cemmap

centre for microdata methods and practice

Honest confidence regions for a regression parameter in logistic regression with a large number of controls

Alexandre Belloni
Victor Chernozhukov
Ying Wei

The Institute for Fiscal Studies
Department of Economics, UCL

cemmap working paper CWP67/13

E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL

An ESRC Research Centre

HONEST CONFIDENCE REGIONS FOR A REGRESSION PARAMETER IN LOGISTIC REGRESSION WITH A LARGE NUMBER OF CONTROLS

ALEXANDRE BELLONI*, VICTOR CHERNOZHUKOV†, AND YING WEI‡

Abstract. This paper considers inference in logistic regression models with high dimensional data. We propose new methods for estimating and constructing confidence regions for a regression parameter of primary interest α_0, a parameter in front of the regressor of interest, such as the treatment variable or a policy variable. These methods allow to estimate α_0 at the root-n rate when the total number p of other regressors, called controls, exceed the sample size n, using the sparsity assumptions. The sparsity assumption means that only s unknown controls are needed to accurately approximate the nuisance part of the regression function, where s is smaller than n. Importantly, the estimators and these resulting confidence regions are "honest" in the formal sense that their properties hold uniformly over s-sparse models. Moreover, these procedures do not rely on traditional "consistent model selection" arguments for their validity; in fact, they are robust with respect to "moderate" model selection mistakes in variable selection steps. Moreover, the estimators are semi-parametrically efficient in the sense of attaining the semi-parametric efficiency bounds for the class of models in this paper.

Key words: uniformly valid inference, instruments, double selection, Neymanization, optimality, sparsity, model selection

1. Introduction

The literature on high-dimensional generalized linear models has experienced rapid development [28, 18]. As in the case of linear mean regression models, a striking result of this literature is the achievement of consistency with the total number of covariates p being potentially much larger than the sample size n. The main underlying assumption for achieving consistency is sparsity, namely that the number of relevant controls is at most s, which is much smaller than n. Much of the interest focuses on ℓ_1-penalized estimators that achieve desirable theoretical and computational properties, at least when the log-likelihood functions are concave. The theoretical properties are analogous to those of the corresponding ℓ_1-penalized least squares estimator for linear mean regression models, called Lasso ([27, 9]). Results include prediction error consistency, consistency of the parameter estimates in ℓ_k-norms, variable selection consistency, and minimax-optimal rates.

Several papers have focused on high-dimensional logistic binary choice models, trying to exploit their structure in detail. ℓ_1-penalized logistic regressions models were studied in [10], [1], and [12]. Group

Date: First version: November 2012, this version August 2013.

*Duke University, e-mail:abn5@duke.edu.

†Massachusetts Institute of Technology, e-mail:vchern@mit.edu.

‡Columbia University, e-mail:yw2148@columbia.edu.

logistic regression were studied in [17] and [12]. Ising models were considered in [24] and connections with robust 1-bit recovery were derived in [21]. These works derive rates of convergence for the coefficients, prediction error consistency, and variable selection consistency under various conditions.

This paper attacks the problem of estimation and inference on a regression coefficient of interest in the logistic regression models, when $p \gg n$ is allowed. Specifically, we construct $\sqrt{n}$-consistent estimators and confidence regions for a parameter of interest α_0, which measures the impact of a regressor of interest – typically a "policy variable"– on the regression function. Importantly, we show that the estimator is $\sqrt{n}$-consistent and the confidence regions achieve the required asymptotic coverage uniformly over many data-generating processes. Moreover, our estimator attains the semi-parametric efficiency for the class of models considered here.

It is important to note that our estimation and inferential results are valid *without assuming* the conventional "separation condition" – namely, without assuming that all the non-zero coefficients are bounded away from zero. Although the separation condition is commonly used and might be appealing in some technometric applications, it is often unrealistic and not credible in econometric, biometric, and other applications of interest, or, even if applicable, it might not lead to accurate approximations to the true finite sample behavior of estimation and inference procedures. Our procedures are robust to violation of the separation condition, and, thus, are robust to "moderate" model selection mistakes, which inevitably occur in this case (the mistakes do occur because it is simply impossible to tell apart the variables with coefficients near zero from the variables with coefficients equal zero).

Our work contributes to the small, yet growing literature that avoids imposing "separation conditions". In the context of instrumental regression, [4] and [2] provide uniformly valid estimation and inference methods for instrumental variable models, using either post-selection and ℓ_1 regularization methods to estimate "optimal instruments". They provide a $\sqrt{n}$ consistent, semi-parametrically efficient estimator of the main low-dimensional structural parameter. In the context of the linear mean regression model, [8, 7] proposed a double selection approach to constructing uniformly valid estimation and inference methods, and [33] used one-step corrections to ℓ_1-regularized estimators, in the context of the linear mean regression model. In either case a $\sqrt{n}$-consistent, semi-parametrically efficient estimator of the main low-dimensional regression parameter is provided. In the context of the linear quantile regression models, [6, 5] provide uniformly $\sqrt{n}$-consistent estimators and uniformly valid inference methods for least absolute deviations and quantile regressions. In an independent and contemporaneous work, [29] propose an approach to inference in generalized linear models, based upon the one-step correction of ℓ_1-penalized estimator, where the pieces of the corrections are estimated via (approximate) Lasso inversion of the sample information matrix; they also provide theoretical analysis under high-level conditions. The approach taken in the present paper is an independent proposal, and relies instead on either "optimal instrument" strategy, which is related to Neyman's approach to dealing with nuisance parameters, or the "double selection" strategy. Both strategies build upon our own previous contributions (in the context of linear instrumental and mean regression) referenced above.

The aforementioned works as well as the current approach deviate substantially from the traditional approach of performing inference based upon perfect model selection results. [16] and [23, 22, 14] have shown that such an inference approach is not robust to violations of the "separation condition", which bounds the magnitude of the non-zero coefficients away from zero. The naive post-selection estimators and inference based upon them break down in the sense of failing to achieve $\sqrt{n}$-consistency and asymptotic normality under the failure of the "separation condition". We shall confirm the failure of such naive post-selection procedures in very simple Monte-Carlo experiments. In sharp contrast our procedure, by construction, is robust to violation of such assumptions. We shall demonstrate this via theoretical results as well as via an extensive set of Monte-Carlo experiments. The theoretical results hold uniformly in the class of s-sparse models (and can be shown also to hold over approximately sparse models), using arguments similar to those used for linear mean and quantile models in [7] and [6, 5].

We construct our estimators and confidence regions via three steps. The first step use post-model selection methods to estimate the nuisance part of the regression – the part of the regression function associated to controls, or non-main regressors. The second step uses post-model selection to estimate an "optimal instrument". The third step suitably combines these estimates to form estimating equations that are immunized against crude estimation of the nuisance functions. Solutions of these equations provide the optimal estimators and lead to our proposed confidence regions. In the third step, we propose one implementation based upon instrumental logistic regression with "optimal instrument" and another implementation based upon "double selection" logistic regression. We verify the uniform validity of these procedures and demonstrate their good properties in a wide variety of experiments. While both implementations perform well, the "double selection" procedure emerged as the clear winner in these experiments. Also, our results and proofs reveal that many different estimators can be used as ingredient in the three steps of the algorithm, as long as a required sparsity and rates for estimating nuisance functions are achieved. For example, the first and second steps can be based not only on post-selection estimators but also on ℓ_1-regularized estimators, while the third step can be alternatively approximated by a "one-step" correction from an initial value. Therefore several implementations having the same asymptotic properties are possible. We narrowed down our formal theoretical analysis to the set of procedures that exhibited the best performance in Monte-Carlo experiments (for example, Lasso methods performed worse than post-Lasso methods for estimating the nuisance parts, and "one-step" corrections performed worse than the exact solution of the estimating equation).

Our constructions of the final estimators and confidence regions mainly make use of the post-model selection estimators in estimating the nuisance part of the regression function as well as the optimal instrument. We focus on using selection as a means of regularization (which is necessary when $p > n$), mainly because compared to other methods of regularization, such as ℓ_1-penalized maximum likelihood, they performed best in a wide set of experiments. In order to develop sharp results for these estimators we must control sparsity effectively. We therefore provide sparsity bounds for ℓ_1-penalized logistic maximum likelihood estimators, which is used for selection, and also derive the rates of convergence for the post-model selection logistic maximum likelihood estimator. These results are of independent interest. Also in the estimation of "optimal instruments", which we use as an ingredient in building the optimal estimating equation, we rely on post-selection least squares estimator with data dependent weights. There too the presence of data-dependent weights, which is needed to create "immunization property" for the final step, creates several interesting technical challenges. Finally, to obtain the asymptotic approximations to the estimators of regression coefficients of interest we rely on empirical process methods, using self-normalized maximal inequalities and entropy calculations that rely on the sparsity of the models selected via data-driven procedures. These proofs are of independent interests in other types of generalized linear models.

