,1)$. The SI confidence interval (CI) of level $1-\zeta$ is of the form $\mathrm{CI}_{\hat{M}j} \equiv [\tilde{q}_l, \tilde{q}_u]$, where $\tilde{q}_l$ and $\tilde{q}_u$ are the solutions to the following equations

$$F(n^{1/2}q, e'_{jp}Z, e'_{jp}\hat{\Sigma}e_{jp}, \mathcal{V}^-(r), \mathcal{V}^+(r)) = \frac{\zeta}{2}, \tag{3.24}$$

$$F(n^{1/2}q, e'_{jp}Z, e'_{jp}\hat{\Sigma}e_{jp}, \mathcal{V}^-(r), \mathcal{V}^+(r)) = 1 - \frac{\zeta}{2}. \tag{3.25}$$

The asymptotic validity of the above CI is established in the following proposition.

Proposition 3.2. *Let Assumption 1 hold with p fixed, $\lambda = Cn^{-1/2}$, where $C = O(1)$, and $H(\theta_0)$ and $I(\theta_0)$ converge to nonsingular matrices. Then, it holds that for $\zeta \in (0,1)$*

$$\liminf_{n\to\infty} \; P[e'_{j\hat{M}}\beta_{\hat{M}0} \in \mathrm{CI}_{\hat{M}j} | \hat{M} = M, \hat{\mathtt{s}}_M = \mathtt{s}_M] = 1 - \zeta.$$

See Appendix A.1 for a proof. The assumption of fixed p is commonly used in the literature on SI (see e.g. Lee et al., 2016; Tian and Taylor, 2017; Taylor and Tibshirani, 2018; Kuchibhotla et al., 2022). Taylor and Tibshirani (2018) provide a heuristic argument for the validity of the selective inference in a homoskedastic GLM. Proposition 3.2 extends their argument to i.n.i.d. and possibly heteroskedastic survey samples. The asymptotic validity of the SI procedures typically entails showing that CLTs that hold before selection extend to selective inference under suitable assumptions (Tian and Taylor, 2017; Kuchibhotla et al., 2022). We establish the asymptotic validity of the selective inference procedure by verifying the conditions given in Kuchibhotla et al. (2022).

Inference on a nonlinear parameter function. Next we consider inference on a scalar nonlinear parameter function $\rho_M(\theta_{M0})$ (which may depend on n) in the selected model with coefficients β_M on the active variables. Such results are especially useful in the context of logit and probit models because the AMEs are often the objects of interest therein. Analogously to (3.6), consider the one-step estimator

$$\tilde{\rho}_M \equiv \rho(\hat{\theta}_M) + \dot{\rho}_M(\hat{\theta}_M)\hat{H}_M(\hat{\theta}_M)^{-1}S_M(\hat{\theta}_M), \quad \dot{\rho}_M(\theta_M) \equiv \frac{\partial \rho_M(\theta_M)'}{\partial \theta_M}. \tag{3.26}$$

Standard arguments yield the distributional approximation

$$n^{1/2}\tilde{\rho}_M \overset{a}{\sim} N\left(n^{1/2}\rho_M(\theta_{M0}), \dot{\rho}_M(\theta_{M0})'\,\Sigma\,\dot{\rho}_M(\theta_{M0})\right). \tag{3.27}$$

Again, an approach similar to those applied to the elements of β allows us to define the augmented variables:

$$A_\rho \equiv \begin{bmatrix} 0 & 0_p' \\ 0_{2p+1-|M|} & A \end{bmatrix} \in \mathbb{R}^{(2p+2-|M|)\times(p+1)}, \quad Z_\rho \equiv \begin{bmatrix} n^{1/2}\tilde{\rho}_M \\ Z \end{bmatrix} \in \mathbb{R}^{p+1}, \tag{3.28}$$

$$b_\rho \equiv \begin{bmatrix} 0 \\ b \end{bmatrix} \in \mathbb{R}^{2p+2-|M|}, \quad \hat{\Sigma}_\rho \equiv \begin{bmatrix} \hat{\Sigma}_{\rho\rho} & \hat{\Sigma}_{\rho\beta} & \hat{\Sigma}_{\rho s} \\ \hat{\Sigma}_{\beta\rho} & \hat{\Sigma}_{\beta\beta} & \hat{\Sigma}_{\beta s} \\ \hat{\Sigma}_{s\rho} & \hat{\Sigma}_{s\beta} & \hat{\Sigma}_{ss} \end{bmatrix} \in \mathbb{R}^{(p+1)\times(p+1)}, \tag{3.29}$$

where Z, A and b are as defined in (3.12), and

$$\begin{aligned}
\hat{\Sigma}_{\rho\rho} &\equiv \dot{\rho}_M(\hat{\theta}_M)'\hat{H}_M(\hat{\theta}_M)^{-1}\hat{I}_M(\hat{\theta}_M)\hat{H}_M(\hat{\theta}_M)^{-1}\dot{\rho}_M(\hat{\theta}_M), \\
\hat{\Sigma}_{\rho\beta} &\equiv \dot{\rho}_M(\hat{\theta}_M)'\hat{H}_M(\hat{\theta}_M)^{-1}\hat{I}_M(\hat{\theta}_M)\hat{H}_M(\hat{\theta}_M)^{-1}[0_{|M|-1}, I_{|M|-1}]', \\
\hat{\Sigma}_{\rho s} &\equiv \dot{\rho}_M(\hat{\theta}_M)'\hat{H}_M(\hat{\theta}_M)^{-1}\hat{I}_{M(-M)}(\hat{\theta}_M) - \dot{\rho}_M(\hat{\theta}_M)'\hat{H}_M(\hat{\theta}_M)^{-1}\hat{I}_M(\hat{\theta}_M)\hat{H}_M(\hat{\theta}_M)^{-1}\hat{H}_{M(-M)}(\hat{\theta}_M).
\end{aligned}$$

Then, for $\zeta \in (0,1)$, the level $1-\zeta$ CI for $\rho_M(\theta_{M0})$ can be constructed as in (3.24) and (3.25) by replacing A, Z, b and e_{jp} by A_ρ, Z_ρ, b_ρ and $e_{j(p+1)}$, respectively, and letting $r_\rho = r(Z_\rho, \hat{\Sigma}_\rho, e_{j(p+1)})$ in (3.17).

We can also infer the parameter $\rho_M(\theta_{M0})$ by conditioning on the sign of the estimated parameter $\rho_{\hat{M}}(\hat{\theta}_{\hat{M}})$ in addition to the event $\{AZ \le b\}$ considered previously in (3.11). To this end, let $\mathtt{s}_M^\rho \equiv \text{sign}(\rho_M(\theta_{M0}))$ and $\hat{\mathtt{s}}_M^\rho \equiv \text{sign}(\rho_M(\hat{\theta}_M))$ and redefine

$$A_\rho \equiv \begin{bmatrix} -\mathtt{s}_M^\rho & 0_p' \\ 0_{2p+1-|M|} & A \end{bmatrix} \in \mathbb{R}^{(2p+2-|M|)\times(p+1)},\ b_\rho \equiv \begin{bmatrix} -\mathtt{s}_M^\rho\,\dot{\rho}_M(\hat{\theta}_M)'\hat{H}_M(\hat{\theta}_M)^{-1}(0, \lambda\,\mathtt{s}_M')' \\ b \end{bmatrix} \in \mathbb{R}^{2p+2-|M|},$$

and keep Z_ρ and $\hat{\Sigma}_\rho$ as defined in (3.28) and (3.29). Then, we can rewrite the event

$\{\mathtt{s}_M^\rho = \hat{\mathtt{s}}_M^\rho\} = \{\mathtt{s}_M^\rho = \text{sign}(\rho_M(\hat{\theta}_M))\}$ as

$$\begin{aligned}\{\mathtt{s}_M^\rho = \text{sign}(\rho_M(\hat{\theta}_M))\} &= \{\mathtt{s}_M^\rho\, \rho_M(\hat{\theta}_M) > 0\} \\ &= \{\mathtt{s}_M^\rho\, (\tilde{\rho}_M - \dot{\rho}_M(\hat{\theta}_M)'\hat{H}_M(\hat{\theta}_M)^{-1}(0, \lambda \mathtt{s}_M')') > 0\} \\ &= \{-\mathtt{s}_M^\rho\, \tilde{\rho}_M < -\lambda\, \mathtt{s}_M^\rho\, \dot{\rho}_M(\hat{\theta}_M)'\hat{H}_M(\hat{\theta}_M)^{-1}(0, \mathtt{s}_M')'\}. \qquad (3.30)\end{aligned}$$

Therefore, the event $\{\hat{M} = M, \hat{\mathtt{s}}_{\hat{M}} = \mathtt{s}_M, \hat{\mathtt{s}}_{\hat{M}}^\rho = \mathtt{s}_M^\rho\}$ is equivalent to the affine constraint $A_\rho Z_\rho \leq b_\rho$. We proceed similarly to the subvector case considered previously to obtain the SI CI for $\rho_M(\theta_{M0})$. Let $r_\rho = r(Z_\rho, \hat{\Sigma}_\rho, e_{j(p+1)})$ as in (3.17), and fix $\zeta \in (0,1)$. The SI CI of level $1-\zeta$ for $\rho_M(\theta_{M0})$ is given by $\text{CI}_{\hat{M}}^\rho \equiv [\tilde{q}_l^\rho, \tilde{q}_u^\rho]$, where $\tilde{q}_l^\rho$ and $\tilde{q}_u^\rho$ are the solutions respectively to the following equations

$$F(n^{1/2}q, n^{1/2}\tilde{\rho}_M, e_{j(p+1)}'\hat{\Sigma}_\rho e_{j(p+1)}, \mathcal{V}^-(r_\rho), \mathcal{V}^+(r_\rho)) = \frac{\zeta}{2}, \qquad (3.31)$$

$$F(n^{1/2}q, n^{1/2}\tilde{\rho}_M, e_{j(p+1)}'\hat{\Sigma}_\rho e_{j(p+1)}, \mathcal{V}^-(r_\rho), \mathcal{V}^+(r_\rho)) = 1 - \frac{\zeta}{2}. \qquad (3.32)$$

We summarize the asymptotic validity of the above CI in the next corollary which follows from the arguments similar to the proof of Proposition 3.2.

Corollary 3.3. *Suppose that the conditions of Proposition 3.2 hold, and the scalar non-linear parameter function $\rho_M(\theta_M)$ is continuously differentiable in a neighborhood of θ_{M0} with $\dot{\rho}_M(\theta_{M0})'\dot{\rho}_M(\theta_{M0}) > \lambda_l > 0$, Then, it holds that for $\zeta \in (0,1)$*

$$\liminf_{n\to\infty}\ P[\rho_{\hat{M}} \in \text{CI}_{\hat{M}}^\rho | \hat{M} = M, \hat{\mathtt{s}}_{\hat{M}} = \mathtt{s}_M, \hat{\mathtt{s}}_{\hat{M}}^\rho = \mathtt{s}_M^\rho] = 1 - \zeta.$$

3.2 Debiased Lasso inference

The debiased Lasso method of Zhang and Zhang (2014) and Javanmard and Montanari (2014) is based on the one-step estimator constructed from the initial Lasso estimator $\hat{\theta}$:

$$\tilde{\theta} = \hat{\theta} + \hat{H}(\hat{\theta})^{-1}S(\hat{\theta}). \qquad (3.33)$$

This particular variant of the debiased Lasso that employs the standard Hessian is proposed by Xia et al. (2021) for a homoskedastic GLM. Similarly, we use $\hat{I}(\hat{\theta})$ to estimate the asymptotic variance of $n^{1/2}S(\theta_0)$ and $I(\theta_0)$. To show the consistency of $\hat{H}(\hat{\theta})$ and $\hat{I}(\hat{\theta})$, we first extend Corollary 5.50 of Vershynin (2010) to random matrices i.n.i.d. rows with non-identical second moment matrices in the following lemma.

Lemma 3.4 (Covariance matrix consistency for i.n.i.d. random vectors.)**.** *Let A be an $n \times p$ matrix whose rows A_i' are independent sub-Gaussian random vectors in $\mathbb{R}^p$ with $\mathrm{E}[A_i] = \mu_i$, $\mathrm{E}[A_i A_i'] = \Sigma_i$ and $0 < \lambda_l < \lambda_{\min}(\bar{\Sigma}_n) < \infty$, where $\bar{\Sigma}_n \equiv n^{-1}\sum_{i=1}\Sigma_i$. Then for every $t \geq 0$, with probability at least $1 - 2\exp(-t^2)$ it holds that*

$$\|n^{-1}A'A - \bar{\Sigma}_n\|_2 \leq C_K \max(\delta, \delta^2)\|\bar{\Sigma}_n\|_2, \quad \delta \equiv c\left(\sqrt{\frac{p}{n}} + \frac{t}{\sqrt{n}}\right), \tag{3.34}$$

where c is an absolute constant and $C_K > 0$ is a constant that depend only on the sub-Gaussian norm $K = \max_i \|A_i\|_{\psi_2} < \infty$ of the rows and λ_l.

See Appendix A.2 for a proof. Relative to Theorem 5.39 and Corollary 5.50 of Vershynin (2010), the invertibility of $\bar{\Sigma}_n$ is required in Lemma 3.4, but the rows of the matrix A can be heterogeneous with non-identical second moment matrices $\Sigma_i, i = 1, \ldots, n$. Lemma 3.4 together with Lemma S2 of Xia et al. (2021) yields the following result.

Lemma 3.5 (The rate of convergence of the Hessian and information matrices)**.** *Under Assumption 1,*

$$\|\hat{H}(\hat{\theta}) - H(\theta_0)\|_2 = O_p\left(\sqrt{\frac{p}{n}} + m_0\lambda\right), \tag{3.35}$$

$$\|\hat{H}(\hat{\theta})^{-1} - H(\theta_0)^{-1}\|_2 = O_p\left(\sqrt{\frac{p}{n}} + m_0\lambda\right), \tag{3.36}$$

$$\|\hat{I}(\hat{\theta}) - I(\theta_0)\|_2 = O_p\left(\sqrt{\frac{p}{n}} + m_0\lambda\right), \tag{3.37}$$

$$\|\hat{I}(\hat{\theta})^{-1} - I(\theta_0)^{-1}\|_2 = O_p\left(\sqrt{\frac{p}{n}} + m_0\lambda\right). \tag{3.38}$$

The proof is provided in Appendix A.3 which essentially verifies that the argument of Xia et al. (2021) goes through with i.n.i.d. data.

For inference on a $r \times 1$ vector nonlinear parameter function $\rho(\theta)$ (which may depend

on n), we define a debiased Lasso (one-step estimator) as

$$\tilde{\rho} \equiv \rho(\hat{\theta}) + \dot{\rho}(\hat{\theta})'\hat{H}(\hat{\theta})^{-1}S(\hat{\theta}), \quad \dot{\rho}(\theta) \equiv \frac{\partial \rho(\theta)'}{\partial \theta}. \tag{3.39}$$

We establish the asymptotic validity of Wald-type inference based on the debiased Lasso estimator above in the proposition below.

Proposition 3.6 (Asymptotic validity of Survey Debiased Lasso test)**.** *Let Assumption 1 hold and assume that* $\lambda = C\sqrt{\frac{\log p}{n}}$ *with* $C = O(1)$ *and* $p \geq 1$, $p^2/n \to 0$ *and* $m_0 \log p\sqrt{\frac{p}{n}} \to 0$ *as* $n \to \infty$*. If the* $r \times 1$ *function* $\rho(\theta)$ *is differentiable in a neighborhood of* θ_0 *with a locally Lipschitz Jacobian* $\dot{\rho}(\theta)$ *and* $\lambda_{\min}(\dot{\rho}(\theta_0)'\dot{\rho}(\theta_0)) > \lambda_l > 0$*, where* $r(< (p+1))$ *is fixed, then*

$$\left(\dot{\rho}(\hat{\theta})'\hat{H}(\hat{\theta})^{-1}\hat{I}(\hat{\theta})\hat{H}(\hat{\theta})^{-1}\dot{\rho}(\hat{\theta})\right)^{-1/2} n^{1/2}(\tilde{\rho} - \rho(\theta_0)) \xrightarrow{d} N(0, I_r). \tag{3.40}$$

The proof is given in Appendix A.4. $\lambda = C\sqrt{\frac{\log p}{n}}$ is a standard assumption in the literature (see e.g. Bühlmann and van de Geer, 2011; Negahban et al., 2012; van de Geer et al., 2014; Hastie et al., 2015). The assumptions imposed on the number of covariates p, and the model sparsity m_0 are the same as those in Xia et al. (2021). In particular, while the condition $m_0 \log p\sqrt{\frac{p}{n}} \to 0$ is stronger than the condition $m_0 \frac{\log p}{\sqrt{n}} \to 0$ assumed by van de Geer et al. (2014), no assumption is imposed directly on the sparsity of the inverse Hessian (and information matrix) i.e. $\max_j m_j = o(n/\log p)$, where $m_j \equiv |\{k \neq j : (H(\theta_0)^{-1})_{jk} \neq 0\}|$ is the number of non-zero elements of the j-th row of $H(\theta_0)^{-1}$, as in van de Geer et al. (2014). As noted by Xia et al. (2021), the condition $p^2/n \to 0$ is weaker than the condition $\max_j m_j = o(n/\log p)$, when m_j is of the order p.

The assumption of locally Lipschitz Jacobian $\dot{\rho}(\theta)$ is slightly stronger than the usual continuous differentiability assumption required for testing nonlinear hypotheses (see e.g. Section 9 of Newey and McFadden (1994) and Hansen (2022*a*,*b*)). Under this assumption, an error term $n^{1/2}(\dot{\rho}(\hat{\theta}) - \dot{\rho}(\bar{\theta}))'(\hat{\theta} - \theta_0)$, where $\bar{\theta}$ is a mean-value between $\hat{\theta}$ and θ_0, that results from the estimation of θ_0 and $\rho(\theta_0)$ becomes negligible.

Using Proposition 3.6, we obtain confidence intervals for the elements of θ_0 as well as the vector nonlinear parameter function $\rho(\theta_0)$. One can also consider a plug-in estimator $\rho(\tilde{\theta})$, where $\tilde{\theta}$ is the one-step estimator defined in (3.33). This estimator is asymptotically

equivalent to the one-step estimator $\tilde{\rho}$ in (3.39). The proof is actually similar to that of Proposition 3.6, thus is omitted. In addition, multi-step estimators of θ_0 and $\rho(\theta_0)$ can also be considered.

3.3 $C(\alpha)$/Orthogonalization inference

Belloni et al. (2016) develop subvector inference procedure in a high-dimensional GLM that satisfies sparsity assumptions. They construct an estimating equation orthogonalized against the direction of the nuisance parameter estimation which also underlies the Neyman (1959)'s $C(\alpha)$ test. Here, we consider a survey version of the $C(\alpha)$-type statistic for the $r \times 1$ nonlinear parameter functon $\rho(\theta)$ defined as

$$C_\alpha(\rho_0) \equiv n\, S(\tilde{\theta}^*)'\hat{H}(\tilde{\theta}^*)^{-1}\dot{\rho}(\tilde{\theta}^*)\left(\dot{\rho}(\tilde{\theta}^*)'\hat{H}(\tilde{\theta}^*)^{-1}\hat{I}(\tilde{\theta}^*)\hat{H}(\tilde{\theta}^*)^{-1}\dot{\rho}(\tilde{\theta}^*)\right)^{-1}\dot{\rho}(\tilde{\theta}^*)'\hat{H}(\tilde{\theta}^*)^{-1}S(\tilde{\theta}^*), \tag{3.41}$$

where $\tilde{\theta}^*$ is an auxiliary estimate that satisfies $\rho(\tilde{\theta}^*) = \rho_0$. This test statistic is proposed, in a regular likelihood context, by Smith (1987) and studied further by Dufour et al. (2016) among others.

Proposition 3.7 (Asymptotic validity of Survey $C(\alpha)$ test)**.** *Let Assumption 1 hold and assume that $\lambda = C\sqrt{\frac{\log p}{n}}$ with $C = O(1)$, $p^2/n \to 0$ and $m_0 \log p\sqrt{\frac{p}{n}} \to 0$ as $n \to \infty$. Let $\tilde{\theta}^*$ be an auxiliary estimator that satisfies $\|\tilde{\theta}^* - \theta_0\|^2 = O_p(m_0\lambda^2)$ and $\rho(\tilde{\theta}^*) = \rho_0$, where the nonlinear parameter function $\rho(\theta) \in \mathbb{R}^r$ satisfies the conditions given in Proposition 3.6. Then, under $H_0 : \rho(\theta_0) = \rho_0$*

$$C_\alpha(\rho_0) \xrightarrow{d} \chi^2_r. \tag{3.42}$$

The proof is given in Appendix A.5. In general, determining an auxiliary estimator that satisfies the constraint $\rho(\tilde{\theta}^*) = \rho_0$ may be difficult. However, as we show in the next section, when testing a restriction on the AME of a binary regressor in the logit model, such an estimator can be readily obtained.

4 Survey logit

This section applies the results established in the previous sections to inference on the logit model estimated by the Lasso from survey data. The standard logit specification for a dependent variable $y_i, i = 1, \ldots, n$, is

$$P[y_i = 1|x_i] = \Lambda(x_i'\theta), \tag{4.1}$$

where $x_i = (1, \tilde{x}_i')' \in \mathbb{R}^{p+1}$, $\tilde{x}_i = (\tilde{x}_{i1}, \ldots, \tilde{x}_{ip})' \in \mathbb{R}^p$, and $\theta = (\alpha, \beta')' \in \mathbb{R}^{p+1}$, $\alpha \in \mathbb{R}$, $\beta \in \mathbb{R}^p$. Given the survey weights $\{w_i\}_{i=1}^n$ on the observations $\{(y_i, x_i')'\}_{i=1}^n$, the weighted log-likelihood function is

$$L(\theta) = n^{-1} \sum_{i=1}^{n} w_i (y_i x_i'\theta - \log(1 + \exp(x_i'\theta))). \tag{4.2}$$

The score function, the sample information and negative Hessian functions are given by

$$S(\theta) = \frac{\partial L(\theta)}{\partial \theta} = n^{-1} \sum_{i=1}^{n} w_i x_i (y_i - \Lambda(x_i'\theta)), \tag{4.3}$$

$$\hat{I}(\theta) = n^{-1} \sum_{i=1}^{n} w_i^2 x_i x_i' \Lambda(x_i'\theta)(1 - \Lambda(x_i'\theta)), \tag{4.4}$$

$$\hat{H}(\theta) = -\frac{\partial^2 L(\theta)}{\partial \theta \partial \theta'} = n^{-1} \sum_{i=1}^{n} w_i x_i x_i' \Lambda(x_i'\theta)(1 - \Lambda(x_i'\theta)). \tag{4.5}$$

4.1 Inference on average marginal effects

In the context of the logit model, a key parameter of interest is the AME which is a nonlinear function of the model parameters. As such, this section focuses on the inference on AMEs. The marginal effect (ME) of a binary regressor $\tilde{x}_{ij}, j = 1, \ldots, p, i = 1, \ldots, n$, with a coefficient $\theta_{(j)}$ is calculated by the change in $P[y_i = 1|x_i]$ when the regressor $\tilde{x}_{ij}$ is switched from 0 to 1 holding all other variables constant:

$$\mathrm{ME}_{ij}(\theta) \equiv \Lambda(x_i'\theta)|_{\tilde{x}_{ij}=1} - \Lambda(x_i'\theta)|_{\tilde{x}_{ij}=0}. \tag{4.6}$$

The AME of the j-th regressor is defined as

$$\mathrm{AME}_j = \mathrm{AME}_j(\theta_0) \equiv \mathrm{E}\left[\frac{1}{\sum_{i=1}^n w_i}\sum_{i=1}^n w_i \mathrm{ME}_{ij}(\theta_0)\right],$$

where θ_0 denotes the true value of θ and the expectation is taken with respect to the distribution of the regressors.

Let us first consider the debiased Lasso inference for the AMEs. A natural estimator of $\mathrm{AME}_j(\theta_0)$ is

$$\widehat{\mathrm{AME}}_j(\hat{\theta}) \equiv \frac{1}{\sum_{i=1}^n w_i}\sum_{i=1}^n w_i \left(\Lambda(x_i'\hat{\theta})|_{\tilde{x}_{ij}=1} - \Lambda(x_i'\hat{\theta})|_{\tilde{x}_{ij}=0}\right),$$

where $\hat{\theta} = (\hat{\alpha}, \hat{\beta}')'$ is an estimator of θ_0 e.g. the survey-weighted Lasso estimator. In the current context, the one-step estimator defined in (3.33) specializes to

$$\widetilde{\mathrm{AME}}_j = \widehat{\mathrm{AME}}_j(\hat{\theta}) + \frac{\partial \widehat{\mathrm{AME}}_j(\hat{\theta})}{\partial \theta'}\hat{H}(\hat{\theta})^{-1}S(\hat{\theta}),$$

where

$$\frac{\partial \widehat{\mathrm{AME}}_j(\hat{\theta})}{\partial \theta} \equiv \frac{1}{\sum_{i=1}^n w_i}\sum_{i=1}^n w_i \left\{\left[x_i\Lambda(x_i'\hat{\theta})(1-\Lambda(x_i'\hat{\theta}))\right]|_{\tilde{x}_{ij}=1} - \left[x_i\Lambda(x_i'\hat{\theta})(1-\Lambda(x_i'\hat{\theta}))\right]|_{\tilde{x}_{ij}=0}\right\}.$$

To obtain a confidence interval for $\mathrm{AME}_j(\theta_0), j = 2, \ldots, p+1$, we can then use

$$\left(\frac{\partial \widehat{\mathrm{AME}}_j(\hat{\theta})}{\partial \theta'}\hat{H}(\hat{\theta})^{-1}\hat{I}(\hat{\theta})\hat{H}(\hat{\theta})^{-1}\frac{\partial \widehat{\mathrm{AME}}_j(\hat{\theta})}{\partial \theta}\right)^{-1/2} n^{1/2}(\widetilde{\mathrm{AME}}_j - \mathrm{AME}_j) \xrightarrow{d} N(0,1).$$

Next, we turn to the SI. Let $\mathrm{AME}_M(\theta_M) = [\mathrm{AME}_{M2}(\theta_M), \ldots, \mathrm{AME}_{MM}(\theta_M)]' \in \mathbb{R}^{|M|-1}$ denote the AMEs for the active variables selected by the survey-weighted Lasso with coefficients β_M. Then, from (3.26) the SI for the AMEs in the selected model is based on the one-step estimator

$$\widetilde{\mathrm{AME}}_M = \widehat{\mathrm{AME}}_M(\hat{\theta}_M) + \frac{\partial \widehat{\mathrm{AME}}_M(\hat{\theta}_M)}{\partial \theta_M'}\hat{H}_M(\hat{\theta}_M)^{-1}S_M(\hat{\theta}_M). \tag{4.7}$$

Finally, we consider the $C(\alpha)$ statistic. Let $C_\alpha(\mathrm{AME}_{0j})$ denote the $C(\alpha)$ statistic for testing

$H_0 : \text{AME}_j = \text{AME}_{0j}$. To obtain an auxiliary estimate that satisfies $\text{AME}_j(\tilde{\theta}^*) = \text{AME}_{0j}$, we only need to solve for a scalar $\theta_{(j)}$ in the following equation:

$$\frac{1}{\sum_{i=1}^n w_i} \sum_{i=1}^n w_i \left(\Lambda \left(\theta_{(j)} + x'_{i(-j)} \hat{\theta}_{(-j)} \right) - \Lambda \left(x'_{i(-j)} \hat{\theta}_{(-j)} \right) \right) = \text{AME}_{0j}.$$

Testing the zero restriction $H_0 : \text{AME}_j = 0$ is particularly simple. First, note that $\text{AME}_j = 0$ if $\theta_{(j)} = 0$. Furthermore, the Jacobian used in the $C_\alpha(\text{AME}_{0j})$ statistic is

$$\begin{aligned} \frac{\partial \widehat{\text{AME}}_j(\theta)}{\partial \theta} &\equiv \frac{1}{\sum_{i=1}^n w_i} \sum_{i=1}^n w_i \left[0, \ldots, \left(\Lambda(x'_i\theta)(1 - \Lambda(x'_i\theta)) \right) |_{\tilde{x}_{ij}=1}, \ldots, 0 \right]' \\ &= \frac{1}{\sum_{i=1}^n w_i} \sum_{i=1}^n w_i \Lambda(x'_i\theta)(1 - \Lambda(x'_i\theta))|_{\tilde{x}_{ij}=1} e_{j(p+1)}. \end{aligned}$$

Let $\tilde{\theta}^*$ denote the estimator when the j-th element of the Lasso estimator $\hat{\theta}$ is replaced by 0. Then, we have

$$\begin{aligned} & C_\alpha(\text{AME}_{0j}) \\ & = n\, S(\tilde{\theta}^*)' \hat{H}(\tilde{\theta}^*)^{-1} e_{j(p+1)} \left(e'_{j(p+1)} \hat{H}(\tilde{\theta}^*)^{-1} \hat{I}(\tilde{\theta}^*) \hat{H}(\tilde{\theta}^*)^{-1} e_{j(p+1)} \right)^{-1} e'_{j(p+1)} \hat{H}(\tilde{\theta}^*)^{-1} S(\tilde{\theta}^*) \\ & = C_\alpha(\theta_{0(j)}), \end{aligned}$$

where $C_\alpha(\theta_{0(j)})$ is the $C(\alpha)$ statistic for testing the coefficient $H_0 : \theta_{(j)} = 0$. We summarize this simple observation in the following lemma.

Lemma 4.1. *The $C(\alpha)$ statistic for testing the coefficient $H_0 : \theta_{(j)} = 0$ on a binary regressor $\tilde{x}_{ij}$ based on the auxiliary estimator $\tilde{\theta}^*$ is equivalent to the $C(\alpha)$ statistic for testing the corresponding AME $H_0 : \text{AME}_j = 0$ based on $\tilde{\theta}^*$.*

5 Simulations

This section presents a simulation evidence on the performance of the proposed procedures. We consider a logit model where the regressors and the dependent variables are generated as follows:

$$y_i \sim \text{Bernoulli}(\pi_i), \tag{5.1}$$

where $\theta_0 = (1, 1, 1, 0_{1\times(p-2)})'$, $\tilde{x}_{ij} \sim i.i.d.$ Bernoulli(prob), $j = 1, \ldots, p$, $i = 1, \ldots, N$, $x_i = (1, \tilde{x}_i')'$ and $\pi_i = x_i'\theta_0$. We set the size of the population equal to $N = 10,000$. Two sampling schemes are considered: standard stratified sampling and exogenous stratification with prob $= 0.5$ and prob $= 0.4$, respectively. For each scheme, we create 4 strata and consider two cases: $(n_s, n) \in \{(50, 200), (100, 400)\}$, where n_s observations are drawn from each stratum with replacement yielding a stratified sample of size n.

In the standard stratified sampling with prob $= 0.5$, the population is stratified into 4 strata of sizes $N_1 = 1000, N_2 = 2000, N_3 = 3000$ and $N_4 = 4000$, respectively. As a result, the weights on the observations are $w_i = 0.1, 0.2, 0.3, 0.4$ corresponding to the four strata. In the exogenous stratification with prob $= 0.4$, the population is stratified according to the values of the first two non-constant regressors: $(\tilde{x}_{i1}, \tilde{x}_{i2}) \in \{(0, 0), (0, 1), (1, 0), (1, 1)\}$. The weights on the observations in the above four strata are $w_i = 0.36, 0.24, 0.24, 0.16$, respectively.

To assess the effect of the dimension of the regressors, the values for p are set such that $\frac{p}{n} \in \{0.01, 0.025, 0.05, 0.1, 0.25, 0.5\}$ for each $n \in \{200, 400\}$. The true value of AME corresponding to the coefficient $\theta_{(2)} = \beta_1$ is 0.11. The empirical size of the tests is examined by testing the following two restrictions separately:

$$H_0 : \theta_{(2)} = 1, \quad H_0 : \text{AME}_2 = 0.11. \tag{5.2}$$

To test the hypothesis on AME, we implement the two SI approaches, labeled as SI and SI2, with or without conditioning on the sign of the estimated AME, respectively, described in Section 3.1. For the auxiliary estimate $\tilde{\theta}^*$ in the $C(\alpha)$ statistic, we used the one-step iteration of $(1, \hat{\theta}_{(-2)}')'$, where 1 corresponds to the tested value and $\hat{\theta}_{(-2)}$ is the (weighted) logistic Lasso estimate of $\theta_{(-2)}$, the model coefficients other than $\theta_{(2)}$. Moreover, whenever the sample Hessian evaluated at $\tilde{\theta}^*$ in the $C(\alpha)$ statistic is found to be singular, we used the Moore-Penrose inverse. There was no such issue in the other test statistics.

The model (2.6) is fit using the `R` package `glmnet`. For the tuning parameter λ, we use the default value of the package which is chosen by 10-fold cross validation with loss function "`auc`" (area under the ROC curve).

Tables 1 and 2 report the empirical sizes of the tests under standard stratified sampling

and exogenous stratification, respectively. The results under both sampling schemes are qualitatively similar. All tests show reasonable size control when the number of regressors is moderate i.e. $p/n = 0.01, 0.025, 0.05, 0.1, 0.25$ for both hypotheses. We can also see that the SI tests tend to underreject in most cases while the $C(\alpha)$ test does so when $p/n = 0.5$. The size distortions of the SI method could potentially be alleviated by considering an appropriate form of bootstrap. When $p/n = 0.5$, that is, the number of covariates is large relative to the sample size, all tests tend to underreject. This may be attributed to the conditions imposed on the growth rate of p relative to the degree of sparsity, the tuning parameter and the sample size which are needed for the asymptotic validity of the DB and $C(\alpha)$ tests given in Propositions 3.6 and 3.7.