We organize the remainder of the paper as follows. In Section 2, we present the model and the proposed estimator. In Section 3 we provide primitive conditions and the statements of our main results on the uniform validity of the estimators and confidence regions. We present the proofs of these results in Appendix B, which we base on carefully verifying the high-level conditions of a general result in Appendix A. In Appendix C we collect results on Lasso and Post-Lasso with estimated weights (Appendix C.1) as well as results on ℓ_1-penalized Logistic regression and post model selection Logistic regression (Appendix C.2). In Appendix D we present auxiliary inequalities.

1.1. **Notation.** Denote by (Ω, P) the underlying probability space. The notation $\mathbb{E}_n[\cdot]$ denotes the average over index $1 \leqslant i \leqslant n$, i.e., it simply abbreviates the notation $n^{-1}\sum_{i=1}^n[\cdot]$. For example, $\mathbb{E}_n[x_{ij}^2] = n^{-1}\sum_{i=1}^n x_{ij}^2$. Moreover, we use the notation $\bar{\mathrm{E}}[\cdot] = \mathbb{E}_n[\mathrm{E}[\cdot]]$. For example, $\bar{\mathrm{E}}[v_i^2] = n^{-1}\sum_{i=1}^n \mathrm{E}[v_i^2]$. For a function $f : \mathbb{R}\times\mathbb{R}\times\mathbb{R}^p \to \mathbb{R}$, we write $\mathbb{G}_n(f) = n^{-1/2}\sum_{i=1}^n (f(y_i, d_i, x_i) - \mathrm{E}[f(y_i, d_i, x_i)])$. We denote the l_1-norm as $\|\cdot\|_1$, l_2-norm as $\|\cdot\|$, l_∞-norm as $\|\cdot\|_\infty$, and the "l_0-norm" as $\|\cdot\|_0$ to denote the number of non-zero components of a vector. For a sequence $(t_i)_{i=1}^n$, we denote $\|t_i\|_{2,n} = \sqrt{\mathbb{E}_n[t_i^2]}$. For example, for a vector $\delta \in \mathbb{R}^p$, $\|x_i'\delta\|_{2,n} = \sqrt{\mathbb{E}_n[(x_i'\delta)^2]}$ denotes the prediction norm of δ. Given a vector $\delta \in \mathbb{R}^p$, and a set of indices $T \subseteq \{1, \ldots, p\}$, we denote by $\delta_T \in \mathbb{R}^p$ the vector such that $(\delta_T)_j = \delta_j$ if $j \in T$ and $(\delta_T)_j = 0$ if $j \notin T$. The support of δ as $\mathrm{support}(\delta) = \{j \in \{1, ..., p\} : \delta_j \neq 0\}$. We use the notation $(a)_+ = \max\{a, 0\}$, $a \vee b = \max\{a, b\}$, and $a \wedge b = \min\{a, b\}$. We also use the notation $a \lesssim b$ to denote $a \leqslant cb$ for some constant $c > 0$ that does not depend on n; and $a \lesssim_P b$ to denote $a = O_P(b)$. We assume that the quantities such as p, s, $y_i, d_i, x_i, \beta_0, \theta_0, T$ and T_{θ_0} are all dependent on the sample size n, and allow for the case where $p = p_n \to \infty$ and $s = s_n \to \infty$ as $n \to \infty$. We omit the dependence of these quantities on n for notational convenience.

2. Setup and Method

Consider a binary regression model, where the binary outcome of interest y_i relates to a scalar main regressor (e.g. a treatment or a policy variable) d_i and p-dimensional controls x_i through a link function G, namely

$$\mathrm{E}[y_i \mid x_i, d_i] = G(d_i\alpha_0 + x_i'\beta_0). \tag{2.1}$$

Here α_0 is the main target parameter, and $x_i'\beta_0$ is the nuisance regression function, which is assumed to be sparse, namely $\|\beta_0\|_0 \leqslant s$. We require s will be small relative to n in the sense that will be specified below, in condition L, in particular

$$\frac{s^2 \log p}{n} \to 0$$

is required. This conditions allows for estimation of the nuisance function at the rate of $o(n^{-1/4})$. This work studies the logistic regression in which the link function is given by

$$G(t) := \exp(t)/\{1+\exp(t)\}.$$

Let $\{(y_i, d_i, x_i) \ : \ i = 1, \ldots, n\}$ be a random sample, independent across i, obeying the model (2.1) with $\|\beta_0\|_0 \leqslant s$. We aim to perform statistical inference on the coefficient α_0 that is robust to moderate model selection mistakes. Our proposed methods rely (implicitly or explicitly) on an instrument $z_{0i} = z_0(d_i, x_i)$ such that:

$$\mathrm{E}[\{y_i - G(d_i\alpha_0 + x_i'\beta_0)\}z_{0i}] \qquad = 0, \tag{2.2}$$

$$\frac{\partial}{\partial\beta}\mathrm{E}[\{y_i - G(d_i\alpha_0 + x_i'\beta)\}z_{0i}]\Big|_{\beta=\beta_0} \quad = 0. \tag{2.3}$$

The first relation provides an estimating equation for α_0. The second relation states that this equation must be insensitive with respect to first order perturbations of the nuisance function $x_i'\beta_0$, which we call "immunity". Such immunization ideas can be traced to Neyman's approach to dealing with nuisance parameters, as we explain in detail in Section 5.

Our methods have three different steps:

1. The first step computes an estimate for the nuisance function $x_i'\beta_0$.
2. The second step estimates the instrument z_{0i}.
3. The third step combines these estimates to estimate the parameter of interest α_0.

Estimation of nuisance functions $x_i'\beta_0$ and the instrument z_{0i} has an asymptotically negligible effect, due to the "immunization properties" of the estimating equations. Several different choices for these procedures and for instruments are possible. Next we provide detailed recommendations for their choices.

In general, we construct a valid, optimal instrument based on the following decomposition for the weighted main regressor:

$$f_i d_i = f_i x_i'\theta_0 + v_i, \quad \text{with} \ \ \mathrm{E}\left[f_i v_i \mid x_i\right] = 0, \tag{2.4}$$

where

$$f_i := \frac{w_i}{\sigma_i}, \quad w_i := G'(d_i\alpha_0 + x_i'\beta_0), \quad \sigma_i^2 := \mathrm{Var}(y_i|d_i, x_i), \tag{2.5}$$

where $G'(t) = \frac{\partial}{\partial t}G(t)$. The optimal instrument is given by

$$z_{0i} := \frac{v_i}{\sigma_i}. \tag{2.6}$$

Here too we shall impose a sparsity condition in (2.4), namely that $\|\theta_0\|_0 \leqslant s$. The use of sparsity in the main equation and this auxiliary equation can be generalized to approximate sparsity, with all results in this paper extending to this case.

The weights $f_i = w_i/\sigma_i$'s are used to achieve the orthogonality condition (2.3):

$$\frac{\partial}{\partial\beta}\mathrm{E}[\{y_i - G(d_i\alpha_0 + x_i'\beta)\}z_{0i}]\Big|_{\beta=\beta_0} = \mathrm{E}[w_i z_{0i} x_i] = \mathrm{E}\left[f_i v_i x_i\right] = 0; \tag{2.7}$$

and this condition "immunizes" the estimation of the main parameter α_0 against crude estimation of the nuisance function $x_i'\beta_0$, in particular via post-selection estimators. The selection steps make unavoidable "moderate" model selection mistakes, which translate into vanishing estimation error, which has an asymptotic negligible effect on the estimator based on the sample analog of the equation (2.2). The orthogonality (2.3) condition is therefore a critical ingredient in achieving asymptotic uniform validity of the estimator and the confidence regions. Among all instruments that provide such immunization, the instrument given in (2.6) minimizes the asymptotic variance of the asymptotically normal and $\sqrt{n}$-consistent estimator based on the estimating equations (2.2). Other valid (but sub-optimal) choices of instruments are discussed in Comment 2.1.