Moreover, when $p/n = 0.5$, the $C(\alpha)$ test exhibits a substantial size distortion, while the DB and SI tests show somewhat better performance despite the fact that, in this case, the number of covariates are too high relative to the sample size for our results to hold. It is also clear that the rejection rate of the survey t-test, denoted as t_{svy}, deteriorates as the ratio p/n grows, which is expected as the test is not robust to increasing number of covariates.

6 Empirical application

This section applies the proposed methods to Canadian Internet Use Survey (CIUS) 2020 data, and examines what demographic factors affect a person's access to a government program or service.[3] The dependent variable is a binary variable where respondents answered 1) *yes*; 2) *no*; 3) *not stated* to the question "During the past 12 months, what activities did you perform on the Internet to interact with the government in Canada? Was it: Accessed an account for a government program or service?"

The covariates in this analysis are income, education, employment status, aboriginal identity, visible minority status, immigration status, gender, type of household, language spoken at home, and province. All have two or more categories. There are $n = 17,031$ observations in the survey.

The collection of CIUS 2020 is based on a stratified design employing probability sam-

[3] Available at `https://www150.statcan.gc.ca/n1/daily-quotidien/210622/dq210622b-eng.htm`

Table 1: Empirical rejection frequencies of the tests for $H_0 : \theta_{(2)} = 1$ and $H_0 : \text{AME}_2 = 0.11$ at 5% level. Standard stratified sampling.

Tests	$p = 2$	$p = 5$	$p = 10$	$p = 20$	$p = 50$	$p = 100$
$H_0 : \theta_{(2)} = 1$, $n = 200$						
DB	5.0	4.4	3.7	3.1	4.5	3.3
$C(\alpha)$	5.5	4.1	3.1	2.8	4.2	0.3
SI	3.9	2.5	2.3	2.6	3.6	3.6
t_{svy}	6.2	6.4	8.0	8.7	36.0	94.9
$H_0 : \text{AME}_2 = 0.11$, $n = 200$						
DB	5.4	5.3	4.6	3.7	3.5	1.4
$C(\alpha)$	6.1	6.4	4.9	5.4	4.2	1.3
SI	4.2	2.6	2.2	2.8	3.6	4.7
SI2	4.2	2.6	2.4	2.8	3.5	4.3
t_{svy}	5.7	7.7	7.4	8.2	50.9	93.3

Tests	$p = 4$	$p = 10$	$p = 20$	$p = 40$	$p = 100$	$p = 200$
$H_0 : \theta_{(2)} = 1$, $n = 400$						
DB	4.8	4.4	6.0	3.7	5.6	3.9
$C(\alpha)$	4.7	4.2	4.2	4.2	5.7	0.5
SI	5.5	3.6	3.7	3.0	3.9	2.9
t_{svy}	5.0	5.1	6.3	15.9	40.4	98.3
$H_0 : \text{AME}_2 = 0.11$, $n = 400$						
DB	4.5	4.9	5.8	5.0	4.6	3.3
$C(\alpha)$	5.2	6.9	8.5	3.1	3.4	1.0
SI	5.1	4.4	4.6	2.8	3.9	3.2
SI2	5.1	4.4	4.6	2.9	3.7	2.8
t_{svy}	5.3	6.9	9.1	10.8	46.8	93.7

Notes: $n = 200, 400$ and 1000 simulation replications. DB, $C(\alpha)$, SI and t_{svy} denote the debiased Lasso, $C(\alpha)$, selective inference and standard survey-weighted t tests respectively. For the restriction $H_0 : \text{AME}_2 = 0.11$, SI is conditional on the sign of the estimated AME in addition to $\hat{M} = M, \hat{\mathbf{s}}_{\hat{M}} = \mathbf{s}_M$ while SI2 is conditional on the latter only.

Table 2: Empirical rejection frequencies of the tests for $H_0 : \theta_{(2)} = 1$ and $H_0 : \text{AME}_2 = 0.11$ at 5% level. Exogenous stratification.

Tests	$p = 2$	$p = 5$	$p = 10$	$p = 20$	$p = 50$	$p = 100$
$H_0 : \theta_{(2)} = 1$, $n = 200$						
DB	4.9	4.8	3.1	4.1	7.3	4.6
$C(\alpha)$	6.4	5.3	4.0	4.0	8.4	1.8
SI	4.4	2.1	2.7	2.2	2.9	4.1
t_{svy}	5.1	5.1	6.6	6.1	31.8	95.1
$H_0 : \text{AME}_2 = 0.11$, $n = 200$						
DB	5.4	4.9	5.3	3.8	5.6	4.3
$C(\alpha)$	6.3	5.5	3.9	4.5	6.3	1.7
SI	4.1	1.9	3.0	2.4	2.8	5.2
SI2	4.1	2.0	3.1	2.6	2.7	5.0
t_{svy}	5.9	6.1	8.8	7.6	43.4	93.6

Tests	$p = 4$	$p = 10$	$p = 20$	$p = 40$	$p = 100$	$p = 200$
$H_0 : \theta_{(2)} = 1$, $n = 400$						
DB	5.7	4.9	8.4	4.7	7.1	4.0
$C(\alpha)$	5.1	4.6	5.7	4.7	9.1	7.0
SI	7.0	5.7	4.1	4.2	3.1	2.9
t_{svy}	4.8	5.2	6.0	9.2	29.9	98.4
$H_0 : \text{AME}_2 = 0.11$, $n = 400$						
DB	4.7	5.7	4.3	5.1	4.8	4.6
$C(\alpha)$	4.6	5.3	5.6	4.5	4.3	0.3
SI	3.8	4.2	3.3	3.0	3.3	5.1
SI2	3.8	4.2	3.3	3.0	3.2	4.8
t_{svy}	6.0	6.9	5.3	8.9	43.9	91.0

Notes: $n = 200, 400$ and 1000 simulation replications. DB, $C(\alpha)$, SI and t_{svy} denote the debiased Lasso, $C(\alpha)$, selective inference and standard survey-weighted t tests respectively. For the restriction $H_0 : \text{AME}_2 = 0.11$, SI is conditional on the sign of the estimated AME in addition to $\hat{M} = M, \hat{\mathbf{s}}_{\hat{M}} = \mathbf{s}_M$ while SI2 is conditional on the latter only.

pling; the stratification is done at the province/census metropolitan area (CMA) and census agglomeration (CA) level where each of the ten Canadian provinces were divided into strata/geographic areas.[4] Each record on a sampling frame used in CIUS 2020 is a group of one or several telephone numbers associated with the same address from the Census and various administrative sources with Statistics Canada's dwelling frame. The records—the groups of telephone numbers—were sampled independently without replacement from each stratum.

The initial weight on observations is the inverse of an adjusted version of the probability of selection equal to the number of records sampled in the stratum divided by the number of records in the stratum from the survey frame. The final person weight w_i is an adjusted version of the initial weight that takes into account the household size and survey-response among others.[5]

The base categories are omitted in each model as the comparison category for the logit model. The representative individual in the base category has the following characteristics – *male, non-aboriginal, neither English nor French* (e.g. English and non-official language) speaker, *not employed, some post-secondary education, not a visible minority, family household with children under 18,* income of $\$44,120$-$\$75,321$, *landed immigrant* (recent immigrant), and from the province Alberta.

Table 3 reports the inference results for the logit coefficients. The survey logit Lasso selects *French, Employed, High school or less, University degree, Visible Minority, Family household with no children under 18*, and *Single*. The magnitude of the Lasso estimates are in line with the survey logit estimates, and the signs of the estimates also appear reasonable. All inference methods indicate that the coefficients on *Employed, University degree* and *Visible minority* are highly significant. The variable *French* is selected by the Lasso but the inference results show that its coefficient is far from being significant.

It is interesting to note that although New Brunswick (NB) is not selected by the Lasso, its debiased Lasso estimate -0.32 is almost identical to the survey GLM estimate -0.33

[4]There are 151 strata with the largest stratum, Toronto, having 2,235,145 private dwellings and the smallest stratum, Elliot Lake, having 6,259 private dwellings as of 2016, see `https://www12.statcan.gc.ca/census-recensement/2016/dp-pd/hlt-fst/pd-pl/Table.cfm?Lang=Eng&T=201&SR=1&S=3&O=D&RPP=9999&PR=0`

[5]Further details of the weighting procedure can be found in Section 10 of Microdata user Guide, CIUS 2020 at `https://www23.statcan.gc.ca/imdb/p2SV.pl?Function=getSurvey$&$SDDS=4432$#$a2`

and highly significant.

Table 4 displays the inference results for the AMEs. The employed are about 8-11 percentage points more likely to use the government online service than those not employed. Moreover, the use of government online service in NB appears to be 6-7 percentage points lower than the level of Alberta (AB).

The debiased Lasso and $C(\alpha)$ test results in Table 4 show that the family household without children under age 18 is less likely to use the government service than those with children under age 18. Moreover, low educational attainment and high income negatively affect the likelihood of an individual using the government online services. The variables with the largest (in absolute value) AMEs on whether a person uses government online services are whether or not a person is employed, whether or not a person is single, and if their educational attainment was a *High school or less* or a *University degree*.

7 Conclusion

This paper has provided two main results. First, we have extended Lasso inference methods to a GLM with survey weights and/or heteroskedasticity, and established their asymptotic validity. Second, we have considered inference on nonlinear parameter functions. The proposed extended inference methods were applied to the logit model and remain reliable when p/n increases as illustrated in a simulation study with standard stratified sampling and exogenous stratification. An empirical illustration based on the CIUS 2020 data also confirms the relevance of the proposed approach.

Table 3: Point estimates of θ_0 and test p-values for $H_0 : \theta_{0(j)} = 0$ in Lasso Logistic Regression for Government Online Service Access.

	Estimator				p-values			
Variable	GLM	Lasso	DB	SI	t_{svy}	DB	$C(\alpha)$	SI
Intercept	−0.53	−0.51	−0.52	−0.37	0.23	0.18	0.18	0.00
Female	−0.02	0.00	−0.02	–	0.61	0.61	0.61	–
Aboriginal	0.11	0.00	0.10	–	0.43	0.45	0.45	–
Aboriginal n.s.	0.85	0.00	0.71	–	0.11	0.16	0.16	–
English	0.30	0.00	0.28	–	0.49	0.44	0.44	–
French	−0.18	−0.22	−0.15	−0.59	0.68	0.70	0.78	1.00
English and French	0.48	0.00	0.46	–	0.27	0.22	0.22	–
Language n.s.	−0.15	0.00	−0.14	–	0.80	0.79	0.79	–
Employed	0.36	0.30	0.36	0.34	0.00	0.00	0.00	0.00
Employment n.s.	0.35	0.00	0.32	–	0.29	0.33	0.33	–
High school or less	−0.53	−0.41	−0.51	−0.51	0.00	0.00	0.00	1.00
University degree	0.37	0.32	0.37	0.35	0.00	0.00	0.00	0.00
Education n.s.	−0.99	0.00	−0.78	–	0.01	0.06	0.06	–
Visible minority	0.27	0.17	0.27	0.25	0.00	0.00	0.00	0.00
Visible minority n.s.	0.40	0.00	0.33	–	0.33	0.41	0.41	–
Family household w.o.c.u 18	−0.29	−0.08	−0.28	−0.28	0.00	0.00	0.00	1.00
Single	−0.72	−0.31	−0.70	−0.65	0.00	0.00	0.00	1.00
Other household type	−0.07	0.00	−0.07	–	0.62	0.64	0.64	–
Family n.s.	−0.17	0.00	−0.17	–	0.48	0.48	0.48	–
$44,119 and less	−0.01	0.00	−0.01	–	0.92	0.94	0.94	–
$75,322–$109,431	−0.04	0.00	−0.04	–	0.62	0.62	0.62	–
$109,432–$162,799	−0.02	0.00	−0.02	–	0.80	0.80	0.80	–
$162,800 and higher	−0.22	0.00	−0.21	–	0.01	0.01	0.01	–
Non-landed immigrant	0.01	0.00	0.01	–	0.86	0.87	0.87	–
Immigration n.s.	0.90	0.00	0.91	–	0.29	0.31	0.31	–
NL	0.14	0.00	0.14	–	0.20	0.20	0.20	–
PEI	−0.03	0.00	−0.03	–	0.78	0.78	0.78	–
NS	−0.21	0.00	−0.21	–	0.05	0.06	0.06	–
NB	−0.33	0.00	−0.32	–	0.01	0.00	0.00	–
QC	−0.25	0.00	−0.25	–	0.02	0.02	0.02	–
ON	−0.16	0.00	−0.16	–	0.06	0.06	0.06	–
MB	−0.21	0.00	−0.20	–	0.07	0.07	0.07	–
SK	−0.01	0.00	−0.01	–	0.90	0.90	0.90	–
BC	0.04	0.00	0.04	–	0.65	0.65	0.65	–

Notes: $n = 17,031$. GLM, Lasso, DB and SI in the columns 2-5 denote the survey GLM, survey Lasso, debiased Lasso and SI one-step estimates of $\theta_{0(j)}$, respectively. The columns 6-9 report the p-values of the survey GLM, DB, $C(\alpha)$ and SI tests for $\theta_{0(j)} = 0$, respectively. " – " means "not computed". n.s. and w.o.c.u. abbreviate "not stated" and "without children under".

Table 4: Point estimates of AME and test p-values for $H_0 : \text{AME}_j = 0$ in Lasso Logistic Regression for Government Online Service Access.

	Estimator			p-values			
Variable	GLM	DB	SI	t_{svy}	DB	$C(\alpha)$	SI
Female	0.02	−0.01	–	0.43	0.63	0.61	–
Aboriginal	0.19	0.02	–	0.11	0.46	0.45	–
Aboriginal n.s.	0.01	0.16	–	0.65	0.16	0.16	–
English	−0.22	0.06	–	0.01	0.46	0.44	–
French	0.08	−0.03	−0.13	0.29	0.70	0.78	1.00
English and French	0.08	0.10	–	0.00	0.23	0.22	–
Language n.s.	0.07	−0.03	–	0.49	0.80	0.79	–
Employed	0.11	0.08	0.08	0.27	0.00	0.00	0.00
Employment n.s.	−0.04	0.07	–	0.48	0.34	0.33	–
High school or less	−0.06	−0.11	−0.11	0.00	0.00	0.00	1.00
University degree	−0.01	0.08	0.08	0.61	0.00	0.00	0.00
Education n.s.	−0.04	−0.16	–	0.68	0.10	0.06	–
Visible minority	−0.12	0.06	0.06	0.00	0.00	0.00	0.00
Visible minority n.s.	0.20	0.08	–	0.29	0.42	0.41	–
Family household w.o.c.u 18	0.00	−0.06	−0.06	0.92	0.00	0.00	1.00
Single	−0.01	−0.15	−0.14	0.62	0.00	0.00	1.00
Other household type	0.00	−0.02	–	0.80	0.65	0.64	–
Family n.s.	−0.05	−0.04	–	0.01	0.50	0.48	–
$44,119 and less	−0.03	0.00	–	0.80	0.94	0.94	–
$75,322–$109,431	−0.05	−0.01	–	0.07	0.63	0.62	–
$109,432–$162,799	−0.07	0.00	–	0.01	0.81	0.80	–
$162,800 and higher	0.03	−0.05	–	0.20	0.01	0.01	–
Non-landed immigrant	0.00	0.00	–	0.86	0.88	0.87	–
Immigration n.s.	−0.05	0.21	–	0.05	0.31	0.31	–
NL	−0.04	0.03	–	0.06	0.21	0.20	–
PEI	−0.02	−0.01	–	0.62	0.79	0.78	–
NS	−0.01	−0.05	–	0.78	0.07	0.06	–
NB	−0.06	−0.07	–	0.02	0.01	0.00	–
QC	−0.16	−0.05	–	0.00	0.03	0.02	–
ON	0.00	−0.04	–	0.90	0.07	0.06	–
MB	0.08	−0.04	–	0.00	0.08	0.07	–
SK	0.06	0.00	–	0.00	0.90	0.90	–
BC	0.09	0.01	–	0.33	0.66	0.65	–

Notes: $n = 17,031$. GLM, DB and SI in the columns 2-4 denote the survey GLM, debiased Lasso and SI one-step estimates of AME_j. The columns 5-8 report the p-values of the survey GLM, DB, $C(\alpha)$ and SI tests for $\text{AME}_j = 0$, respectively. The $C(\alpha)$ test p-values are identical to those reported in Table 3 (Lemma 4.1). The p-values of SI2 were identical to those of SI, thus not shown. " − " means "not computed".

References

Belloni, A., Chernozhukov, V. and Wei, Y. (2016), 'Post-Selection Inference for Generalized Linear Models with Many Controls', *Journal of Business & Economic Statistics* **34**(4), 606–619.

Bühlmann, P. and van de Geer, S. (2011), *Statistics for High-Dimensional Data: Methods, Theory and Applications*, Springer Science & Business Media.

Cameron, A. C. and Trivedi, P. K. (2009), *Microeconometrics: Methods and Evaluations*, Cambridge University Press.

Dufour, J.-M., Trognon, A. and Tuvaandorj, P. (2016), Generalized $C(\alpha)$ Tests in Estimating Functions with Serial Dependence, *in* W. K. Li, D. Stanford and H. Yu, eds, 'Advances in Time Series Methods and Applications: the McLeod Festschrift', Springer, Berlin and New York, pp. 151–178.

Eaton, M. L. and Tyler, D. E. (1991), 'On Wielandt's Inequality and its Application to the Asymptotic Distribution of the Eigenvalues of a Random Symmetric Matrix', *The Annals of Statistics* **19**, 260–271.

Fuller, W. A. (2011), *Sampling Statistics*, John Wiley & Sons.

Hansen, B. (2022*a*), *Econometrics*, Princeton University Press.

Hansen, B. (2022*b*), *Probability and Statistics for Economists*, Princeton University Press.

Hastie, T., Tibshirani, R. and Wainwright, M. (2015), *Statistical Learning with Sparsity: The Lasso and Generalizations*, CRC press.

Javanmard, A. and Montanari, A. (2014), 'Confidence Intervals and Hypothesis Testing for High-Dimensional Regression', *The Journal of Machine Learning Research* **15**(1), 2869–2909.

Jin, C., Netrapalli, P., Ge, R., Kakade, S. M. and Jordan, M. I. (2019), 'A Short Note on Concentration Inequalities for Random Vectors with SubGaussian Norm', *arXiv preprint arXiv:1902.03736* .

Kuchibhotla, A. K., Kolassa, J. E. and Kuffner, T. A. (2022), 'Post-Selection Inference', *Annual Review of Statistics and Its Application* **9**, 505–527.

Lee, J. D., Sun, D. L., Sun, Y. and Taylor, J. E. (2016), 'Exact Post-Selection Inference, with Application to the Lasso', *The Annals of Statistics* **44**(3), 907–927.

Lehmann, E. L. and Romano, J. P. (2005), *Testing Statistical Hypotheses*, 3 edn, Springer.

Manski, C. F. and Lerman, S. R. (1977), 'The Estimation of Choice Probabilities from Choice Based Samples', *Econometrica: Journal of the Econometric Society* **45**, 1977–1988.

Markovic, J., Xia, L. and Taylor, J. (2017), 'Unifying Approach to Selective Inference with Applications to Cross-Validation', *arXiv preprint arXiv:1703.06559* .

McConville, K. S., Breidt, F. J., Lee, T. and Moisen, G. G. (2017), 'Model-Assisted Survey Regression Estimation with the Lasso', *Journal of Survey Statistics and Methodology* **5**(2), 131–158.

Negahban, S. N., Ravikumar, P., Wainwright, M. J. and Yu, B. (2012), 'A Unified Framework for High-Dimensional Analysis of M-Estimators with Decomposable Regularizers', *Statistical Science* **27**(4), 538 – 557.
URL: *https://doi.org/10.1214/12-STS400*

Newey, W. K. and McFadden, D. (1994), Large Sample Estimation and Hypothesis Testing, *in* R. F. Engle and D. L. McFadden, eds, 'Handbook of Econometrics, Volume 4', Amsterdam, chapter 36, pp. 2111–2245.

Neyman, J. (1959), Optimal Asymptotic Tests of Composite Statistical Hypotheses, *in* U. Grenander, ed., 'Probability and Statistics, the Harald Cramér Volume', Almqvist and Wiksell, Uppsala, Sweden, pp. 213–234.

Smith, R. J. (1987), 'Alternative Asymptotically Optimal Tests and their Application to Dynamic Specification', **LIV**, 665–680.

Taylor, J. and Tibshirani, R. (2018), 'Post-Selection Inference for l1-Penalized Likelihood Models', *Canadian Journal of Statistics* **46**(1), 41–61.

Thompson, S. K. (2012), *Sampling*, Vol. 755, John Wiley & Sons.

Tian, X. and Taylor, J. (2017), 'Asymptotics of Selective Inference', *Scandinavian Journal of Statistics* **44**(2), 480–499.

van de Geer, S. A. (2008), 'High-Dimensional Generalized Linear Models and the Lasso', *The Annals of Statistics* **36**(2), 614 – 645.
URL: *https://doi.org/10.1214/009053607000000929*

van de Geer, S., Bühlmann, P., Ritov, Y. and Dezeure, R. (2014), 'On Asymptotically Optimal Confidence Regions and Tests for High-Dimensional Models', *The Annals of Statistics* **42**(3), 1166 – 1202.
URL: *https://doi.org/10.1214/14-AOS1221*

Vershynin, R. (2010), 'Introduction to the Non-Asymptotic Analysis of Random Matrices', *arXiv preprint arXiv:1011.3027* .

Vershynin, R. (2018), *High-Dimensional Probability: An Introduction with Applications in Data Science*, Vol. 47, Cambridge University Press.

Wooldridge, J. M. (2001), 'Asymptotic Properties of Weighted M-Estimators for Standard Stratified Samples', *Econometric Theory* **17**(2), 451–470.

Wooldridge, J. M. (2010), *Econometric Analysis of Cross Section and Panel Data*, MIT press.

Xia, L., Nan, B. and Li, Y. (2021), 'Debiased Lasso for Generalized Linear Models with a Diverging Number of Covariates', *Biometrics* **forthcoming**.
URL: *https://onlinelibrary.wiley.com/doi/abs/10.1111/biom.13587*

Zhang, C.-H. and Zhang, S. S. (2014), 'Confidence Intervals for Low Dimensional Parameters in High Dimensional Linear Models', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **76**(1), 217–242.

A Proofs

A.1 Proposition 3.2

We will verify the assumptions of Algorithm 2 of Kuchibhotla et al. (2022). Let $\hat{\beta}_{Mj}$ and β_{M0j} denote the j-th elements of $\hat{\beta}_M$ and β_{M0}, respectively. Set in Assumptions (A1)–(A4) and Algorithm 2 of Kuchibhotla et al. (2022) that $\hat{\theta}_q = \hat{\beta}_{Mj}$, $\theta_q = \beta_{M0j}$, $D_{n,q} = AZ - b$, where A, Z and b are as defined in (3.12), and

$$\mu_{n,q} = An^{1/2}(\beta'_{M0}, 0'_{p+1-|M|})' - b = [n^{1/2}(-\mathrm{diag}(\mathtt{s}_M)\beta_{M0})', 0'_{2p+2-2|M|}]' - b.$$

Rewrite (2.7) as

$$\begin{aligned}\{\mathtt{s}_M = \mathrm{sign}(\hat{\beta}_M)\} &= \{\mathrm{diag}(\mathtt{s}_M)\hat{\beta}_M > 0\} \\ &= \{\mathrm{diag}(\mathtt{s}_M)(\tilde{\beta}_M - \hat{H}_M(\hat{\theta}_M)^{-1}(0, \lambda\mathtt{s}'_M)') > 0\} \\ &= \{-\mathrm{diag}(\mathtt{s}_M)\tilde{\beta}_M < -\lambda\,\mathrm{diag}(\mathtt{s}_M)\hat{H}_M(\hat{\theta}_M)^{-1}(0, \mathtt{s}'_M)'\}. \end{aligned} \tag{A.1}$$

The constraint (2.8) can be rewritten as

$$\begin{aligned}\{\|\mathtt{u}\|_\infty < 1\} &= \{\|\lambda^{-1}S_{-M}(\hat{\theta}_M)\|_\infty < 1\} \\ &= \{\|\lambda^{-1}\left(\tilde{S}_{-M}(\hat{\theta}_M) + \hat{H}_{-MM}(\hat{\theta}_M)\hat{H}_M(\hat{\theta}_M)^{-1}S_M(\hat{\theta}_M)\right)\|_\infty < 1\} \\ &= \{-\mathbf{1}_{p+1-|M|} \le \lambda^{-1}\left(\tilde{S}_{-M}(\hat{\theta}_M) + \hat{H}_{-MM}(\hat{\theta}_M)\hat{H}_M(\hat{\theta}_M)^{-1}S_M(\hat{\theta}_M)\right) \le \mathbf{1}_{p+1-|M|}\} \\ &= \{\tilde{S}_{-M}(\hat{\theta}_M) \le \lambda(\mathbf{1}_{p+1-|M|} - \hat{H}_{-MM}(\hat{\theta}_M)\hat{H}_M(\hat{\theta}_M)^{-1}(0, \mathtt{s}'_M)'), \\ &\quad -\tilde{S}_{-M}(\hat{\theta}_M) \le \lambda(\mathbf{1}_{p+1-|M|} + \hat{H}_{-MM}(\hat{\theta}_M)\hat{H}_M(\hat{\theta}_M)^{-1}(0, \mathtt{s}'_M)')\}, \end{aligned} \tag{A.2}$$

where the fourth equality uses (2.7). Thus, $\{\hat{M} = M, \mathrm{sign}(\hat{\beta}_M) = \mathtt{s}_M\} = \{AZ \le b\}$ and Assumption (A1) of Kuchibhotla et al. (2022) is satisfied.

Assumption (A2) therein is verified as follows. Consider the first $(|M|-1)\times 1$ nonzero subvector of $\mu_{n,q} = [n^{1/2}(-\mathrm{diag}(\mathtt{s}_M)\beta_{M0})', 0'_{2p+2-2|M|}]' - b$. Clearly, $-n^{1/2}\mathrm{diag}(\mathtt{s}_M)\beta_{M0} = -n^{1/2}|\beta_{M0}| \to -\infty$ as $n \to \infty$. Furthermore, $b = O_p(1)$ because $n^{-1/2}\lambda = C = O(1)$ and $\hat{H}_{-MM}(\hat{\theta}_M) - H_{-MM}(\theta_{M0}) \xrightarrow{p} 0$ and $\hat{H}_M(\hat{\theta}_M)^{-1} - H_M(\theta_{M0})^{-1} \xrightarrow{p} 0$ by and Assumption 1 and Lemma 3.5. Thus, the first $(|M|-1)\times 1$ nonzero subvector of $\mu_{n,q}$ diverges to $-\infty$

in probability. Combined with (A.6) shown below, by Slutsky's lemma (Corollary 11.2.3 and Problem 11.36 of Lehmann and Romano (2005)) we have

$$\liminf_{n\to\infty} \; P[D_{n,q} \le 0] = \liminf_{n\to\infty} \; P[D_{n,q} - \mu_{n,q} \le -\mu_{n,q}] > 0.$$

This verifies Assumption (A2) of Kuchibhotla et al. (2022).

Assumption (A3) of Kuchibhotla et al. (2022) holds as follows. From Lemma B.4 and the fact p is fixed, $n^{1/2}(\hat{\theta}_M - \theta_{M0}) = O_p(1)$. By the mean-value expansion,

$$S_M(\hat{\theta}_M) = S_M(\theta_{M0}) - \hat{H}_M(\theta_M^*)(\hat{\theta}_M - \theta_{M0}), \tag{A.3}$$

$$S_{-M}(\hat{\theta}_M) = S_{-M}(\theta_{M0}) - \hat{H}_{-MM}(\bar{\theta}_M)(\hat{\theta}_M - \theta_{M0}), \tag{A.4}$$

where θ_M^* and $\bar{\theta}_M$ are the mean-value between θ_{M0} and $\hat{\theta}_M$. Hence,

$$\begin{aligned} n^{1/2}(\tilde{\theta}_M - \theta_{M0}) &= n^{1/2}(\hat{\theta}_M - \theta_{M0}) - \hat{H}_M(\hat{\theta}_M)^{-1}\hat{H}_M(\theta_M^*)n^{1/2}(\hat{\theta}_M - \theta_{M0}) + n^{1/2}\hat{H}_M(\hat{\theta}_M)^{-1}S_M(\theta_{M0}), \\ &= n^{1/2}H_M(\theta_{M0})^{-1}S_M(\theta_{M0}) + o_p(1), \end{aligned}$$

where we used $\hat{H}_M(\hat{\theta}_M)^{-1}\hat{H}_M(\theta_M^*) \xrightarrow{p} I_{|M|}$ and $\hat{H}_M(\hat{\theta}_M)^{-1} - H_M(\theta_{M0})^{-1} \xrightarrow{p} 0$ which follow from Lemma 3.5, the convergence of the Hessian assumption, $n^{1/2}S_M(\theta_{M0}) = O_p(1)$ and the CMT. Moreover, using (A.3) and (A.4)

$$\begin{aligned} n^{1/2}\tilde{S}_{-M}(\hat{\theta}_M) &= n^{1/2}S_{-M}(\hat{\theta}_M) - \hat{H}_{-MM}(\hat{\theta}_M)\hat{H}_M(\hat{\theta}_M)^{-1}n^{1/2}S_M(\hat{\theta}_M) \\ &= n^{1/2}S_{-M}(\theta_{M0}) - \hat{H}_{-MM}(\bar{\theta}_M)(\hat{\theta}_M - \theta_{M0}) \\ &\quad - \hat{H}_{-MM}(\hat{\theta}_M)\hat{H}_M(\hat{\theta}_M)^{-1}n^{1/2}[S_M(\theta_{M0}) - \hat{H}_M(\theta_M^*)(\hat{\theta}_M - \theta_{M0})] \\ &= n^{1/2}S_{-M}(\theta_{M0}) - H_{-MM}(\theta_{M0})H_M(\theta_{M0})^{-1}n^{1/2}S_M(\theta_{M0}) + o_p(1) \\ &= n^{1/2}\tilde{S}_{-M}(\theta_{M0}) + o_p(1), \end{aligned}$$

where the second equality uses $\hat{H}_{-MM}(\hat{\theta}_M)^{-1} - H_{-MM}(\theta_{M0})^{-1} \xrightarrow{p} 0$ which follows from Lemma 3.5, and and the convergence of the Hessian above.. Therefore, by the Lyapunov's CLT applied to $[n^{1/2}S_M(\theta_{M0})', n^{1/2}\tilde{S}_{-M}(\theta_{M0})']'$ (see the proof of Lemma B.1) and Slutsky's

lemma

$$\begin{bmatrix} n^{1/2}(\tilde{\beta}_M - \beta_{M0}) \\ n^{1/2}\tilde{S}_{-M}(\hat{\theta}_M) \end{bmatrix} = \begin{bmatrix} [0_{|M|-1}, I_{|M|-1}]H_M(\theta_{M0})^{-1}n^{1/2}S_M(\theta_{M0}) \\ n^{1/2}\tilde{S}_M(\theta_{M0}) \end{bmatrix} + o_p(1) \xrightarrow{d} N(0, \Sigma). \tag{A.5}$$

Then, by Slutsky's lemma for $j = 1, \ldots, |M| - 1$

$$\begin{bmatrix} n^{1/2}e'_{j(|M|-1)}(\hat{\beta}_M - \beta_{M0}) \\ D_{n,q} - \mu_{n,q} \end{bmatrix} = \begin{pmatrix} e'_{jp} \\ A \end{pmatrix} \begin{bmatrix} n^{1/2}(\tilde{\beta}_M - \beta_{M0}) \\ n^{1/2}\tilde{S}_{-M}(\hat{\theta}_M) \end{bmatrix} \xrightarrow{d} N\left[0, \begin{pmatrix} e'_{jp} \\ A \end{pmatrix} \Sigma \begin{pmatrix} e'_{jp} \\ A \end{pmatrix}'\right]. \tag{A.6}$$

Assumption (A3) of Kuchibhotla et al. (2022) thus holds. Finally, by the CMT and Lemma 3.5

$$\hat{\Sigma} \xrightarrow{p} \Sigma. \tag{A.7}$$

This verifies Assumption (A4) of Kuchibhotla et al. (2022) and the result follows.

A.2 Lemma 3.4

We prove the result in 4 steps. In the first step, we show that $\bar{\Sigma}_n^{-1/2}A_i$ is sub-Gaussian. The second step reduces to the problem into bounding a sum of zero mean, independent sub-exponential random variables. The third step applies Bernstein's inequality to the average determined in the second step. Finally, the fourth step completes the proof.