In the present paper we apply the above principle to the case of logistic regression, in which case $G(t) = \exp(t)/\{1 + \exp(t)\}$, and the following simplification occurs: w_i in (2.5) becomes the conditional variance of the outcome σ_i^2, namely

$$w_i = \sigma_i^2 = G(d_i\alpha_0 + x_i'\beta_0)\{1 - G(d_i\alpha_0 + x_i'\beta_0)\},$$

so that the decomposition (2.4) and the optimal instrument (2.6) become

$$\sqrt{w_i}d_i = \sqrt{w_i}x_i'\theta_0 + v_i, \quad \text{with } \mathrm{E}\,[\sqrt{w_i}v_i \mid x_i] = 0 \quad \text{and} \quad z_{0i} = v_i/\sqrt{w_i}. \tag{2.8}$$

We describe two resulting (recommended) implementations in Tables 1 and 2. In these tables we denote the (negative) log-likelihood function associated with the logistic link function as

$$\Lambda(\alpha, \beta) = \mathbb{E}_n[\log\{1 + \exp(d_i\alpha + x_i'\beta)\} - y_i(d_i\alpha + x_i'\beta)]. \tag{2.9}$$

Table 1 displays an implementation based on the optimal instrument. The estimation in Step 1 is based on post-selection logistic regression where the model is selected based on ℓ_1-penalized logistic regression. Step 2 is based on a post-selection least squares with weights estimated based upon Step 1. It estimates the optimal instrument. Step 3 is based on an instrumental logistic regression, with estimates of nuisance functions (control function $x_i'\beta_0$ and the instrument z_{0i}) obtained in Steps 1 and 2. The use of post-selection estimators in first two steps instead of penalized estimators was motivated by a (much) better finite sample performance in our experiments. We also provide two confidence regions for α_0 in this table: the direct confidence region $\mathcal{CR}_D$ is based on the asymptotic normality of the estimator $\check{\alpha}$. The indirect confidence region $\mathcal{CR}_I$ is based on the asymptotic $\chi^2(1)$ law of the statistic $nL_n(\alpha_0)$.

Table 2 describes an alternative implementation, which builds upon the idea of the double selection method proposed in [7] for partial linear mean regression models. The method replaces Step 3 in Table 1 with a logistic regression of the outcome on the main regressors as well as the union of controls selected in two selection steps – Steps 1 and 2. This approach creates an optimal instrument implicitly. In fact, inspection of the proof shows that the double-selection estimator can be seen as an infinitely iterated version of the previous method. We refer to Section 5.1 for further connections and discussions.

Comment 2.1 (Other Valid Instruments)**.** An instrument z_0 is valid if it has the orthogonality property

$$\frac{\partial}{\partial\beta}\mathrm{E}[\{y_i - G(d_i\alpha_0 + x_i'\beta)\}z_{0i}]\bigg|_{\beta=\beta_0} = \mathrm{E}[w_i z_{0i} x_i] = 0$$

Estimators and Honest Confidence Regions based on Optimal Instrument

Step 1 Run Post-Lasso-Logistic of y_i on d_i and x_i:

$$(\widehat{\alpha},\widehat{\beta}) \in \arg\min_{\alpha,\beta} \Lambda(\alpha,\beta) + \tfrac{\lambda_1}{n}\|(\alpha,\beta)\|_1$$
$$(\widetilde{\alpha},\widetilde{\beta}) \in \arg\min_{\alpha,\beta} \Lambda(\alpha,\beta) \ : \ \text{support}(\beta) \subseteq \text{support}(\widehat{\beta})$$

For $i = 1,\ldots,n$, keep the value $x_i'\widetilde{\beta}$ and weight

$$\widehat{f}_i := \widehat{w}_i/\widehat{\sigma}_i, \ \text{where } \widehat{w}_i = G'(d_i\widetilde{\alpha} + x_i'\widetilde{\beta}), \quad \widehat{\sigma}_i^2 = G(d_i\widetilde{\alpha} + x_i'\widetilde{\beta})\{1 - G(d_i\widetilde{\alpha} + x_i'\widetilde{\beta})\}.$$

Step 2 Run Post-Lasso-OLS of $\widehat{f}_i d_i$ on $\widehat{f}_i x_i$:

$$\widehat{\theta} \in \arg\min_{\theta} \mathbb{E}_n[\widehat{f}_i^2(d_i - x_i'\theta)^2] + \tfrac{\lambda_2}{n}\|\widehat{\Gamma}\theta\|_1$$
$$\widetilde{\theta} \in \arg\min_{\theta} \mathbb{E}_n[\widehat{f}_i^2(d_i - x_i'\theta)^2] \ : \ \text{support}(\theta) \subseteq \text{support}(\widehat{\theta})$$

Keep the residual $\widehat{v}_i := \widehat{f}_i(d_i - x_i'\widetilde{\theta})$ and instrument $\widehat{z}_i := \widehat{v}_i/\sqrt{\widehat{\sigma}_i}$, $i = 1,\ldots,n$.

Step 3 Run Instrumental Logistic Regression of $y_i - x_i'\widetilde{\beta}$ on d_i using $\widehat{z}_i$ as the instrument for d_i

$$\check{\alpha} \in \arg\inf_{\alpha\in\mathcal{A}} L_n(\alpha), \quad \text{where } \ L_n(\alpha) = \frac{|\ \mathbb{E}_n[\ \{y_i - G(d_i\alpha + x_i'\widetilde{\beta})\}\widehat{z}_i\ ]\ |^2}{\mathbb{E}_n[\ \{y_i - G(d_i\alpha + x_i'\widetilde{\beta})\}^2\widehat{z}_i^2\ ]}$$

where $\mathcal{A} = \{\alpha \in \mathbb{R} : |\alpha - \widetilde{\alpha}| \leqslant C/\log n\}$. Define the confidence regions with asymptotic coverage $1 - \xi$

$$\mathcal{CR}_D = \{\alpha \in \mathbb{R} \ : |\alpha - \check{\alpha}| \leqslant \widehat{\Sigma}_n\Phi^{-1}(1-\xi/2)/\sqrt{n}\}$$
$$\mathcal{CR}_I = \{\alpha \in \mathcal{A} : nL_n(\alpha) \leqslant (1-\xi)\text{−quantile of } \chi^2(1)\}.$$

Table 1. The algorithm has three steps: (1) initial estimation of the regression function via post-selection logistic regression, (2) estimation of instruments which are orthogonal to the weighted controls via a weighted post-selection least squares, and (3) estimation of α_0 based on the nuisance estimates obtain in (1) and (2). This step guards against "moderate model selection" mistakes and constructs semi-parametrically efficient estimator. We assume the normalization $\mathbb{E}_n[x_{ij}^2] = 1$ and $\mathbb{E}_n[d_i^2] = 1$, and penalty parameters $\lambda_1 = \frac{1.1}{2}\sqrt{n}\Phi^{-1}(1 - 0.05/\{n \vee p\log n\})$, $\lambda_2 = 1.1\sqrt{n}2\Phi^{-1}(1 - 0.05/\{n \vee p\log n\})$ and $\widehat{\Gamma}$ is defined in the appendix, see (C.45). The estimator of the variance is given by $\widehat{\Sigma}_n^2 = \max\{\widehat{\Sigma}_{1n}^2, \widehat{\Sigma}_{2n}^2\}$ where $\widehat{\Sigma}_{1n}^2 = \{\mathbb{E}_n[\widehat{w}_i d_i \widehat{z}_i]\}^{-1}\mathbb{E}_n[\{y_i - G(d_i\check{\alpha} + x_i'\widetilde{\beta})\}^2\widehat{z}_i^2]\{\mathbb{E}_n[\widehat{w}_i d_i \widehat{z}_i]\}^{-1}$ and $\widehat{\Sigma}_{2n}^2 = \mathbb{E}_n[\widehat{v}_i^2]$.

and is non-trivial, namely $\bar{\mathrm{E}}[w_i d_i z_{0i}] \neq 0$. A valid, non-trivial instrument is optimal if it minimizes the asymptotic variance of the final estimator of α_0. The algorithm stated in Table 1 uses the optimal instrument $z_{0i} := v_i/\sqrt{w_i}$. Estimation of this instrument requires that in Step 2 a Lasso method is applied in the weighted equation (2.4). Since the weights w_i's in the resulting weighted Lasso problem are estimated, with estimation errors depending upon the response variable d_i, estimation of the optimal instrument creates interesting technical challenges in the analysis of Lasso or post-Lasso that are dealt with in the Appendix. Thus estimation of the optimal instruments poses an interesting problem in its own right.