Step 1: Sub-Gaussian norm bound for $\bar{\Sigma}_n^{-1/2}A_i$.
We first verify that $\bar{\Sigma}_n^{-1/2}A_i$ is sub-Gaussian. Because A_i is sub-Gaussian with $\|A_i\|_{\psi_2} \leq K$, by Remark 5.18 of Vershynin (2010) $\|A_i - \mu_i\|_{\psi_2} \leq 2K$. Hence, there exists an absolute constant $C > 0$ such that for all $t \in \mathbb{R}^p$

$$\mathrm{E}[\exp(t'(A_i - \mu_i))] \leq \exp(C\|t\|^2\|A_i - \mu_i\|_{\psi_2}) \leq \exp(2CK\|t\|^2), \tag{A.8}$$

see Section 5.2.3 of Vershynin (2010) and Jin et al. (2019). Hence,

$$
\begin{aligned}
\mathrm{E}[\exp(t'\bar{\Sigma}_n^{-1/2}(A_i-\mu_i))] &\le \exp(2CK\|\bar{\Sigma}_n^{-1/2}t\|^2) \\
&\le \exp(2CK\|t\|^2\lambda_{\max}(\bar{\Sigma}_n^{-1})) \\
&= \exp(2CK\|t\|^2/\lambda_{\min}(\bar{\Sigma}_n)) \\
&< \exp(2CK\|t\|^2/\lambda_l). \qquad \text{(A.9)}
\end{aligned}
$$

It follows that for some absolute constant $C_1 > 0$

$$
\|\bar{\Sigma}_n^{-1/2}(A_i-\mu_i)\|_{\psi_2} \le C_1 K \equiv K_1. \qquad \text{(A.10)}
$$

Let $S^{p-1} \equiv \{x \in \mathbb{R}^p, \|x\|^2 = 1\}$. Next we will bound

$$
\|\bar{\Sigma}_n^{-1/2}A_i\|_{\psi_2} \equiv \sup_{x\in S^{p-1}} \|A_i'\bar{\Sigma}_n^{-1/2}x\|_{\psi_2} = \sup_{x\in S^{p-1}} \sup_{m\ge 1} m^{-1/2}\left(\mathrm{E}[|A_i'\bar{\Sigma}_n^{-1/2}x|^m]\right)^{1/m}. \qquad \text{(A.11)}
$$

For $m \ge 1$, it holds that

$$
\begin{aligned}
\left(\mathrm{E}[|A_i'\bar{\Sigma}_n^{-1/2}x|^m]\right)^{1/m} &= \left(\mathrm{E}[|(A_i-\mu_i)'\bar{\Sigma}_n^{-1/2}x + \mu_i'\bar{\Sigma}_n^{-1/2}x|^m]\right)^{1/m} \\
&\le \left(\mathrm{E}[|(A_i-\mu_i)'\bar{\Sigma}_n^{-1/2}x|^m]\right)^{1/m} + \left(\mathrm{E}[|\mu_i'\bar{\Sigma}_n^{-1/2}x|^m]\right)^{1/m} \\
&\le \left(\mathrm{E}[|(A_i-\mu_i)'\bar{\Sigma}_n^{-1/2}x|^m]\right)^{1/m} + \mathrm{E}[|A_i'\bar{\Sigma}_n^{-1/2}x|]. \qquad \text{(A.12)}
\end{aligned}
$$

where the first inequality is by Minkowski's inequality and the second inequality is by Jensen's inequality on noting that $\left(\mathrm{E}[|\mu_i'\bar{\Sigma}_n^{-1/2}x|^m]\right)^{1/m} = |\mu_i'\bar{\Sigma}_n^{-1/2}x| = |\,\mathrm{E}[A_i'\bar{\Sigma}_n^{-1/2}x]|$. Consider the second term in (A.12). Since A_i is sub-Gaussian with $\sup_{x\in S^{p-1}} \frac{\sqrt{\mathrm{E}[(A_i'x)^2]}}{\sqrt{2}} \le K$,

$$
\lambda_{\max}(\mathrm{E}[A_iA_i']) = \sup_{x\in S^{p-1}} x'\,\mathrm{E}[A_iA_i']x = \sup_{x\in S^{p-1}} \mathrm{E}[(A_i'x)^2] \le (\sqrt{2}K)^2. \qquad \text{(A.13)}
$$

Then,

$$\begin{aligned}
\sup_{x\in S^{p-1}} \mathrm{E}[|A_i'\bar{\Sigma}_n^{-1/2}x|] &\leq \sup_{x\in S^{p-1}} \left(\mathrm{E}[|A_i'\bar{\Sigma}_n^{-1/2}x|^2]\right)^{1/2} \\
&\leq \sup_{x\in S^{p-1}} \left(\lambda_{\max}(\mathrm{E}[A_iA_i'])\lambda_{\max}(\bar{\Sigma}_n^{-1})\|x\|^2\right)^{1/2} \\
&= \sup_{x\in S^{p-1}} \left(\lambda_{\max}(\mathrm{E}[A_iA_i'])/\lambda_{\min}(\bar{\Sigma}_n)\right)^{1/2} \\
&\leq \sqrt{\frac{2}{\lambda_l}}K. \qquad \text{(A.14)}
\end{aligned}$$

where the first inequality is by Jensen's inequality, the second inequality is the extremal property of the maximum eigenvalue and the eigenvalue product inequality (see Hansen (2022*a*), Appendix B), and the third is by $\lambda_{\min}(\bar{\Sigma}_n) > \lambda_l$ and (A.13). Finally,

$$\begin{aligned}
\|\bar{\Sigma}_n^{-1/2}A_i\|_{\psi_2} &\leq \sup_{x\in S^{p-1}} \sup_{m\geq 1} m^{-1/2}\left[\left(\mathrm{E}[|(A_i-\mu_i)'\bar{\Sigma}_n^{-1/2}x|^m]\right)^{1/m} + \mathrm{E}[|A_i'\bar{\Sigma}_n^{-1/2}x|]\right] \\
&\leq K_1 + \sqrt{(2/\lambda_l)}K \sup_{m\geq 1} m^{-1/2} \leq C_1K + \sqrt{(2/\lambda_l)}K \equiv K_2, \qquad \text{(A.15)}
\end{aligned}$$

where the first inequality follows from (A.11) and (A.12), and the second inequality is by (A.10) and (A.14).

Step 2: Reduction to an average of sub-exponential random variables. Given K_2 defined in (A.15), let

$$\epsilon \equiv 8K_2^2 \max(\delta, \delta^2), \qquad \text{(A.16)}$$

Below, we will show that with probability at least $1 - 2\exp(-t^2)$

$$\|n^{-1}\bar{\Sigma}_n^{-1/2}A'A\bar{\Sigma}_n^{-1/2} - I_p\|_2 \leq \max(\delta, \delta^2) = \frac{\epsilon}{8K_2^2}, \qquad \text{(A.17)}$$

Let $\mathcal{N}$ denote the 1/4-net of S^{p-1}. Since $n^{-1}\bar{\Sigma}_n^{-1/2}A'A\bar{\Sigma}_n^{-1/2} - I_p = n^{-1}\sum_{i=1}^n \bar{\Sigma}_n^{-1/2}(A_iA_i' - \mathrm{E}[A_iA_i'])\bar{\Sigma}_n^{-1/2}$, by Lemma 5.4 of Vershynin (2010)

$$\begin{aligned}\|n^{-1}\bar{\Sigma}_n^{-1/2}A'A\bar{\Sigma}_n^{-1/2} - I_p\|_2 &\leq 2\max_{x\in\mathcal{N}} |n^{-1}\sum_{i=1}^n x'\bar{\Sigma}_n^{-1/2}(A_iA_i' - \mathrm{E}[A_iA_i'])\bar{\Sigma}_n^{-1/2}x| \\ &= 2\max_{x\in\mathcal{N}} |n^{-1}\sum_{i=1}^n (Z_i^2 - \mathrm{E}[Z_i^2])|, \end{aligned} \tag{A.18}$$

where $Z_i \equiv x'\bar{\Sigma}_n^{-1/2}A_i$. To show (A.17), for $\epsilon > 0$ defined in (A.17) we will upper bound the probability

$$P\left[\max_{x\in\mathcal{N}} |n^{-1}\sum_{i=1}^n (Z_i^2 - \mathrm{E}[Z_i^2])| \geq \epsilon/2\right].$$

Step 3: Concentration. Fix $x \in S^{n-1}$. It is clear that $\{Z_i^2 - \mathrm{E}[Z_i^2]\}_{i=1}^n$ are centered and independent. In addition, by Remark 5.18 and Lemma 5.14 of Vershynin (2010), $\{Z_i^2 - \mathrm{E}[Z_i^2]\}_{i=1}^n$ are sub-exponential random variables with $\|Z_i^2 - \mathrm{E}[Z_i^2]\|_{\psi_1} \leq 2\|Z_i^2\|_{\psi_1} \leq 4\|Z_i\|_{\psi_2}^2 \leq 4K_2^2$, where the last inequality is due to (A.15). By Bernstein's inequality (Corollary 5.17 of Vershynin (2010), Corollary 2.8.3 of Vershynin (2018)), for an absolute constant $c_1 > 0$

$$\begin{aligned} P\left[\left|n^{-1}\sum_{i=1}^n (Z_i^2 - \mathrm{E}[Z_i^2])\right| \geq \frac{\epsilon}{2}\right] &\leq 2\exp\left[-c_1 \min\left(\frac{\epsilon^2}{64K_2^4}, \frac{\epsilon}{8K_2^2}\right) n\right] \\ &= 2\exp(-c_1\delta^2 n) \\ &= 2\exp\left[-c_1c^2(\sqrt{p}+t)^2\right] \\ &\leq 2\exp(-c_1c^2(p+t^2)) \end{aligned} \tag{A.19}$$

where the first equality holds by the definition of ϵ in (3.34), the second equality is by the definition of δ, and the last inequality is due to the fact that $(a+b)^2 \geq a^2+b^2$ for $a, b \geq 0$.

Step 4: Union bound. By Corollary 4.2.13 of Vershynin (2018), there exists a 1/4-net $\mathcal{N}$ of S^{p-1} with cardinality $|\mathcal{N}| \leq 9p$. Taking the union bound and using (A.19) give

$$P\left[\max_{x\in\mathcal{N}}\left|n^{-1}\sum_{i=1}^n (Z_i^2 - \mathrm{E}[Z_i^2])\right| \geq \frac{\epsilon}{2}\right] \leq 9^p\, 2\exp\left[-c_1c^2(p+t^2)\right] \leq 2\exp(-t^2), \tag{A.20}$$

where the second inequality is by the choice $c = \sqrt{\frac{\log 9}{c_1}}$. Next we note that

$$
\begin{aligned}
&P[\|n^{-1}A'A - \bar{\Sigma}_n\|_2 < 8K_2^2 \max(\delta, \delta^2)\|\bar{\Sigma}_n\|_2] \\
&= P[\|n^{-1}A'A - \bar{\Sigma}_n\|_2 < \epsilon\|\bar{\Sigma}_n\|_2] \\
&= P[\|\bar{\Sigma}_n^{1/2}(n^{-1}\bar{\Sigma}_n^{-1/2}A'A\bar{\Sigma}_n^{-1/2} - I_p)\bar{\Sigma}_n^{1/2}\|_2 < \epsilon\|\bar{\Sigma}_n\|_2] \\
&\geq P[\|\bar{\Sigma}_n^{1/2}\|_2\|n^{-1}\bar{\Sigma}_n^{-1/2}A'A\bar{\Sigma}_n^{-1/2} - I_p\|_2\|\bar{\Sigma}_n^{1/2}\|_2 < \epsilon\|\bar{\Sigma}_n\|_2] \\
&= P\left[\|n^{-1}\bar{\Sigma}_n^{-1/2}A'A\bar{\Sigma}_n^{-1/2} - I_p\|_2 < \epsilon\right] \\
&\geq P\left[\max_{x\in\mathcal{N}}\left|n^{-1}\sum_{i=1}^{n}(Z_i^2 - \mathrm{E}[Z_i^2])\right| < \frac{\epsilon}{2}\right] \\
&\geq 1 - 2\exp(-t^2),
\end{aligned}
$$

where the first two equalities hold trivially, the first inequality is by the Cauchy-Schwarz inequality, the second equality holds by the definition of the spectral norm and $\bar{\Sigma}_n$ is symmetric, the second inequality holds by (A.18) and the last is by (A.18). This completes the proof.

A.3 Lemma 3.5

By the triangle inequality for spectral norm,

$$
\|\hat{H}(\hat{\theta}) - H(\theta_0)\|_2 \leq \|\hat{H}(\hat{\theta}) - \hat{H}(\theta_0)\|_2 + \|\hat{H}(\theta_0) - H(\theta_0)\|_2. \tag{A.21}
$$

Let $A_i = x_i\sqrt{w_i\ddot{g}(y_i, x_i'\theta_0)}$ and $t = s\sqrt{p}$ in Lemma 3.4. Since x_i is sub-Gaussian and $\sqrt{w_i\ddot{g}(y_i, x_i'\theta_0)} \leq C_u$ a.s. by Assumption 1(a) and (c) (the condition (3.4)), using Assumption 1(a) once again

$$
\|A_i\|_{\psi_2} = \sup_{\|b\|=1}\sup_{m\geq 1} m^{-1/2}\left(\mathrm{E}\left[\left|x_i'b\sqrt{w_i\ddot{g}(y_i, x_i'\theta_0)}\right|^m\right]\right)^{1/m} \leq C_u^2. \tag{A.22}
$$

Then, $n^{-1}A'A = n^{-1}\sum_{i=1}^{n} A_i A_i' = \hat{H}(\theta_0)$, $\bar{\Sigma}_n = H(\theta_0)$ and $\delta = c\left(\sqrt{\frac{p}{n}} + \frac{t}{\sqrt{n}}\right) = (s+1)c\sqrt{\frac{p}{n}}$, and Lemma 3.4 gives

$$P\left[\|n^{-1}A'A - \bar{\Sigma}_n\|_2 < 8K_2^2(s+1)c\sqrt{\frac{p}{n}}\lambda_u\right] \geq 1 - 2\exp(-s^2 p).$$

Therefore, $\|n^{-1}A'A - \bar{\Sigma}_n\|_2 = O_p\left(\sqrt{\frac{p}{n}}\right)$ or equivalently

$$\|\hat{H}(\theta_0) - H(\theta_0)\|_2 = O_p\left(\sqrt{\frac{p}{n}}\right). \tag{A.23}$$

By Lemma 3.4 and Assumption 1(b), $\|n^{-1}X'X - \mathrm{E}[n^{-1}X'X]\|_2 \xrightarrow{p} 0$ and

$$\lambda_{\max}(n^{-1}X'X) - \lambda_{\max}(\mathrm{E}[n^{-1}X'X]) \leq \|n^{-1}X'X - \mathrm{E}[n^{-1}X'X]\|_2 \xrightarrow{p} 0. \tag{A.24}$$

Hence

$$\lambda_{\max}(n^{-1}X'X) = O_p(1), \tag{A.25}$$

Furthermore, letting $W(\theta) \equiv -\mathrm{diag}(w_1\ddot{g}(y_1, x_1'\theta), \ldots, w_n\ddot{g}(y_n, x_n'\theta))$

$$\begin{aligned}
\|\hat{H}(\hat{\theta}) - \hat{H}(\theta_0)\|_2 &= \|n^{-1}X'(W(\hat{\theta}) - W(\theta_0))X\|_2 \\
&\leq n^{-1}\|X'\|_2\|X\|_2\|W(\hat{\theta}) - W(\theta_0)\|_2 \\
&= \lambda_{\max}(n^{-1}X'X)\|W(\hat{\theta}) - W(\theta_0)\|_2 \\
&\leq \lambda_{\max}(n^{-1}X'X)\max_i |w_i||\ddot{g}(y_i, x_i'\hat{\theta}) - \ddot{g}(y_i, x_i'\theta_0)| \\
&\leq \lambda_{\max}(n^{-1}X'X)C_u L_0 \max_i |x_i(\hat{\theta} - \theta_0)| \\
&\leq \lambda_{\max}(n^{-1}X'X)C_u^2 L_0\|\hat{\theta} - \theta_0)\|_1 \\
&= O_p(m_0\lambda),
\end{aligned} \tag{A.26}$$

where the last equality uses Lemma B.4. Thus, combining (A.23) and (A.26) with (A.21) gives

$$\|\hat{H}(\hat{\theta}) - H(\theta_0)\|_2 = O_p\left(\sqrt{\frac{p}{n}} + m_0\lambda\right). \tag{A.27}$$

To show (3.36), note first that by Cauchy-Schwarz inequality for spectral norm (Hansen, 2022*a*)

$$\begin{aligned}\|\hat{H}(\hat{\theta})^{-1} - H(\theta_0)^{-1}\|_2 &= \|\hat{H}(\hat{\theta})^{-1}(\hat{H}(\hat{\theta}) - H(\theta_0))H(\theta_0)^{-1}\|_2 \\ &\leq \|\hat{H}(\hat{\theta})^{-1}\|_2 \|\hat{H}(\hat{\theta}) - H(\theta_0)\|_2 \|H(\theta_0)^{-1}\|_2. \end{aligned} \tag{A.28}$$

For the third term on the right-hand side of (A.28), by Assumption 1(b)

$$\|H(\theta_0)^{-1}\|_2 = 1/\lambda_{\max}(H(\theta_0)) = O(1). \tag{A.29}$$

Finally consider the third factor in (A.28). By Weyl's inequality (see Eaton and Tyler (1991), Lemma 2.1), $\lambda_{\min}(\hat{H}(\hat{\theta}) - H(\theta_0)) \leq \lambda_{\min}(\hat{H}(\hat{\theta})) - \lambda_{\min}(H(\theta_0)) \leq \lambda_{\max}(\hat{H}(\hat{\theta}) - H(\theta_0))$. Combining this with the fact that

$$\|\hat{H}(\hat{\theta}) - H(\theta_0)\|_2 = \max\{-\lambda_{\min}(\hat{H}(\hat{\theta}) - H(\theta_0)), \lambda_{\max}(\hat{H}(\hat{\theta}) - H(\theta_0))\},$$

we obtain

$$|\lambda_{\min}(\hat{H}(\hat{\theta})) - \lambda_{\min}(H(\theta_0))| \leq \|\hat{H}(\hat{\theta}) - H(\theta_0)\|_2. \tag{A.30}$$

Fix $0 < \epsilon < \lambda_l$. Since $\|(\hat{H}(\hat{\theta}))^{-1}\|_2 = 1/\lambda_{\min}(\hat{H}(\hat{\theta}))$, using (A.30)

$$\begin{aligned} P\left[\|\hat{H}(\hat{\theta})^{-1}\|_2 \geq \frac{1}{\lambda_{\min}(H(\theta_0)) - \epsilon}\right] &= P\left[\frac{1}{\lambda_{\min}(\hat{H}(\hat{\theta}))} \geq \frac{1}{\lambda_{\min}(H(\theta_0)) - \epsilon}\right] \\ &= P\left[\lambda_{\min}(H(\theta_0)) - \lambda_{\min}(\hat{H}(\hat{\theta})) \geq \epsilon\right] \\ &\leq P\left[|\lambda_{\min}(H(\theta_0)) - \lambda_{\min}(\hat{H}(\hat{\theta}))| \geq \epsilon\right] \\ &\leq P\left[\|H(\theta_0) - \hat{H}(\hat{\theta})\|_2 \geq \epsilon\right] \\ &\to 0, \end{aligned}$$

where the last line follows from (A.27). Thus,

$$\|\hat{H}(\hat{\theta})^{-1}\|_2 = O_p(1). \tag{A.31}$$

Combining (A.27), (A.29) and (A.31) in (A.28), we obtain (3.36). The convergence results in (3.37) and (3.38) follow similarly by setting $A_i = x_i w_i \dot{g}(y_i, x_i'\theta_0)$ in Lemma 3.4 and repeating the argument above.

A.4 Proposition 3.6

By the mean value expansion,

$$S(\theta_0) = S(\hat{\theta}) + \hat{H}(\theta^*)(\hat{\theta} - \theta_0) = S(\hat{\theta}) + \hat{H}(\hat{\theta})(\hat{\theta} - \theta_0) + R, \tag{A.32}$$

where θ^* is the mean-value between $\hat{\theta}$ and θ_0, and $R = [R_1, \ldots, R_{p+1}]'$ with

$$R_j \equiv n^{-1} \sum_{i=1}^{n} (\ddot{g}(y_i, x_i'\theta^*) - \ddot{g}(y_i, x_i'\hat{\theta})) w_i x_{ij} x_i'(\theta_0 - \hat{\theta}). \tag{A.33}$$

Note that since $\dot{\rho}(\theta)$ is locally Lipschitz in a neighborhood of θ_0, with probability approaching 1 $\|\dot{\rho}(\bar{\theta}) - \dot{\rho}(\hat{\theta})\| \leq B_0 \|\bar{\theta} - \hat{\theta}\|$ for some $B_0 = O(1)$. Also, since

$$n^{1/2}(\rho(\hat{\theta}) - \rho(\theta_0)) = \dot{\rho}(\bar{\theta})' n^{1/2}(\hat{\theta} - \theta_0), \tag{A.34}$$

where $\bar{\theta}$ is a mean-value between $\hat{\theta}$ and θ_0, we have

$$\begin{aligned} n^{1/2}\|\dot{\rho}(\hat{\theta}) - \dot{\rho}(\bar{\theta})\| \|\hat{\theta} - \theta_0\| &= n^{1/2} B_0 \|\hat{\theta} - \bar{\theta}\| \|\hat{\theta} - \theta_0\| = O_p(n^{1/2} m_0 \lambda^2) \\ &= o_p(1), \end{aligned} \tag{A.35}$$

where the last line is by $n^{1/2} m_0 \lambda^2 = n^{-1/2} m_0 C^2 \log p \leq C^2 m_0 (p/n)^{1/2} \log p = o(1)$. Then,

$$\begin{aligned} & n^{1/2}(\tilde{\rho} - \rho(\theta_0)) \\ &= n^{1/2}(\rho(\hat{\theta}) - \rho(\theta_0)) + \dot{\rho}(\hat{\theta})' \hat{H}(\hat{\theta})^{-1} n^{1/2} S(\hat{\theta}) \\ &= n^{1/2} \dot{\rho}(\bar{\theta})'(\hat{\theta} - \theta_0) + n^{1/2} \dot{\rho}(\hat{\theta})' \hat{H}(\hat{\theta})^{-1} S(\theta_0) - n^{1/2} \dot{\rho}(\hat{\theta})'(\hat{\theta} - \theta_0) - n^{1/2} \dot{\rho}(\hat{\theta})' \hat{H}(\hat{\theta})^{-1} R \\ &= n^{1/2} \dot{\rho}(\hat{\theta})' \hat{H}(\hat{\theta})^{-1} S(\theta_0) - n^{1/2} \dot{\rho}(\hat{\theta})' \hat{H}(\hat{\theta})^{-1} R + o_p(1), \end{aligned}$$

where the first equality is by the definition of $\tilde{\rho}$, the second equality is by (A.32) and (A.34), and the third is by (A.35). Below, the proof will be completed in three steps: the first two steps establish

$$\begin{aligned}
&\dot{\rho}(\hat{\theta})'\hat{H}(\hat{\theta})^{-1}\hat{I}(\hat{\theta})\hat{H}(\hat{\theta})^{-1}\dot{\rho}(\hat{\theta}) - \dot{\rho}(\theta_0)'H(\theta_0)^{-1}I(\theta_0)H(\theta_0)^{-1}\dot{\rho}(\theta_0) = o_p(1), \\
&n^{1/2}\dot{\rho}(\hat{\theta})'\hat{H}(\hat{\theta})^{-1}S(\theta_0) - n^{1/2}\dot{\rho}(\theta_0)'H(\theta_0)^{-1}S(\theta_0) = o_p(1),
\end{aligned} \tag{A.36}$$

and the third step verifies $n^{1/2}\dot{\rho}(\hat{\theta})'\hat{H}(\hat{\theta})^{-1}R = o_p(1)$. It will then follow that

$$\begin{aligned}
&\left[\dot{\rho}(\hat{\theta})'\hat{H}(\hat{\theta})^{-1}\hat{I}(\hat{\theta})\hat{H}(\hat{\theta})^{-1}\dot{\rho}(\hat{\theta})\right]^{-1/2} n^{1/2}(\tilde{\rho} - \rho(\theta_0)) \\
&= \left[\dot{\rho}(\hat{\theta})'\hat{H}(\hat{\theta})^{-1}\hat{I}(\hat{\theta})\hat{H}(\hat{\theta})^{-1}\dot{\rho}(\hat{\theta})\right]^{-1/2} \left[n^{1/2}\dot{\rho}(\hat{\theta})'\hat{H}(\hat{\theta})^{-1}S(\theta_0) - n^{1/2}\dot{\rho}(\hat{\theta})'\hat{H}(\hat{\theta})^{-1}R + o_p(1)\right] \\
&= \left[\dot{\rho}(\theta_0)'H(\theta_0)^{-1}I(\theta_0)H(\theta_0)^{-1}\dot{\rho}(\theta_0)\right]^{-1/2} n^{1/2}\dot{\rho}(\theta_0)'H(\theta_0)^{-1}S(\theta_0) + o_p(1).
\end{aligned}$$

Finally, applying Lemma B.1 and Slutsky's lemma give the desired result.

Step 1: $\dot{\rho}(\hat{\theta})'\hat{H}(\hat{\theta})^{-1}\hat{I}(\hat{\theta})\hat{H}(\hat{\theta})^{-1}\dot{\rho}(\hat{\theta}) - \dot{\rho}(\theta_0)'H(\theta_0)^{-1}I(\theta_0)H(\theta_0)^{-1}\dot{\rho}(\theta_0) = o_p(1)$**.**
First, by the triangle inequality

$$\begin{aligned}
&\|\dot{\rho}(\hat{\theta})'\hat{H}(\hat{\theta})^{-1}\hat{I}(\hat{\theta})\hat{H}(\hat{\theta})^{-1}\dot{\rho}(\hat{\theta}) - \dot{\rho}(\theta_0)'H(\theta_0)^{-1}I(\theta_0)H(\theta_0)^{-1}\dot{\rho}(\theta_0)\|_2 \\
&\leq \|\dot{\rho}(\hat{\theta})'\left[\hat{H}(\hat{\theta})^{-1}\hat{I}(\hat{\theta})\hat{H}(\hat{\theta})^{-1} - H(\theta_0)^{-1}I(\theta_0)H(\theta_0)^{-1}\right]\dot{\rho}(\hat{\theta})\|_2 \\
&\quad + \|\dot{\rho}(\hat{\theta})'H(\theta_0)^{-1}I(\theta_0)H(\theta_0)^{-1}(\dot{\rho}(\hat{\theta}) - \dot{\rho}(\theta_0))\|_2 \\
&\quad + \|(\dot{\rho}(\hat{\theta}) - \dot{\rho}(\theta_0))'H(\theta_0)^{-1}I(\theta_0)H(\theta_0)^{-1}\dot{\rho}(\theta_0)\|_2.
\end{aligned} \tag{A.37}$$

Consider the first term on the right-hand side of (A.37). By Cauchy-Schwarz inequality,

$$\begin{aligned}
&\|\dot{\rho}(\hat{\theta})'\left[\hat{H}(\hat{\theta})^{-1}\hat{I}(\hat{\theta})\hat{H}(\hat{\theta})^{-1} - H(\theta_0)^{-1}I(\theta_0)H(\theta_0)^{-1}\right]\dot{\rho}(\hat{\theta})\|_2 \\
&\leq \|\hat{H}(\hat{\theta})^{-1}\hat{I}(\hat{\theta})\hat{H}(\hat{\theta})^{-1} - H(\theta_0)^{-1}I(\theta_0)H(\theta_0)^{-1}\|_2\|\dot{\rho}(\hat{\theta})\|_2^2,
\end{aligned} \tag{A.38}$$

After rearranging and using the triangle and Cauchy-Schwarz inequalities

$$
\begin{aligned}
&\|\hat{H}(\hat{\theta})^{-1}\hat{I}(\hat{\theta})\hat{H}(\hat{\theta})^{-1} - H(\theta_0)^{-1}I(\theta_0)H(\theta_0)^{-1}\|_2 \\
&= \|(\hat{H}(\hat{\theta})^{-1} - H(\theta_0)^{-1})\hat{I}(\hat{\theta})\hat{H}(\hat{\theta})^{-1} + H(\theta_0)^{-1}(\hat{I}(\hat{\theta})\hat{H}(\hat{\theta})^{-1} - I(\theta_0)H(\theta_0)^{-1})\|_2, \\
&\leq \|\hat{H}(\hat{\theta})^{-1} - H(\theta_0)^{-1}\|_2\|\hat{I}(\hat{\theta})\|_2\|\hat{H}(\hat{\theta})^{-1}\|_2 + \|H(\theta_0)^{-1}\|_2\|\hat{I}(\hat{\theta})\hat{H}(\hat{\theta})^{-1} - I(\theta_0)H(\theta_0)^{-1}\|_2.
\end{aligned} \tag{A.39}
$$

For the first summand of (A.39), by Lemma 3.5

$$
\|\hat{H}(\hat{\theta})^{-1} - H(\theta_0)^{-1}\|_2\|\hat{I}(\hat{\theta})\|_2\|\hat{H}(\hat{\theta})^{-1}\|_2 = o_p(1). \tag{A.40}
$$

For the second factor in the second summand of (A.39), using the triangle and Cauchy-Schwarz inequalities

$$
\begin{aligned}
&\|\hat{I}(\hat{\theta})\hat{H}(\hat{\theta})^{-1} - I(\theta_0)H(\theta_0)^{-1}\|_2 \\
&= \|(\hat{I}(\hat{\theta}) - I(\theta_0))(\hat{H}(\hat{\theta})^{-1} - H(\theta_0)^{-1}) + (\hat{I}(\theta_0) - I(\theta_0))H(\theta_0)^{-1} + I(\theta_0)(\hat{H}(\theta_0)^{-1} - H(\theta_0)^{-1})\|_2 \\
&\leq \|\hat{I}(\hat{\theta}) - I(\theta_0)\|_2\|\hat{H}(\hat{\theta})^{-1} - H(\theta_0)^{-1}\|_2 + \|\hat{I}(\theta_0) - I(\theta_0)\|_2\|H(\theta_0)^{-1}\|_2 \\
&\quad + \|I(\theta_0)\|_2\|\hat{H}(\theta_0)^{-1} - H(\theta_0)^{-1}\|_2 \\
&\xrightarrow{p} 0,
\end{aligned} \tag{A.41}
$$

where the last line is by Lemma 3.5 and the CMT. From Lemma B.4, $\|\hat{\theta} - \theta_0\| = O_p(m_0^{1/2}\lambda) = O_p\left(\left(\frac{m_0 \log p}{n}\right)^{1/2}\right) = o_p(1)$. Since $\dot{\rho}(\theta)$ is locally Lipschitz in a neighborhood of θ_0, with probability approaching 1, we have for $B_0 = O(1)$ $\|\dot{\rho}(\hat{\theta}) - \dot{\rho}(\theta_0)\| \leq B_0\|\hat{\theta} - \theta_0\|$. Thus,

$$
\|\dot{\rho}(\hat{\theta}) - \dot{\rho}(\theta_0)\|_2 \leq r^{1/2}\|\dot{\rho}(\hat{\theta}) - \dot{\rho}(\theta_0)\| = O_p\left(\left(\frac{m_0 \log p}{n}\right)^{1/2}\right). \tag{A.42}
$$

By the triangle inequality and (A.42)

$$
\|\dot{\rho}(\hat{\theta})\|_2 \leq \|\dot{\rho}(\hat{\theta}) - \dot{\rho}(\theta_0)\|_2 + \|\dot{\rho}(\theta_0)\|_2 = O_p(1). \tag{A.43}
$$

Therefore, the quantity in (A.38) is $o_p(1)$. Consider the second term on the right-hand side of (A.37). By the triangle inequality and (A.42),

$$\begin{aligned}
&\|\dot{\rho}(\hat{\theta})' H(\theta_0)^{-1} I(\theta_0) H(\theta_0)^{-1} (\dot{\rho}(\hat{\theta}) - \dot{\rho}(\theta_0))\|_2 \\
&\leq \|\dot{\rho}(\hat{\theta})\|_2 \|H(\theta_0)^{-1} I(\theta_0) H(\theta_0)^{-1}\|_2 \|\dot{\rho}(\hat{\theta}) - \dot{\rho}(\theta_0)\|_2 \\
&\xrightarrow{p} 0.
\end{aligned}$$

Similarly, for the third term on the right-hand side of (A.37)

$$\begin{aligned}
&\|(\dot{\rho}(\hat{\theta}) - \dot{\rho}(\theta_0))' H(\theta_0)^{-1} I(\theta_0) H(\theta_0)^{-1} \dot{\rho}(\theta_0)\|_2 \\
&\leq \|\dot{\rho}(\hat{\theta}) - \dot{\rho}(\theta_0)\|_2 \|H(\theta_0)^{-1} I(\theta_0) H(\theta_0)^{-1}\|_2 \|\dot{\rho}(\theta_0)\|_2 \\
&\xrightarrow{p} 0.
\end{aligned}$$

Step 2: $n^{1/2}\dot{\rho}(\hat{\theta})'\hat{H}(\hat{\theta})^{-1}S(\theta_0) - n^{1/2}\dot{\rho}(\theta_0)'H(\theta_0)^{-1}S(\theta_0) = o_p(1)$**.**
Remark that from Assumption 1, $|\dot{g}(y_i, x_i'\theta_0)| \leq C_u$, $|w_i| \leq C_u$ and $\|x_i\|^2 \leq (p+1)C_u^2$ a.s. for all i. Using the independence assumption,

$$\mathrm{E}[\|S(\theta_0)\|_2^2] = \mathrm{E}[\|S(\theta_0)\|^2] = n^{-2}\,\mathrm{E}\left[\sum_{i=1}^{n} w_i^2 \|x_i\|^2 \dot{g}(y_i, x_i'\theta_0)^2\right] \leq n^{-1}(p+1)C_u^6.$$

By Markov's inequality,

$$\|S(\theta_0)\|_2 = O_p\left(\sqrt{\frac{p}{n}}\right). \tag{A.44}$$

Now rewrite

$$\begin{aligned}
&n^{1/2}\dot{\rho}(\hat{\theta})'\hat{H}(\hat{\theta})^{-1}S(\theta_0) - n^{1/2}\dot{\rho}(\theta_0)'H(\theta_0)^{-1}S(\theta_0) \\
&= n^{1/2}(\dot{\rho}(\hat{\theta}) - \dot{\rho}(\theta_0))'\hat{H}(\hat{\theta})^{-1}S(\theta_0) + n^{1/2}\left(\dot{\rho}(\theta_0)'\hat{H}(\hat{\theta})^{-1}S(\theta_0) - \dot{\rho}(\theta_0)'H(\theta_0)^{-1}S(\theta_0)\right).
\end{aligned} \tag{A.45}$$