Estimators and Honest Confidence Region based on Double Selection

Step 1 Run Post-Lasso-Logistic of y_i on d_i and x_i:

$$(\widehat{\alpha},\widehat{\beta}) \in \arg\min_{\alpha,\beta} \ \Lambda(\alpha,\beta) + \tfrac{\lambda_1}{n}\|(\alpha,\beta)\|_1$$
$$(\widetilde{\alpha},\widetilde{\beta}) \in \arg\min_{\alpha,\beta} \ \Lambda(\alpha,\beta) \ : \ \text{support}(\beta) \subseteq \text{support}(\widehat{\beta})$$

For $i = 1,\ldots,n$, construct the weights

$$\widehat{f}_i := \widehat{w}_i/\widehat{\sigma}_i, \ \text{where } \widehat{w}_i = G'(d_i\widetilde{\alpha} + x_i'\widetilde{\beta}), \quad \widehat{\sigma}_i^2 = G(d_i\widetilde{\alpha} + x_i'\widetilde{\beta})\{1 - G(d_i\widetilde{\alpha} + x_i'\widetilde{\beta})\}.$$

Step 2 Run Lasso-OLS of $\widehat{f}_i d_i$ on $\widehat{f}_i x_i$:

$$\widehat{\theta} \in \arg\min_{\theta} \ \mathbb{E}_n[\widehat{f}_i^2(d_i - x_i'\theta)^2] + \tfrac{\lambda_2}{n}\|\widehat{\Gamma}\theta\|_1$$

Step 3 Run Post-Lasso-Logistic of y_i on d_i and the covariates selected in Step 1 and 2:

$$(\check{\alpha},\check{\beta}) \in \arg\min_{\alpha,\beta} \ \Lambda(\alpha,\beta) \ : \ \text{support}(\beta) \subseteq \text{support}(\widehat{\beta}) \cup \text{support}(\widehat{\theta})$$

Define the confidence region with asymptotic coverage $1-\xi$ as

$$\mathcal{CR}_{DS} = \{\alpha \in \mathbb{R} \ : |\alpha - \check{\alpha}| \leqslant \widehat{\Sigma}_n \Phi^{-1}(1-\xi/2)/\sqrt{n}\}.$$

TABLE 2. The double selection algorithm has three steps: (1) use ℓ_1-penalized logistic regression to select covariates; and use post-selection logistic regression to estimate the weights to be used in the next step, (2) select covariates based on the weighted post-selection least squares, where the dependent variable is the main regressor and the independent variables are the rest of the regressors, and (3) run a Logistic regression of the outcome on the main regressors and the union of controls in steps (1) and (2) . We assume the normalization $\mathbb{E}_n[x_{ij}^2] = 1$ and $\mathbb{E}_n[d_i^2] = 1$, and penalty parameters $\lambda_1 = \frac{1.1}{2}\sqrt{n}\Phi^{-1}(1-0.05/\{n \vee p\log n\})$, $\lambda_2 = 1.1\sqrt{n}2\Phi^{-1}(1-0.05/\{n \vee p\log n\})$ and $\widehat{\Gamma}$ is defined in the appendix, see (C.45). The estimator of the variance is given by $\widehat{\Sigma}_n^2 = \max\{\widehat{\Sigma}_{1n}^2, \widehat{\Sigma}_{2n}^2\}$ where $\widehat{\Sigma}_{1n}^2 = \{\mathbb{E}_n[\check{w}_i d_i \widehat{z}_i]\}^{-1}\mathbb{E}_n[\{y_i - G(d_i\check{\alpha} + x_i'\check{\beta})\}^2\widehat{z}_i^2]\{\mathbb{E}_n[\check{w}_i d_i \widehat{z}_i]\}^{-1}$, $\Sigma_{2n}^2 = \{\mathbb{E}_n[\check{w}_i(d_i,\check{x}_i')'(d_i,\check{x}_i')]\}_{11}^{-1}$, $\check{w}_i = G(d_i\check{\alpha} + x_i'\check{\beta})\{1 - G(d_i\check{\alpha} + x_i'\check{\beta})\}$ and $\check{x}_i = x_{i,\text{support}(\check{\beta})}$.

There are other valid instruments that we can rely on, but these instruments are not generally optimal. For example, a valid, yet sub-optimal choice of the instrument is $z_{0i} := (d_i - \mathrm{E}[d_i \mid x_i])/w_i$. The estimation of this instrument is technically simpler, and follows easily from available results. Indeed, in this case, assuming $\mathrm{E}[d_i \mid x_i] = x_i'\theta_d$, with θ_d sparse or approximately sparse, we can estimate z_{0i} by estimating θ_d via standard Lasso of d_i on x_i, and estimating w_i using the estimates of the ℓ_1-penalized logistic regression as in Step 1. Note that since no estimated weights are used in Lasso estimation of θ_d, standard results on the Lasso estimator deliver the required properties. We further discuss the choice of instruments in Section 5. ■

Comment 2.2 (Alternative Implementations via Approximate Instrumental Regression)**.** The instrumental logistic regression can be approximately implemented by a 1-Step estimator from the ℓ_1-penalized logistic estimator $\widehat{\alpha}$ of the form $\check{\alpha} = \widehat{\alpha} + (\mathbb{E}_n[\widehat{w}_i d_i \widehat{z}_i])^{-1}\mathbb{E}_n[\{y_i - G(d_i\widehat{\alpha} + x_i'\widehat{\beta})\}\widehat{z}_i]$. However, we prefer the exact implementations, since the they perform better in an extensive set of Monte-Carlo experiments. ■

3. Main Result

3.1. **Primitive Assumptions.** In this section, we list and discuss primitive conditions that allow us to derive our results. We consider the following quantities associated with the covariates in the sample $\tilde{x}_i = (d_i, x_i')'$, $i = 1, \ldots, n$. We denote the largest value of the components of the controls x_i as $K_x = \max_{i \leqslant n} \|x_i\|_\infty$ and denote the minimum and maximum m-sparse empirical eigenvalues as

$$\phi_{\min}(m) := \min_{1 \leqslant \|\delta\|_0 \leqslant m} \frac{\|\tilde{x}_i'\delta\|_{2,n}^2}{\|\delta\|^2} \quad \text{and} \quad \phi_{\max}(m) := \max_{1 \leqslant \|\delta\|_0 \leqslant m} \frac{\|\tilde{x}_i'\delta\|_{2,n}^2}{\|\delta\|^2}.$$

The following are sufficient primitive conditions.

Fix some sequences of constants, $\delta_n \to 0$, $\ell_n \to \infty$, and $\Delta_n \to 0$, and constants $0 < c < C < \infty$.

Condition L. *(i) Let $(x_i)_{i=1}^n$ denote a sequence of non-stochastic vectors in $\mathbb{R}^p$ of covariates normalized so that $\mathbb{E}_n[x_{ij}^2] = 1$, $j = 1, \ldots, p$, and $\{(y_i, d_i, v_i, w_i) : i = 1, \ldots, n\}$ be independent random vectors that obey the model given by (2.1) and (2.4). There exists $s = s_n$ such that $\|\beta_0\|_0 + \|\theta_0\|_0 \leqslant s$. (ii) The weights w_i satisfy $\min_{i \leqslant n} w_i \geqslant c > 0$ with probability $1 - \Delta_n$, and the following moment conditions hold $0 < c \leqslant \bar{\mathrm{E}}[v_i^2 \mid x_i] \leqslant \max_{i \leqslant n}\{\mathrm{E}[v_i^4/w_i^2 \mid x_i]\}^{1/2} \vee \{\mathrm{E}[d_i^4 \mid x_i]\}^{1/2} \leqslant C$, $\bar{\mathrm{E}}[v_i^8 \mid x_i] \leqslant C$. (iii) The sparse minimal and maximal eigenvalues are bounded, $c \leqslant \phi_{\min}(s\ell_n) \leqslant \phi_{\max}(s\ell_n) \leqslant C$ with probability $1 - \Delta_n$. (iv) The sparsity index s and overall number of controls p obey the following growth conditions $K_x^2 s^2 \log^3(p \vee n) \leqslant n\delta_n$ and $K_x^q \log(p \vee n) \leqslant \delta_n n$ for some fixed $q > 4$.*

Condition L(i) assumes independence across i and the model described in Section 2. Condition L(ii) assumes the conditional variance is bounded away from zero and mild moment conditions. Condition L(iii) assumes that the sparse eigenvalues of size $s\ell_n$ are well behaved. In the case that the variables $(\tilde{x}_i)_{i=1}^n$ have been generated as i.i.d. realizations of some random vector $\tilde{X}_i$, these conditions are implied by a variety of conditions if the population Gram matrix $\mathrm{E}[\tilde{X}_i\tilde{X}_i']$ has the corresponding sparse eigenvalues well behaved; see [26, 25] for detailed discussion. We refer the reader to the Appendix for weaker high-level conditions that also imply our results.

3.2. **Uniformly Valid Estimators and Confidence Regions.** Next we state the main inferential results of the paper. It concerns the (uniform) validity of the different confidence regions for the coefficient α_0 based on the optimal instrument and double selection algorithms.