For the first term of (A.45),

$$\begin{aligned}
\|n^{1/2}(\dot{\rho}(\hat{\theta}) - \dot{\rho}(\theta_0))'\hat{H}(\hat{\theta})^{-1}S(\theta_0)\|_2 &\leq n^{1/2}\|\dot{\rho}(\hat{\theta}) - \dot{\rho}(\theta_0)\|_2\|\hat{H}(\hat{\theta})^{-1}\|_2\|S(\theta_0)\|_2 \\
&= n^{1/2}O_p\left(\sqrt{\frac{m_0 \log p}{n}}\right) O_p(1)\, O_p\left(\sqrt{\frac{p}{n}}\right) \\
&= O_p\left(\sqrt{\frac{p\, m_0 \log p}{n}}\right) \\
&= o_p(1), \qquad \text{(A.46)}
\end{aligned}$$

where the first inequality is by Cauchy-Schwarz, the first equality uses (A.31), (A.42) and (A.44), and the last equality holds because $m_0(\log p)p/n \leq m_0(\log p)(p/n)^{1/2}(p^2/n)^{1/2} \to 0$ by the assumption of the proposition. For the second term of (A.45), we have

$$\begin{aligned}
n^{1/2}\|\dot{\rho}(\theta_0)'\hat{H}(\hat{\theta})^{-1}S(\theta_0) - \dot{\rho}(\theta_0)'H(\theta_0)^{-1}S(\theta_0)\|_2 &\leq n^{1/2}\|\dot{\rho}(\theta_0)\|_2\|\hat{H}(\hat{\theta})^{-1} - H(\theta_0)^{-1}\|_2\|S(\theta_0)\|_2 \\
&= n^{1/2}O_p\left(\sqrt{\frac{p}{n}} + m_0\lambda\right) O_p\left(\sqrt{\frac{p}{n}}\right) \\
&= O_p\left(\sqrt{\frac{p^2}{n}} + \sqrt{p}\, m_0\lambda\right) \\
&= o_p(1), \qquad \text{(A.47)}
\end{aligned}$$

where the first inequality is by Cauchy-Schwarz, the first equality is by Lemma 3.5 and (A.44), and the last equality holds because $p^2/n \to 0$ and $p^{1/2}\, m_0\lambda = Cm_0(p/n)^{1/2}(\log p)^{1/2} \leq Cm_0(p/n)^{1/2}2\log p \to 0$ by the assumption of the proposition. It follows from (A.45), (A.46) and (A.47) that

$$n^{1/2}\dot{\rho}(\hat{\theta})'\hat{H}(\hat{\theta})^{-1}S(\theta_0) - n^{1/2}\dot{\rho}(\theta_0)'H(\theta_0)^{-1}S(\theta_0) = o_p(1).$$

Step 3: $n^{1/2}\dot{\rho}(\hat{\theta})'\hat{H}(\hat{\theta})^{-1}R = o_p(1)$**.**

By Cauchy-Schwarz, $n^{1/2}\|\dot{\rho}(\hat{\theta})'\hat{H}(\hat{\theta})^{-1}R\|_2 \leq n^{1/2}\|\dot{\rho}(\hat{\theta})\|_2\|\hat{H}(\hat{\theta})^{-1}R\|_2$. Remark from (A.43)

that $\|\dot{\rho}(\hat{\theta})\|_2 = O_p(1)$. To show $n^{1/2}\|\hat{H}(\hat{\theta})^{-1}R\|_2 = o_p(1)$, note that

$$\begin{aligned}
\max_{1\le j\le p+1} |R_j| &\le n^{-1}\sum_{i=1}^{n} |\ddot{g}(y_i, x_i'\theta^*) - \ddot{g}(y_i, x_i'\hat{\theta})||w_i| \max_{1\le j\le p+1} |x_{ij}||x_i'(\theta_0 - \hat{\theta})| \\
&\le n^{-1}\sum_{i=1}^{n} L_0|x_i(\theta^* - \hat{\theta})|C_u^2|x_i'(\theta_0 - \hat{\theta})| \\
&\le L_0 C_u^2 n^{-1}\sum_{i=1}^{n} |x_i'(\theta_0 - \hat{\theta})|^2 \\
&= L_0 C_u^2 O_p(m_0\lambda^2) \\
&= O_p(m_0\lambda^2),
\end{aligned} \tag{A.48}$$

where the first inequality is by Assumption 1(c), and the first equality uses Lemma B.4. Since $\|H(\theta_0)\| = O(1)$ and $\|\hat{H}(\hat{\theta}) - H(\theta_0)\| = o_p(1)$, $\|\hat{H}(\hat{\theta})\| = O_p(1)$. Therefore,

$$\begin{aligned}
n^{1/2}\|\hat{H}(\hat{\theta})^{-1}R\|_2 &\le n^{1/2}\|\hat{H}(\hat{\theta})^{-1}\|_2\|R\|_2 \\
&\le n^{1/2}\hat{H}(\hat{\theta})^{-1}(p+1)^{1/2}\|R\|_\infty \\
&= O_p((n(p+1))^{1/2}m_0\lambda^2) \\
&= o_p(1),
\end{aligned} \tag{A.49}$$

where the first equality holds by using (A.48) and the second equality follows on noting that $(n(p+1))^{1/2}m_0\lambda^2 = (n(p+1))^{1/2}m_0C^2(\log p)/n \le (2p/n)^{1/2}m_0C^2\log p = o(1)$.

A.5 Proposition 3.7

Similarly to (A.32), by the mean value expansion

$$S(\theta_0) = S(\tilde{\theta}^*) + \hat{H}(\theta^*)(\tilde{\theta}^* - \theta_0) = S(\tilde{\theta}^*) + \hat{H}(\tilde{\theta}^*)(\tilde{\theta}^* - \theta_0) + R^*, \tag{A.50}$$

where θ^* is a mean-value between $\tilde{\theta}^*$ and θ_0, and $R^* = [R_1^*, \ldots, R_{p+1}^*]'$ with

$$R_j^* \equiv n^{-1}\sum_{i=1}^{n}(\ddot{g}(y_i, x_i'\theta^*) - \ddot{g}(y_i, x_i'\tilde{\theta}^*))w_i x_{ij} x_i'(\theta_0 - \tilde{\theta}^*).$$

Proceeding similarly to Steps 1, 2 and 3 in the proof of Proposition 3.6, we obtain

$$\left(\dot{\rho}(\tilde{\theta}^*)'\hat{H}(\tilde{\theta}^*)^{-1}\hat{I}(\tilde{\theta}^*)\hat{H}(\tilde{\theta}^*)^{-1}\dot{\rho}(\tilde{\theta}^*)\right)^{-1/2} - \left(\dot{\rho}(\theta_0)'H(\theta_0)^{-1}I(\theta_0)H(\theta_0)^{-1}\dot{\rho}(\theta_0)\right)^{-1/2} = o_p(1), \tag{A.51}$$

$$n^{1/2}\dot{\rho}(\tilde{\theta}^*)'\hat{H}(\tilde{\theta}^*)^{-1}S(\theta_0) = n^{1/2}\dot{\rho}(\theta_0)'H(\theta_0)^{-1}S(\theta_0) + o_p(1), \tag{A.52}$$

$$n^{1/2}\dot{\rho}(\tilde{\theta}^*)'\hat{H}(\tilde{\theta}^*)^{-1}R^* = o_p(1). \tag{A.53}$$

By the assumption that $\rho(\tilde{\theta}^*) = \rho(\theta_0)$ and the mean value expansion

$$0 = n^{1/2}(\rho(\tilde{\theta}^*) - \rho(\theta_0)) = \dot{\rho}(\bar{\theta})'n^{1/2}(\tilde{\theta}^* - \theta_0), \tag{A.54}$$

where $\bar{\theta}$ is a mean-value between $\tilde{\theta}^*$ and θ_0. Next, we will show that $\dot{\rho}(\tilde{\theta}^*)'n^{1/2}(\tilde{\theta}^* - \theta_0) = o_p(1)$. Since $\dot{\rho}(\theta)$ is locally Lipschitz in a neighborhood of θ_0, with probability approaching 1 $\|\dot{\rho}(\bar{\theta}) - \dot{\rho}(\tilde{\theta}^*)\| \leq B_0\|\bar{\theta} - \tilde{\theta}^*\|$ for some $B_0 = O(1)$. Thus, using (A.54)

$$\begin{aligned}
\|n^{1/2}\dot{\rho}(\tilde{\theta}^*)'(\tilde{\theta}^* - \theta_0)\| &= \|n^{1/2}(\dot{\rho}(\tilde{\theta}^*) - \dot{\rho}(\bar{\theta}))'(\tilde{\theta}^* - \theta_0)\| \\
&\leq n^{1/2}\|\dot{\rho}(\tilde{\theta}^*) - \dot{\rho}(\bar{\theta})\|\|\tilde{\theta}^* - \theta_0\| \\
&= n^{1/2}B_0\|\tilde{\theta}^* - \bar{\theta}\|\|\tilde{\theta}^* - \theta_0\| \\
&= O_p(n^{1/2}m_0\lambda^2).
\end{aligned}$$

Since $n^{1/2}m_0\lambda^2 = n^{-1/2}m_0C^2\log p \leq C^2m_0(p/n)^{1/2}\log p = o(1)$,

$$n^{1/2}\dot{\rho}(\tilde{\theta}^*)'(\tilde{\theta}^* - \theta_0) = o_p(1). \tag{A.55}$$

Using (A.50), (A.52), (A.53) and (A.55), we have

$$\begin{aligned}
n^{1/2}\dot{\rho}(\tilde{\theta}^*)'\hat{H}(\tilde{\theta}^*)^{-1}S(\tilde{\theta}^*) &= n^{1/2}\dot{\rho}(\tilde{\theta}^*)'\hat{H}(\tilde{\theta}^*)^{-1}\left[S(\theta_0) - \hat{H}(\tilde{\theta}^*)(\tilde{\theta}^* - \theta_0) - R^*\right] \\
&= n^{1/2}\dot{\rho}(\tilde{\theta}^*)'\hat{H}(\tilde{\theta}^*)^{-1}S(\theta_0) - n^{1/2}\dot{\rho}(\tilde{\theta}^*)'(\tilde{\theta}^* - \theta_0) - n^{1/2}\dot{\rho}(\tilde{\theta}^*)'\hat{H}(\tilde{\theta}^*)^{-1}R^* \\
&= n^{1/2}\dot{\rho}(\theta_0)'H(\theta_0)^{-1}S(\theta_0) + o_p(1).
\end{aligned} \tag{A.56}$$

By Lemma B.1

$$\left(\dot{\rho}(\theta_0)'H(\theta_0)^{-1}I(\theta_0)H(\theta_0)^{-1}\dot{\rho}(\theta_0)\right)^{-1/2} n^{1/2}\dot{\rho}(\theta_0)'H(\theta_0)^{-1}S(\theta_0) \xrightarrow{d} N(0, I_r). \tag{A.57}$$

Then,

$$\begin{aligned}
&\left(\dot{\rho}(\tilde{\theta}^*)'\hat{H}(\tilde{\theta}^*)^{-1}\hat{I}(\tilde{\theta}^*)\hat{H}(\tilde{\theta}^*)^{-1}\dot{\rho}(\tilde{\theta}^*)\right)^{-1/2} n^{1/2}\dot{\rho}(\tilde{\theta}^*)'\hat{H}(\tilde{\theta}^*)^{-1}S(\tilde{\theta}^*) \\
&= \left(\dot{\rho}(\theta_0)'H(\theta_0)^{-1}I(\theta_0)H(\theta_0)^{-1}\dot{\rho}(\theta_0)\right)^{-1/2} n^{1/2}\dot{\rho}(\theta_0)'H(\theta_0)^{-1}S(\theta_0) + o_p(1) \\
&\xrightarrow{d} N(0, I_r),
\end{aligned} \tag{A.58}$$

where the equality holds by (A.51) and (A.56), and the convergence follows from (A.57) and Slutsky's lemma. Finally, from (A.58) and the CMT

$$C_\alpha(\rho_0) \xrightarrow{d} \chi^2_r.$$

B Supplementary lemmas

We first prove the following lemma that establishes the asymptotic distribution of a studentized quantity with the expected Hessian and information matrices and the score function evaluated at the true parameters.

Lemma B.1. *Let Assumption 1 hold and $p^{1+\delta_0}/n \to 0$ for some $0 < \delta_0 \leq 1$. Then, as $n \to \infty$*

$$\left(\dot{\rho}(\theta_0)'H(\theta_0)^{-1}I(\theta_0)H(\theta_0)^{-1}\dot{\rho}(\theta_0)\right)^{-1/2}\dot{\rho}(\theta_0)'H(\theta_0)^{-1}n^{1/2}S(\theta_0) \xrightarrow{d} N(0, I_r).$$

Proof of Lemma B.1. Let $s_i(\theta_0) \equiv w_i x_i \dot{g}(y_i, x_i'\theta_0)$, $X_{ni} \equiv n^{-1/2}\dot{\rho}(\theta_0)'H(\theta_0)^{-1}s_i(\theta_0)$ and $\Sigma_n \equiv \mathrm{Var}[\sum_{i=1}^n X_{ni}] = \dot{\rho}(\theta_0)'H(\theta_0)^{-1}I(\theta_0)H(\theta_0)^{-1}\dot{\rho}(\theta_0)$. Let $\nu_n \equiv \lambda_{\min}(\Sigma_n)$. We will verify the conditions of the multivariate Lindeberg-Feller CLT (see e.g. Theorem 9.3 of Hansen (2022*b*)). First note that $\mathrm{E}[X_{ni}] = 0$ because $\mathrm{E}[s_i(\theta_0)|x_i] = -\mathrm{E}[x_i w_i(y_i - \dot{a}(x_i'\theta_0))|x_i] = 0$.

Moreover, we have

$$
\begin{aligned}
\nu_n &= \min_{\tau\in\mathbb{R}^r\setminus\{0\}} \frac{\tau'\dot\rho(\theta_0)'H(\theta_0)^{-1}I(\theta_0)H(\theta_0)^{-1}\dot\rho(\theta_0)\tau}{\tau'\tau} \\
&\geq \min_{\tau\in\mathbb{R}^r\setminus\{0\}} \frac{\tau'\dot\rho(\theta_0)'H(\theta_0)^{-1}I(\theta_0)H(\theta_0)^{-1}\dot\rho(\theta_0)\tau}{\tau'\dot\rho(\theta_0)'\dot\rho(\theta_0)\tau} \min_{\tau\in\mathbb{R}^r\setminus\{0\}} \frac{\tau'\dot\rho(\theta_0)'\dot\rho(\theta_0)\tau}{\tau'\tau} \\
&\geq \lambda_{\min}(H(\theta_0)^{-1}I(\theta_0)H(\theta_0)^{-1})\lambda_{\min}(\dot\rho(\theta_0)'\dot\rho(\theta_0)) \\
&\geq \lambda_{\min}(H(\theta_0)^{-1})\lambda_{\min}(I(\theta_0))\lambda_{\min}(H(\theta_0)^{-1})\lambda_{\min}(\dot\rho(\theta_0)'\dot\rho(\theta_0)) \\
&= \frac{\lambda_{\min}(I(\theta_0))}{(\lambda_{\max}(H(\theta_0))^2}\lambda_{\min}(\dot\rho(\theta_0)'\dot\rho(\theta_0)) \\
&\geq \lambda_l^2/\lambda_u^2.
\end{aligned}
$$

where the first inequality follows from the extremal property of $\lambda_{\min}(\cdot)$, the second inequality is the eigenvalue product inequality (Hansen (2022*a*)) and the last inequality is by Assumption 1(b). Next, we will verify the Lindeberg condition: for $\delta = \frac{2}{\delta_0} > 0$ and any $\epsilon > 0$

$$
\frac{1}{\nu_n^2}\sum_{i=1}^n \mathrm{E}[\|X_{ni}\|^2 1(\|X_{ni}\| \geq (\epsilon\nu_n^2)^{1/2})] \leq \frac{1}{\nu_n^{2+\delta}\epsilon^{\delta/2}}\sum_{i=1}^n \mathrm{E}[\|X_{ni}\|^{2+\delta}] \to 0. \tag{B.1}
$$

First, note that

$$
\begin{aligned}
\|\dot\rho(\theta_0)'H(\theta_0)^{-1}x_i\|^{2+\delta} &\leq \|\dot\rho(\theta_0)\|^{2+\delta}\left(\|H(\theta_0)^{-1}x_i\|^2\right)^{1+\delta/2} \\
&\leq r^{1+\delta/2}\|\dot\rho(\theta_0)\|_2^{2+\delta}\left(\lambda_{\max}(H(\theta_0)^{-1}H(\theta_0)^{-1})\|x_i\|^2\right)^{1+\delta/2} \\
&\leq r^{1+\delta/2}\lambda_u^{2+\delta}\left(\frac{\|x_i\|^2}{(\lambda_{\min}(H(\theta_0)))^2}\right)^{1+\delta/2} \\
&\leq r^{1+\delta/2}\lambda_u^{2+\delta}\frac{(p+1)^{1+\delta/2}C_u^{2+\delta}}{\lambda_l^{2+\delta}}.
\end{aligned} \tag{B.2}
$$

where the first inequality is by Cauchy-Schwarz, the second inequality is by the inequality $\|\dot\rho(\theta_0)\| \leq r^{1/2}\|\dot\rho(\theta_0)\|_2$ and the extremal property of $\lambda_{\max}(\cdot)$, the third inequality is by the eigenvalue product inequality (Hansen (2022*a*), Appendix B), and the last inequality is by

Assumption 1(a) and (b). Thus, using $|w_i|^{2+\delta}|\dot{g}(y_i, x_i'\theta_0)|^{2+\delta} \leq C_u^{4+2\delta}$ and (B.2), we have

$$\begin{aligned}
\sum_{i=1}^{n} \|X_{ni}\|^{2+\delta} &\leq \frac{1}{n^{1+\delta/2}} \sum_{i=1}^{n} \|\dot{\rho}(\theta_0)' H(\theta_0)^{-1} x_i\|^{2+\delta} |w_i|^{2+\delta} |\dot{g}(y_i, x_i'\theta_0)|^{2+\delta} \\
&\leq \frac{1}{n^{\delta/2}} r^{1+\delta/2} \lambda_u^{2+\delta} \frac{(p+1)^{1+\delta/2} C_u^{2+\delta}}{\lambda_l^{2+\delta}} C_u^{4+2\delta} \\
&\leq \left(\frac{(p+1)^{1+\delta_0}}{n} \right)^{1/\delta_0} r^{1+\delta/2} \lambda_u^{2+\delta} \frac{C_u^{6+3\delta}}{\lambda_l^{2+\delta}} \\
&\to 0.
\end{aligned}$$

This verifies (B.1) and the result follows. □

Next, we present several lemmas to establish the consistency of the survey GLM Lasso estimator and confirm that the convergence rate obtained with i.i.d. data in the literature also holds with i.n.i.d. data.