Theorem 1 (Robust Estimation and Inference based on the Optimal IV Estimator)**.** *Consider any triangular array of data $(y_i, d_i, x_i)_{i=1}^n$ (i.e., with all variables implicitly indexed by n, with $n = 1, 2, \ldots$) that obeys Condition L for all $n \geqslant 1$. Then, the IV estimator $\check{\alpha}$, based on the optimal instrument, obeys as $n \to \infty$*

$$\Sigma_n^{-1}\sqrt{n}(\check{\alpha} - \alpha_0) = Z_n + o_{\mathrm{P}}(1), \quad Z_n \rightsquigarrow N(0, 1),$$

where

$$Z_n := \frac{\Sigma_n}{\sqrt{n}} \sum_{i=1}^n (y_i - G(d_i\alpha_0 + x_i'\beta_0))z_{0i} \ \text{and} \ \ \Sigma_n^2 := \bar{\mathrm{E}}[v_i^2]^{-1}.$$

Moreover,

$$nL_n(\alpha_0) = Z_n^2 + o_{\mathrm{P}}(1), \quad Z_n^2 \rightsquigarrow \chi^2(1).$$

Finally, Σ_n^2 can be replaced by either $\widehat{\Sigma}_{1n}^2 = \{\mathbb{E}_n[\widehat{w}_i d_i \widehat{z}_i]\}^{-1}\mathbb{E}_n[\{y_i - G(d_i\check{\alpha} + x_i'\widetilde{\beta})\}^2\widehat{z}_i^2]\{\mathbb{E}_n[\widehat{w}_i d_i \widehat{z}_i]\}^{-1}$ or by $\widehat{\Sigma}_{2n}^2 = \mathbb{E}_n[\widehat{v}_i^2]^{-1}$ without affecting the result, i.e. $\widehat{\Sigma}_{1n}^2/\Sigma_n^2 = 1 + o_{\rm P}(1)$ and $\widehat{\Sigma}_{2n}^2/\Sigma_n^2 = 1 + o_{\rm P}(1)$.

Theorem 1 establishes that the IV estimator $\check{\alpha}$ is $\sqrt{n}$-consistent and asymptotically normal, with large-sample variance coinciding with the semi-parametric efficiency bound for the partially linear logistic regression model (see Section 5 for more discussion on this). The studentized estimator converges to the standard normal law, and the criterion function that this estimator minimized, when evaluated at the true value, converges to the standard chi-squares law with one degree of freedom. These results justify and imply the validity of the confidence regions $\mathcal{CR}_D$ and $\mathcal{CR}_I$ for α_0 proposed in Table 1. We note that these results are achieved despite possible model selection mistakes.

The following result derives similar properties for the post double selection estimator.

Theorem 2 (Robust Estimation and Inference based on Double Selection)**.** *Consider any triangular array of data $(y_i, d_i, x_i)_{i=1}^n$ (i.e., with all variables implicitly indexed by n, with $n = 1, 2, ...$) that obeys Condition L for all $n \geqslant 1$. Then, the post-double selection estimator $\check{\alpha}$ obeys as $n \to \infty$*

$$\Sigma_n^{-1}\sqrt{n}(\check{\alpha} - \alpha_0) = Z_n + o_{\rm P}(1), \quad Z_n \rightsquigarrow N(0,1),$$

where

$$Z_n := \frac{\Sigma_n}{\sqrt{n}}\sum_{i=1}^n (y_i - G(d_i\alpha_0 + x_i'\beta_0))z_{0i} \text{ and } \Sigma_n^2 := \bar{\rm E}[v_i^2]^{-1}.$$

Moreover, Σ_n^2 can be replaced by $\widehat{\Sigma}_{1n}^2 = \{\mathbb{E}_n[\check{w}_i d_i \widehat{z}_i]\}^{-1}\mathbb{E}_n[\{y_i - G(d_i\check{\alpha} + x_i'\check{\beta})\}^2\widehat{z}_i^2]\{\mathbb{E}_n[\check{w}_i d_i \widehat{z}_i]\}^{-1}$ or $\widehat{\Sigma}_{2n}^2 = \{\mathbb{E}_n[\check{w}_i(d_i, \check{x}_i')'(d_i, \check{x}_i')]\}_{11}^{-1}$ without affecting the result, i.e. $\widehat{\Sigma}_{1n}^2/\Sigma_n^2 = 1 + o_{\rm P}(1)$ and $\widehat{\Sigma}_{2n}^2/\Sigma_n^2 = 1 + o_{\rm P}(1)$, where $\check{w}_i = G(d_i\check{\alpha} + x_i'\check{\beta})\{1 - G(d_i\check{\alpha} + x_i'\check{\beta})\}$ and $\check{x}_i = x_{i,{\rm support}(\check{\beta})}$.

The theorem allows for the data-generating processes to change with n, in particular allowing sequences of regression models, with coefficients never perfectly distinguishable from zero, i.e. models where perfect model selection is not possible. In so doing, the results achieved in Theorems 1 and 2 imply validity uniformly over a large class of sparse models, which is referred to as "honesty" in the statistical literature on construction of confidence intervals. In what follows, we formalizes these assertions as corollaries.

Let $\mathcal{Q}_n$ denote a collection of distributions Q_n for the data $\{(y_i, d_i, x_i')'\}_{i=1}^n$ such that condition L hold for the given n. This is the collection of all approximately sparse models where the stated above sparsity conditions, moment conditions, and growth conditions hold. This collections expressly permits models to have non-zero coefficients, and thus does not impose the separation conditions (which we believe to be unreasonable in many applications.) For $Q_n \in \mathcal{Q}_n$, let the notation ${\rm P}_{Q_n}$ mean that under ${\rm P}_{Q_n}$, $\{(y_i, d_i, x_i')'\}_{i=1}^n$ is distributed according to Q_n.

Corollary 1 (Uniform $\sqrt{n}$-Rate of Consistency and Uniform Normality). *Let $\mathcal{Q}_n$ be the collection of all distributions of $\{(y_i, d_i, x_i')'\}_{i=1}^n$ for which Condition L is satisfied for the given $n \geqslant 1$. Then either optimal IV estimator or post-double selection estimators, $\check{\alpha}$, are $\sqrt{n}$-consistent and asymptotically normal uniformly over $\mathcal{Q}_n$, namely*

$$\lim_{n\to\infty} \sup_{Q_n \in \mathcal{Q}_n} \sup_{t\in\mathbb{R}} |{\rm P}_{Q_n}(\Sigma_n^{-1}\sqrt{n}(\check{\alpha} - \alpha_0) \leqslant t) - {\rm P}(N(0,1) \leqslant t)| = 0$$

Moreover, the result continues to hold if Σ_n^2 is replaced by any of the estimators $\widehat{\Sigma}_n^2$ specified in the statements of the preceding theorems.

Corollary 2 (**Uniformly Valid Confidence Regions**). *Let $\mathcal{Q}_n$ be the collection of all distributions of $\{(y_i, d_i, x_i')'\}_{i=1}^n$ for which Condition L is satisfied for the given $n \geqslant 1$. Then confidence regions $\mathcal{CR} \in \{\mathcal{CR}_D, \mathcal{CR}_I, \mathcal{CR}_{DS}\}$ are asymptotically valid uniformly in n, namely*

$$\lim_{n\to\infty} \sup_{Q_n \in \mathcal{Q}_n} |\mathrm{P}_{Q_n}(\alpha_0 \in \mathcal{CR}) - (1-\zeta)| = 0$$

All of the results are new under $s \to \infty$ and $p \to \infty$ asymptotics, and they are new even under the fixed s and p asymptotics.

Comment 3.1 (**Generalization to Approximately Sparse Models**). The results can also be shown to hold, with identical conclusions, in the class of approximately sparse models, following the analysis of the partially linear mean regression model in [8, 7]. Specifically, suppose that x_i is a dictionary with respect to some set of regressors z_i, i.e. $x_i = P(z_i)$, and

$$\mathrm{E}[y_i \mid d_i, x_i] = G(\alpha_0 d_i + x_i'\beta_0 + r_{yi}), \tag{3.10}$$

$$f_i d_i = f_i x_i'\theta_0 + r_{di} + v_i, \qquad \mathrm{E}[f_i v_i | x_i] = 0 \tag{3.11}$$

where r_{yi} and r_{di} are approximation errors, such that

$$\sqrt{\bar{\mathrm{E}}[r_{yi}^2]} \leqslant C\sqrt{s/n}, \quad \sqrt{\bar{\mathrm{E}}[r_{di}^2]} \leqslant C\sqrt{s/n}, \tag{3.12}$$

i.e., the $L^2(\mathrm{P})$ size of the approximation errors does not exceed the size of the estimation error of an oracle estimator. We can show that the results in Theorems 1 and 2 and Corollaries 1 and 2 continue to hold for this approximately sparse model. This means that the results are robust with respect to moderate violations of the sparsity assumption. ■

Comment 3.2 (**Generalization to Using Other Valid Instruments**). In the Appendix, we establish a more general result for the IV estimator based on any valid instrument, as defined in Comment 2.1, under high-level conditions. Given any valid instrument $(z_{0i})_{i=1}^n$, we show that

$$\{\bar{\mathrm{E}}[\sqrt{w_i} v_i z_{0i}]^{-1} \bar{\mathrm{E}}[w_i z_{0i}^2] \bar{\mathrm{E}}[\sqrt{w_i} v_i z_{0i}]^{-1}\}^{-1/2} \sqrt{n}(\check{\alpha} - \alpha_0) \rightsquigarrow N(0,1).$$

Therefore, the choice of instrument can be guided by efficiency considerations. ■

Comment 3.3 (**The Case of Testing $H_0 : \alpha_0 = 0$**). In some applications, the main goal is on testing if the policy variable d has an impact, i.e. the null hypothesis $H_0 : \alpha_0 = 0$. Under H_0, the conditional variance w_i of the outcome no longer varies with the policy variable. Specifically, when $\alpha_0 = 0$, we have

$$\mathrm{E}[\sqrt{w_i} v_i \mid x_i] = \sqrt{w_i}\mathrm{E}[v_i \mid x_i] = 0, i = 1, \ldots, n.$$

In the Logistic model associated with (2.1), we have $w_i > 0$ which makes the condition above equivalent to $\mathrm{E}[v_i \mid x_i] = 0$, $i = 1, \ldots, n$. Therefore, one can estimate θ_0 in (2.4) using $\widehat{f}_i = 1$. ■

4. Monte Carlo

Here we provide a simulation study of the finite sample properties of the proposed estimators and confidence intervals. We compare their performance with the post-naive selection estimator, which is defined by applying the logistic regression performed on the model selected by the ℓ_1-penalized logistic regression.

Our simulations are based on the model:

$$\mathrm{E}[y \mid d, x] = G(d\alpha_0 + x'\{c_y\nu_y\}), \qquad d = x'\{c_d\nu_d\} + \tilde{v},$$

where the coefficient vectors ν_y and ν_d are set to

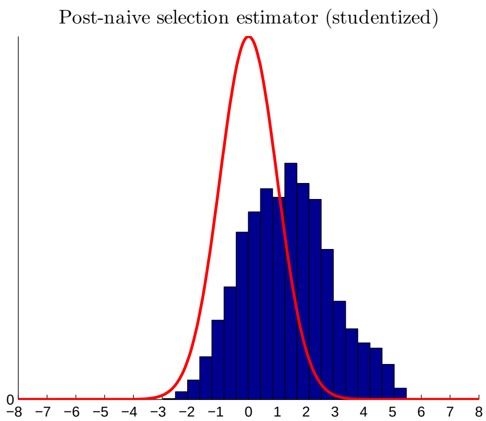
$$\nu_y = (1, 1/2, 1/3, 1/4, 1/5, 0, 0, 0, 0, 0, 1, 1/2, 1/3, 1/4, 1/5, 0, 0, \dots, 0)',$$
$$\nu_d = (1, 1/2, 1/3, 1/4, 1/5, 1/6, 1/7, 1/8, 1/9, 1/10, 0, 0, \dots, 0)',$$

$x = (1, z')'$ consists of an intercept and covariates $z \sim N(0, \Theta)$, and the error $\tilde{v}$ is i.i.d. as $N(0, 1)$. The dimension p of the covariates x is 250, and the sample size n is 200. In this set-up, the coefficients feature a declining pattern, with smallest coefficients that may be hard to detect from zero in the given sample size. Therefore, we expect that the ℓ_1-based model selectors will be making selection mistakes on variables with the smaller coefficients. (Additional simulations are provided in the Supplementary Material where we also consider an approximately sparse model for which all 250 coefficients are non-zero. Those experiments demonstrate that the results are robust with respect to moderate deviations away from exactly sparse models.)

The regressors are correlated with covariance $\Theta_{ij} = \rho^{|i-j|}$ and $\rho = 0.5$. The coefficient c_d is used to control the R^2, denoted R^2_d, in the equation relating main regressor to the controls, and c_y is used to control the R^2, denoted R^d_y, for the regression equation: $\tilde{y} - d\alpha_0 = x'\{c_y\nu_y\} + \epsilon$, where ϵ is logistic noise with unit variance. In the simulations, below we will use different values of α_0, c_y and c_d, which induce different data-generating processes (DGPs). In every repetition, we draw new errors v_i's and controls x_i's. The regression functions $x'(c_y\nu_y)$ and $x'(c_d\nu_d)$ are sparse. As we vary the coefficients c_y and c_d, we induce different amounts of "signal" strength, making it easier or harder for the Lasso-type methods to detect the controls with non-zero coefficients.

In Figure 1 we consider a DGP with $\alpha_0 = 0.2$ and $R^2_d = R^2_y = 0.75$, induced by setting $c_d = 1$ and $c_y = 0.75$. We performed 5000 monte-carlo simulations. The figure summarizes the performance and displays the distribution of the following estimators, which are centered by the true value α_0 and studentized by the standard deviation:

1. Post-naive selection estimator – estimator of α_0 based on logistic regression after the naive selection using ℓ_1-penalized logistic regression,
2. Optimal IV estimator – estimator of α_0 based on the instrumental logistic regression with the optimal instrument, as defined in Table 1,
3. Post-double selection estimator – estimator of α_0 based on the logistic regression after double selection, as defined in Table 2.



estimator	bias	variance	rmse	rp(0.05)
post-naive selection	.173	.041	.267	.350
optimal IV	.038	.036	.193	.043
post-double selection	.024	.039	.199	.051

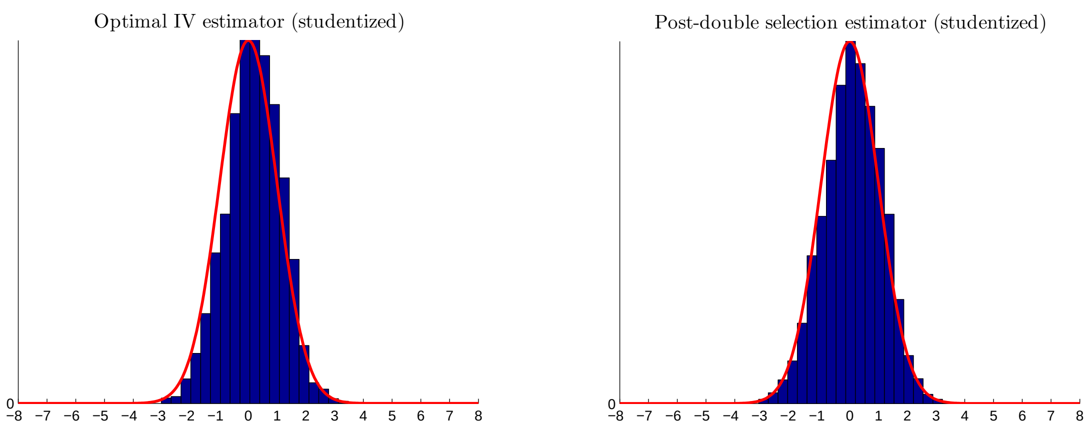


FIGURE 1. The top right panel display bias, variance, RMSE, and rejection frequency for a 5%-level test. The figures display the distribution of the post-naive selection estimator (top left panel) and the two proposed estimators optimal instruments (bottom left panel) and double selection (bottom right).

The optimal IV estimator and the post-double selection estimator have distribution approximately centered at the true value, with distribution agreeing closely with the standard normal distribution. They have low biases, low (compared to naive) root mean squared errors, and confidence regions have rejection rates close to the nominal level of 5%. This good performance is well aligned with our theoretical results that we have developed in the previous section. In sharp contrast, the distribution of the post naive selection estimator seems to deviate substantially from the normal distribution. This estimator exhibits large bias and high root mean squared error compared to the former procedures. This occurs because in this DGP, perfect selection is not achieved, and the resulting "moderate" selection mistakes create a large omitted variable bias. Thus, if we use naive post selection estimator with the standard normal distribution for constructing confidence intervals or performing hypothesis testing, we shall end up with rather misleading inference. This poor performance is well aligned with theoretical predictions of [16, 15, 14] (given in the context of a linear model).

We now examine the performance more systematically by varying

$$(R_d^2, R_y^d) \in \{0, .1, .2, .3, .4, .5, .6, .7, .8, .9\}^2 \text{ and } \alpha_0 \in \{0, .25, .5\}. \tag{4.13}$$

This gives us 300 different data-generating processes, for each of which we performed 1000 Monte-Carlo simulations. In Figures 2-4 display the rejection frequencies of confidence regions with (nominal) significance level of 0.05 and in Figure 5 we display the root mean squared errors of the estimators of α_0. The goal of this exercise is to verify numerically how good the uniformity claims derived in Corollaries 1 and 2 are, and also confirm that the previous conclusions continue to hold across a wide set of data-generating processes.