To obtain the convergence rate of the Lasso estimator, following Bühlmann and van de Geer (2011) we define the empirical process associated with the negative log-likelihood, its local supremum, and the excess risk as:

$$v_n(\theta) \equiv n^{-1} \sum_{i=1}^{n} \left(w_i g(y_i, x_i'\theta) - \mathrm{E}[w_i g(y_i, x_i'\theta)] \right), \quad \theta \in \mathbb{R}^{p+1}, \tag{B.3}$$

$$\boldsymbol{Z}_R \equiv \sup_{\|\theta-\theta_0\|_1 \leq R} |v_n(\theta) - v_n(\theta_0)|, \tag{B.4}$$

$$\mathcal{E}(\theta) \equiv \mathrm{E}\left[n^{-1} \sum_{i=1}^{n} (w_i g(y_i, x_i'\theta) - w_i g(y_i, x_i'\theta_0)) \right]. \tag{B.5}$$

By Jensen's inequality,

$$\begin{aligned}
\mathcal{E}(\theta) &= \mathrm{E}\left[n^{-1} \sum_{i=1}^{n} (w_i g(y_i, x_i'\theta) - w_i g(y_i, x_i'\theta_0)) \right] = n^{-1} \sum_{i=1}^{n} \mathrm{E}\left[w_i \log \frac{f(y_i|x_i, \theta)}{f(y_i|x_i, \theta_0)} \right] \\
&\geq n^{-1} \sum_{i=1}^{n} w_i \log \mathrm{E}\left[\frac{f(y_i|x_i, \theta)}{f(y_i|x_i, \theta_0)} \right] = 0.
\end{aligned}$$

Therefore,

$$\theta_0 = \arg\min_{\theta \in \mathbb{R}^{p+1}} \mathcal{E}(\theta) = \arg\min_{\theta \in \mathbb{R}^{p+1}} \mathrm{E}\left[n^{-1} \sum_{i=1}^{n} w_i g(y_i, x_i'\theta) \right]. \tag{B.6}$$

The following lemma shows that $\boldsymbol{Z}_R$ is proportional to R and follows from Lemma 14.20

of Bühlmann and van de Geer (2011).

Lemma B.2 (Concentration inequality)**.** *Let Assumption 1 hold. Then, for all $R \leq \eta/C_u$*

$$\mathrm{E}[\boldsymbol{Z}_R] \leq 4Ra_n, \quad a_n \equiv C_u^2(C_u^2 + 0.5\eta C_u^2)\left(\frac{2\log(2(p+1))}{n}\right)^{1/2}. \tag{B.7}$$

Proof of Lemma B.2. Let $\gamma(y_i, w_i, s) = w_i g(y_i, s), i = 1, \ldots, n$, in Lemma 14.20 of Bühlmann and van de Geer (2011). Note that $|x_i'(\theta - \theta_0)| \leq \max_{1\leq j\leq p+1} |x_{ij}| \|\theta - \theta_0\|_1 \leq C_u R \leq C_u\eta/C_u = \eta$. By the second-order Taylor expansion and Assumption 1

$$w_i g(y_i, x_i'\theta) - w_i g(y_i, x_i'\theta_0) = w_i \dot{g}(y_i, x_i'\theta_0)x_i'(\theta - \theta_0) + 0.5 w_i x_i'(\theta - \theta_0)\ddot{g}(y_i, x_i'\theta^*)x_i'(\theta - \theta_0), \tag{B.8}$$

where θ^* is between θ and θ_0. By the triangle inequality and Assumption 1,

$$\begin{aligned} |w_i(g(y_i, x_i'\theta) - g(y_i, x_i'\theta_0))| &\leq |(w_i\dot{g}(y_i, x_i'\theta_0) + 0.5 w_i x_i'(\theta - \theta_0)\ddot{g}(y_i, x_i'\theta^*))\, x_i'(\theta - \theta_0)| \\ &\leq (C_u^2 + 0.5RC_u^3)|x_i'(\theta - \theta_0)|. \end{aligned} \tag{B.9}$$

Hence $\gamma(y_i, w_i, s) = w_i g(y_i, s)$ is Lipschitz, and Lemma 14.20 of Bühlmann and van de Geer (2011) and Assumption 1 yield

$$\begin{aligned} \mathrm{E}[\boldsymbol{Z}_R] &\leq 4R(C_u^2 + 0.5RC_u^3)\left(\frac{2\log(2(p+1))}{n}\right)^{1/2} \mathrm{E}\left[\max_{1\leq j\leq p+1} n^{-1}\sum_{i=1}^n x_{ij}^2\right] \\ &\leq 4RC_u^2(C_u^2 + 0.5RC_u^3)\left(\frac{2\log(2(p+1))}{n}\right)^{1/2}. \end{aligned}$$

□

We first recall the compatibility condition for a subset of indices $M \subseteq \{1, \ldots, p+1\}$ which represents the compatibility between a positive definite matrix (of the expected Hessian-type) and the sparsity of the model coefficients.

Assumption 2 (Compatibility Condition (CC))**.** *For a subset of indices $M \subseteq \{1, \ldots, p+1\}$, there exists $\kappa(M) > 0$ such that for all $\theta \in \mathbb{R}^{p+1}$ satisfying $\|\theta_{-M}\|_1 \leq 3\|\theta_M\|_1$ it holds that $\|\theta_M\|_1^2 \leq (\theta' H\theta)|M|/\kappa^2(M)$ as $n \to \infty$, where H is a positive definite fixed matrix.*

Related to the CC are the *restricted eigenvalue condition* (Hansen, 2022*a*, Chapter 29) and the *restricted isometry condition* (Negahban et al., 2012). For detailed discussions, we refer to p.129 and Sections 6.12 and 6.13 of Bühlmann and van de Geer (2011). The next assumption concerns the the quadratic behaviour of the excess risk around the true parameter.

Assumption 3 (Quadratic Margin Condition (QMC))**.** *There exist constants $\eta > 0$, $c > 0$ and a positive definite matrix H such that $\mathcal{E}(\theta) \geq c\|H^{1/2}(\theta - \theta_0)\|^2$ for all θ satisfying $\|X(\theta - \theta_0)\|_\infty \leq \eta$.*

For $c > 0$ in Assumption 3, define the oracle parameter vector θ^* as

$$\theta^* \equiv \arg \min_{\theta : M_\theta \subseteq \{1,\ldots,p+1\}} \left(3\,\mathcal{E}(\theta) + \frac{8\lambda^2 m_\theta}{\kappa^2(M_\theta)c} \right), \tag{B.10}$$

where $M_\theta \equiv \{1\} \cup \{j : \beta_j \neq 0\}$ and $m_\theta \equiv |M_\theta|$ denotes the cardinality of the subset M_θ. Moreover, let

$$\varepsilon^* \equiv \frac{3}{2}\mathcal{E}(\theta^*) + \frac{8\lambda^2 m_*}{2\kappa_*^2 c}, \tag{B.11}$$

where $m_* = |M_{\theta^*}|$ and $\kappa_* = \kappa(M_{\theta^*})$.

Assumption 4 ($\|\cdot\|_\infty$ neighborhood)**.** *For θ^* and ε^* defined in (B.10) and (B.11), assume that $\|X(\theta^* - \theta_0)\|_\infty \leq \eta$ and $\|X(\theta - \theta_0)\|_\infty \leq \eta$ for all $\|\theta - \theta^*\|_1 \leq R$, where $R \equiv \frac{\varepsilon^*}{\lambda_0}$ for some $\lambda_0 > 0$, and $\eta > 0$ is given in Assumption 3.*

Next, we recall Theorem 6.4 of Bühlmann and van de Geer (2011) (see also Corollary 6.6 therein) to derive the consistency and rate of convergence of the GLM Lasso estimator. The key condition for the result, in addition to Assumptions 2–4, is the convexity of the loss function (i.e. the convexity of ρ_f in f in Bühlmann and van de Geer (2011)'s notation) which holds because $wg(y,t)$ is convex in t.

Proposition B.3 (Theorem 6.4 of Bühlmann and van de Geer (2011))**.** *Suppose that there exist $\eta > 0$, $c > 0$ and a positive definite matrix H such that*

(a) Assumption 2 holds for all subsets of indices $M \subseteq \{1, \ldots, p+1\}$;

(b) Assumptions 3 and 4 hold;

(c) The function $g(y,t)$ is convex in t for all y;

(d) λ satisfies $\lambda \geq 8\lambda_0$.

Then on the set

$$\mathcal{F} = \{\boldsymbol{Z}_R \leq \lambda_0 R\} = \{\boldsymbol{Z}_R \leq \varepsilon^*\}, \tag{B.12}$$

where $\boldsymbol{Z}_R$ is defined in (B.4), it holds that

$$\mathcal{E}(\hat{\theta}) + \lambda\|\hat{\theta} - \theta^*\|_1 \leq 6\,\mathcal{E}(\theta^*) + \frac{16\lambda^2 m_*}{c\kappa_*^2}. \tag{B.13}$$

In the following lemma, we obtain the rate convergence of the Lasso estimator by verifying the conditions of Proposition B.3.

Lemma B.4. *Under Assumption 1 and the conditions of Proposition 3.6, $\|\hat{\theta} - \theta_0\|_1 = O_p(m_0\lambda)$, $\|\hat{\theta} - \theta_0\|^2 = O_p(m_0\lambda^2)$ and $n^{-1}\|X(\hat{\theta} - \theta_0)\|_2^2 = O_p(m_0\lambda^2)$.*

Proof of Lemma B.4. Following the remark of Bühlmann and van de Geer (2011) preceding Corollary 6.6 therein, let us set $M_\theta = \{1\} \cup \widetilde{M}_\theta$, where $\widetilde{M}_\theta \subseteq \{2, \ldots, p+1\}$ in the definition of the oracle (B.10). As a result, the unpenalized intercept α is kept in the oracle. When $\widetilde{M}_\theta = \widetilde{M}_{\theta_0}$, that is, $M_\theta = \{1\} \cup \widetilde{M}_{\theta_0}$, we have $\theta^* = \theta_0$, $\mathcal{E}(\theta^*) = \mathcal{E}(\theta_0) = 0$ and $\varepsilon^* = \frac{4\lambda^2 m_0}{\kappa_0^2 c}$.

The proof consists of three steps. The first step verifies the assumption of Proposition Proposition B.3. The second step provides a lower bound for $P[\mathcal{F}]$, where the event $\mathcal{F}$ is defined in (B.12). The final step completes the proof.

Step 1: Verifying the assumptions of Proposition B.3.

We will verify that the conditions of Proposition B.3 hold under Assumption 1. Since $\lambda_{\min}(H) > \lambda_l > 0$, by Lemma 6.23 of Bühlmann and van de Geer (2011) the adaptive restricted eigenvalue condition holds. The latter, in turn, implies that Assumption 2 holds for all index sets $M \subset \{1, \ldots, p+1\}$ (see Bühlmann and van de Geer (2011), p.162). Assumption 3 holds by the condition in (3.1) in Assumption 1. Next, we verify Assumption 4. Let $\lambda_0 = \frac{\lambda}{8} = \frac{C}{8}\sqrt{\frac{\log p}{n}}$. If $\|\theta - \theta_0\|_1 \leq R$, since $m_0\lambda \to 0$ from the rate assumption in Proposition 3.6, for n large

$$\|X(\theta - \theta^*)\|_\infty = \|X(\theta - \theta_0)\|_\infty \leq C_u\|\theta - \theta_0\|_1 \leq C_u R = \frac{4C_u\lambda^2 m_0}{\kappa_0^2 c\lambda_0} = \frac{32C_u\lambda m_0}{\kappa_0^2 c} \leq \eta. \tag{B.14}$$

The conditions of Proposition B.3 are therefore satisfied, and (B.13) implies that

$$\mathcal{E}(\hat{\theta}) + \lambda\|\hat{\theta} - \theta_0\|_1 \leq \frac{16\lambda^2 m_0}{c\kappa_0^2}, \tag{B.15}$$

hence on $\mathcal{F}$

$$\lambda\|\hat{\theta} - \theta_0\|_1 \leq \frac{16\lambda^2 m_0}{c\kappa_0^2}. \tag{B.16}$$

Step 2: Bounding $P[\mathcal{F}]$ for $\mathcal{F}$ defined in (B.12).

Set in Theorem A.1 of van de Geer (2008) that $\gamma(Z_i) = w_i[g(y_i, x_i'\theta) - g(y_i, x_i'\theta_0)]$. Following (B.9) for $\|\theta - \theta_0\|_1 \leq R$

$$\begin{aligned}
|w_i(g(y_i, x_i'\theta) - g(y_i, x_i'\theta_0))| &\leq |(w_i\dot{g}(y_i, x_i'\theta_0) + 0.5w_i x_i'(\theta - \theta_0)\ddot{g}(y_i, x_i'\theta^*))\, x_i'(\theta - \theta_0)| \\
&\leq (C_u^2 + 0.5RC_u^3)|x_i'(\theta - \theta_0)| \\
&\leq \eta(C_u^2 + 0.5\eta C_u^2).
\end{aligned}$$

Therefore,

$$\begin{aligned}
\|\gamma\|_\infty &\leq \eta(C_u^2 + 0.5\eta C_u^2) \equiv b_n, \qquad \text{(B.17)} \\
n^{-1}\sum_{i=1}^n \mathrm{Var}[\gamma(Z_i)] &\leq n^{-1}\sum_{i=1}^n \mathrm{E}[\gamma(Z_i)^2] \leq \eta^2(C_u^2 + 0.5\eta C_u^2)^2 = b_n^2.
\end{aligned}$$

Then, by the Bousquet's inequality (see Theorem A.1 of van de Geer (2008)) followed by Lemma B.2

$$\begin{aligned}
e^{-nt^2} &\geq P\left[\boldsymbol{Z}_R \geq \mathrm{E}[\boldsymbol{Z}_R] + t\sqrt{2(b_n^2 + 2b_n\,\mathrm{E}[\boldsymbol{Z}_R])} + 2t^2 b_n/3\right] \\
&\geq P\left[\boldsymbol{Z}_R \geq 4Ra_n + t\sqrt{2(b_n^2 + 8a_n b_n R)} + 2t^2 b_n/3\right],
\end{aligned}$$

where a_n is defined in (B.7). Replacing t by $4Rt$ in the above inequality yields

$$\begin{aligned}
P\left[\boldsymbol{Z}_R \leq 4R(a_n + t\sqrt{2(b_n^2 + 8a_n b_n R)} + 8b_n R^2 t^2/3)\right] &= P\left[\boldsymbol{Z}_R \leq \lambda_0 R\right] \\
&\geq 1 - e^{-16R^2 nt^2}, \qquad \text{(B.18)}
\end{aligned}$$

where $\lambda_0 = 4(a_n + t\sqrt{2(b_n^2 + 8a_n b_n R)} + 8b_n R^2 t^2/3)$.

Step 3: Completing the proof.

Since $nR^2 = O(m_0^2 \log p)$, with a suitable choice of t (hence with a suitable choice of C in λ_0 and λ), we obtain from (B.16) and (B.18) that $\|\hat{\theta} - \theta_0\|_1 = O_p(m_0\lambda)$. By Corollary 6.4 of Bühlmann and van de Geer (2011), $\|\hat{\theta} - \theta_0\|_2^2 = \|\hat{\theta} - \theta_0\|^2 = O_p(m_0\lambda^2)$. Combining the latter with (A.25), we obtain

$$\|X(\hat{\theta} - \theta_0)\|_2^2/n \leq \lambda_{\max}(n^{-1}X'X)\|\hat{\theta} - \theta_0\|_2^2 = O_p(m_0\lambda^2). \tag{B.19}$$

□