In Figures 2-4 we consider the rejection (non-coverage) frequencies of confidence regions based on: post naive selection logistic estimator[1], optimal IV ($\mathcal{CR}_D$ and $\mathcal{CR}_I$) and double selection ($\mathcal{CR}_{DS}$). These figures illustrates the uniformity properties of the confidence regions based on the discussed estimators. The ideal figure would be a flat surface with the rejection frequency of the true value equal to the nominal level of .05. The confidence regions based on the post-naive selection perform very poorly, and deviate strongly away from the ideal level of .05 throughout large parts of the model space (induced by (4.13)). In contrast, the confidence regions based on optimal IV and double selection seem to be substantially closer to the ideal level, which is in line with our theoretical results in Section 3. The post-double selection method seems to clearly outperform the optimal IV method (this can be seen by looking at the rejection rates and the RMSE for the case with $\alpha = .5$, where optimal IV procedure tends to perform noticeably worse.)[2] Thus, based on the theoretical results and on the Monte-Carlo results, we recommend the use of the post-double selection estimator over the optimal IV estimator and, certainly, over the post-naive selection estimator.

5. Discussion

5.1. **Relation between Double Selection and Optimal Instrument.** In this section, we provide a more formal connection between the two proposed methods. It turns out that the construction of the double selection estimator implicitly approximates the optimal instrument $z_{0i} = v_i/\sqrt{w_i}$. This occurs because the model selection procedure in Step 2 associated with (2.4) allows the estimator to achieve uniformity properties. To see that, using the notation in Table 1 where $\widehat{\beta}$, $\widehat{\theta}$ and $\widetilde{\theta}$ are defined, let $\widehat{T}^*$ denote the variables selected in Step 1 and 2, $\widehat{T}^* = \text{support}(\widehat{\beta}) \cup \text{support}(\widehat{\theta})$. By the first order conditions of the double selection logistic regression problem we have

$$\mathbb{E}_n[\{y_i - G(d_i\check{\alpha} + x_i'\check{\beta})\}(d_i,\ x'_{i\widehat{T}^*})'] = 0$$

[1]This region is given by $\{|\alpha - \widetilde{\alpha}| \leqslant \{\mathbb{E}_n[\widehat{w}_i(d_i, x'_{i\text{support}(\widetilde{\beta})})'(d_i, x'_{i\text{support}(\widetilde{\beta})})]\}_{11}^{-1/2}\Phi^{-1}(1-\xi/2)/\sqrt{n}\}$ where $(\widetilde{\alpha}, \widetilde{\beta})$ is the post naive selection logistic estimator.

[2]In the next section we discuss an interpretation of the double selection procedure as an iterated version of the optimal IV procedure, which might provide some intuition for its better finite-sample performance. Note that the two procedures are first-order asymptotically equivalent.

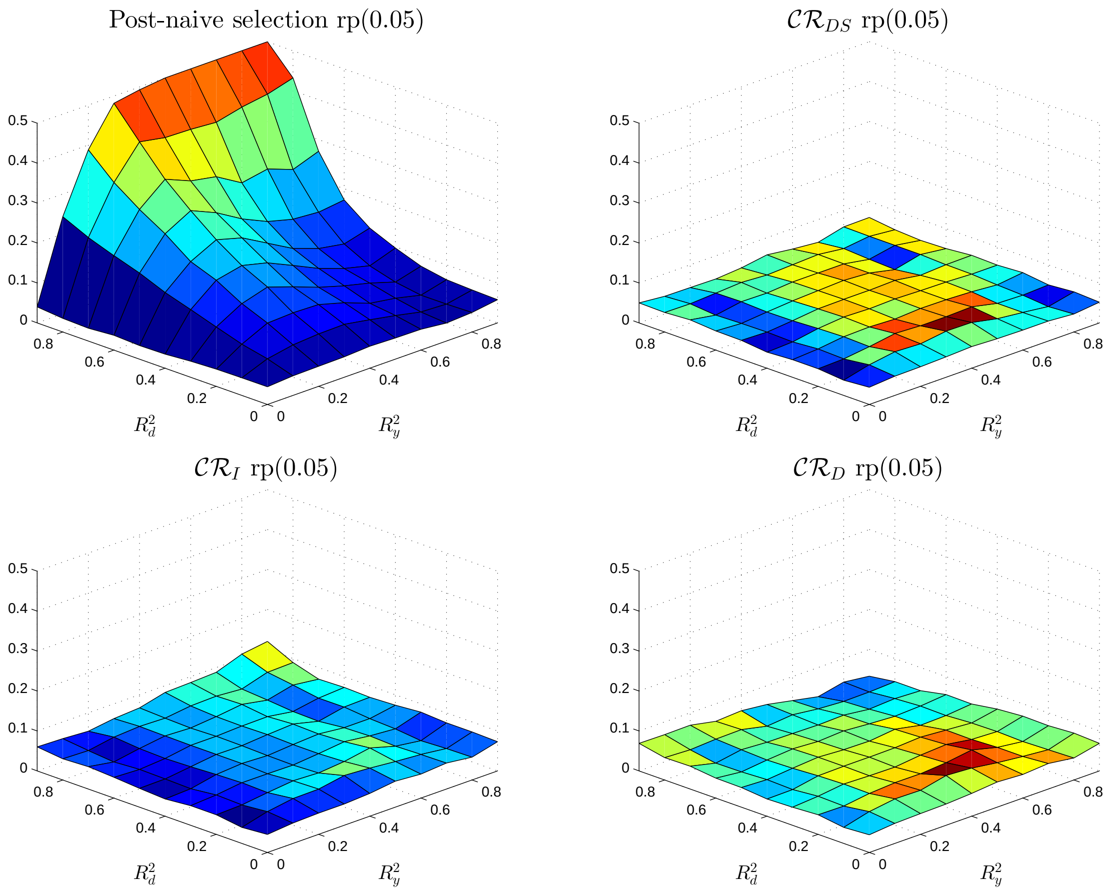


FIGURE 2. The figures display the rp(0.05) of the confidence region based on post-naive selection and the proposed confidence regions based on optimal instrument ($\mathcal{CR}_D$ and $\mathcal{CR}_I$) and double selection (CR_{DS}). There are a total of 100 different designs with $\alpha_0 = 0$. The results are based on 1000 replications for each design.

which creates an orthogonal relation to any linear combination of $(d_i,\ x'_{i\widehat{T}^*})'$. In particular, by taking the linear combination $(d_i,\ x'_{i\widehat{T}^*})(1, -\widetilde{\theta}')' = d_i - x_i'\widetilde{\theta} = \widehat{z}_i$, we have

$$\mathbb{E}_n[\{y_i - G(d_i\check{\alpha} + x_i'\check{\beta})\}\widehat{z}_i] = 0.$$

Therefore the post-double selection estimator $\check{\alpha}$ minimizes

$$\widetilde{L}_n(\alpha) = \frac{\|\mathbb{E}_n[\{y_i - G(d_i\alpha + x_i'\check{\beta})\}\widehat{z}_i]\|^2}{\mathbb{E}_n[\{y_i - G(d_i\alpha + x_i'\check{\beta})\}^2\widehat{z}_i^2]},$$

where $\widehat{z}_i$ is the instrument implicitly created. Thus, the double selection estimator can be seen as an iterated version of the method based on instruments where $\widetilde{\beta}$ is replaced with $\check{\beta}$. Although their first order asymptotic properties coincide, in finite sample, the double selection method seems to obtain better estimates of the weights leading a more robust performance.

5.2. **Relation to Neyman's $C(\alpha)$ test.** Next we discuss connections between the proposed approach and Neyman's $C(\alpha)$ test [19, 20] (here we draw on the discussion in [6]). For the sake of exposition,

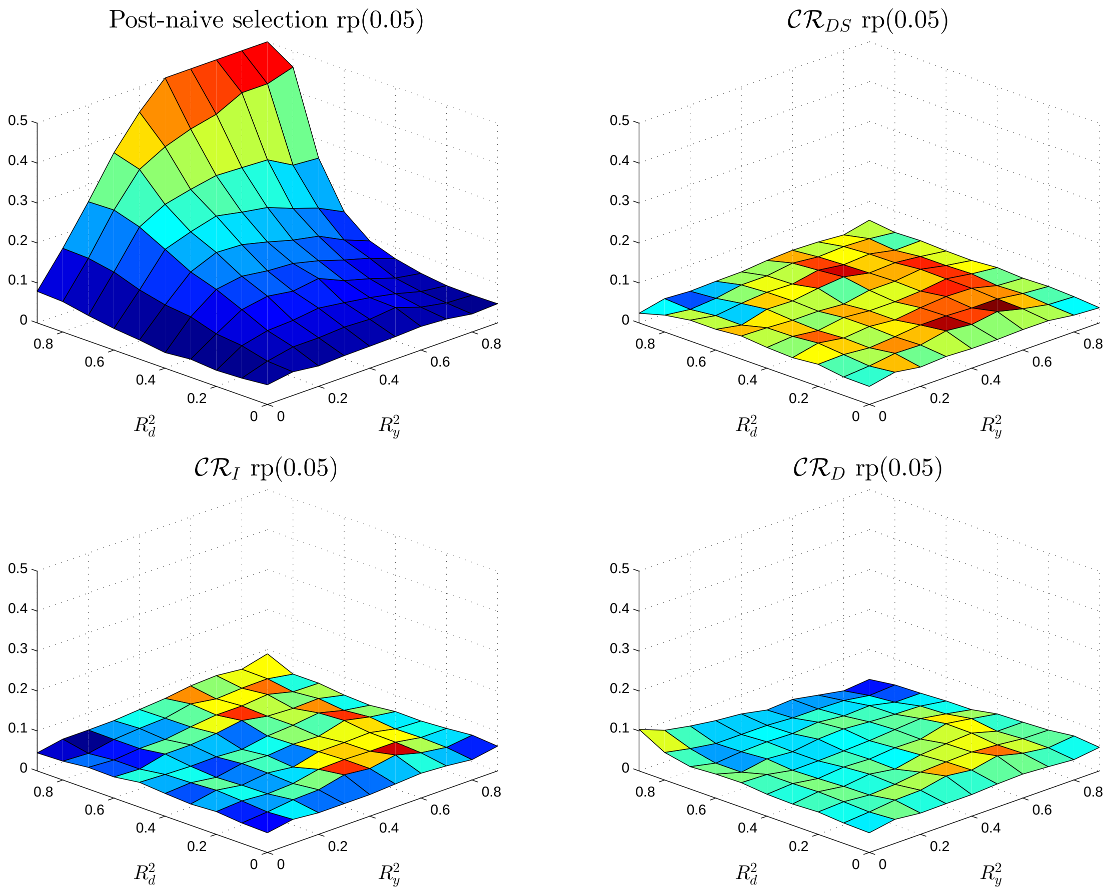


FIGURE 3. The figures display the rp(0.05) of the confidence region based on post-naive selection and the proposed confidence regions based on optimal instrument ($\mathcal{CR}_D$ and $\mathcal{CR}_I$) and double selection (CR_{DS}). There are a total of 100 different designs with $\alpha_0 = 0.25$. The results are based on 1000 replications for each design.

we assume the instruments are known. As stated in (2.2) and (2.3) we need to construct instruments satisfying the two equations:

$$\mathrm{E}[\{y_i - G(d_i\alpha_0 + x_i'\beta_0)\}z_{0i}] = 0 \quad \text{and} \quad \frac{\partial}{\partial\beta}\mathrm{E}[\{y_i - G(d_i\alpha_0 + x_i'\beta)\}z_{0i}]\bigg|_{\beta=\beta_0} = \mathrm{E}[w_i z_{0i} x_i] = 0.$$

Thanks to the second condition, using the first equation, we can construct regular, $\sqrt{n}$-consistent estimators of α_0, despite the fact that nonregular, non $\sqrt{n}$-consistent estimator for β_0 are being used to cope with high-dimensionality; in particular, regularized or post model selection estimators can be used as estimators of β_0. Neyman's $C(\alpha)$ test was motivated by the same idea which motivates the use of the term "Neymanization" to describe such procedure (see also [6]). Although there will be many instruments z_{0i} that can achieve the property stated above, the choice $z_{0i} = v_i/\sqrt{w_i}$ proposed in Section 2 is optimal as it minimizes the asymptotic variance of the resulting IV estimators.

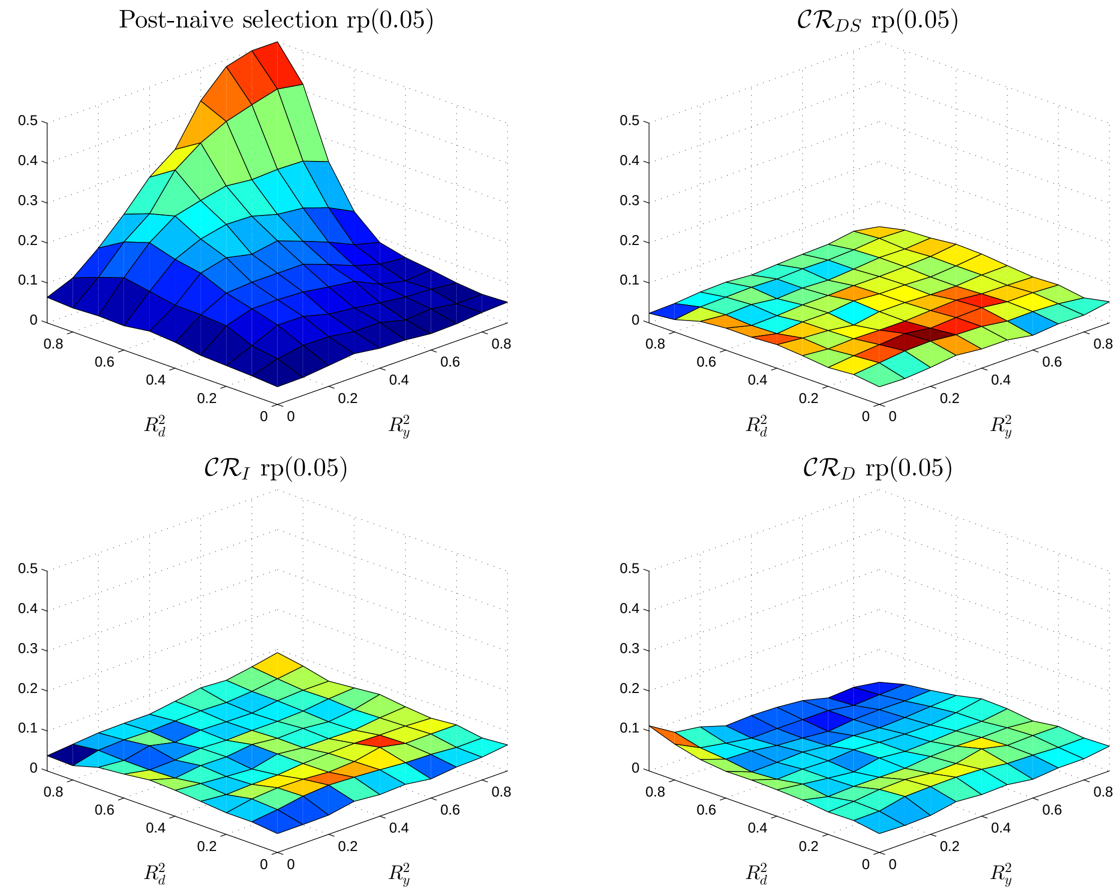


FIGURE 4. The figures display the rp(0.05) of the confidence region based on post-naive selection and the proposed confidence regions based on optimal instrument ($\mathcal{CR}_D$ and $\mathcal{CR}_I$) and double selection ($\mathcal{CR}_{DS}$). There are a total of 100 different designs with $\alpha_0 = 0.5$. The results are based on 1000 replications for each design.

Generally, valid (but not necessarily optimal) instruments can be constructed by generalizing the weighted equation (2.4) to

$$f_i d_i = f_i m_0(x_i) + \tilde{v}_i, \quad \mathrm{E}[f_i \tilde{v}_i | x_i] = 0, \tag{5.14}$$

where $f_i = f(d_i, z_i)$ is a nonnegative weight, and setting the instrument as $z_{0i} := f_i \tilde{v}_i / w_i$. Because of the zero-mean condition in (5.14), and provided that $m_0 \in \mathcal{H}$, the function $m_0(x_i)$ in (5.14) is the solution of the following weighted least squares problem

$$\min_{h \in \mathcal{H}} \bar{\mathrm{E}} \left[f_i^2 \{d_i - h(x_i)\}^2 \right], \tag{5.15}$$

where $\mathcal{H}$ denotes the set of measurable functions h satisfying $\mathrm{E}[f_i^2 h^2(x_i)] < \infty$ for each i. In the current high-dimensional setting, it is assumed that $m_0(x_i)$ can be written as a sparse combination of the controls, namely $m_0(x_i) = x_i' \theta_0$ with $\|\theta_0\|_0 \leqslant s$, so that

$$f_i d_i = f_i x_i' \theta_0 + \tilde{v}_i, \quad \mathrm{E}[f_i \tilde{v}_i | x_i] = 0. \tag{5.16}$$

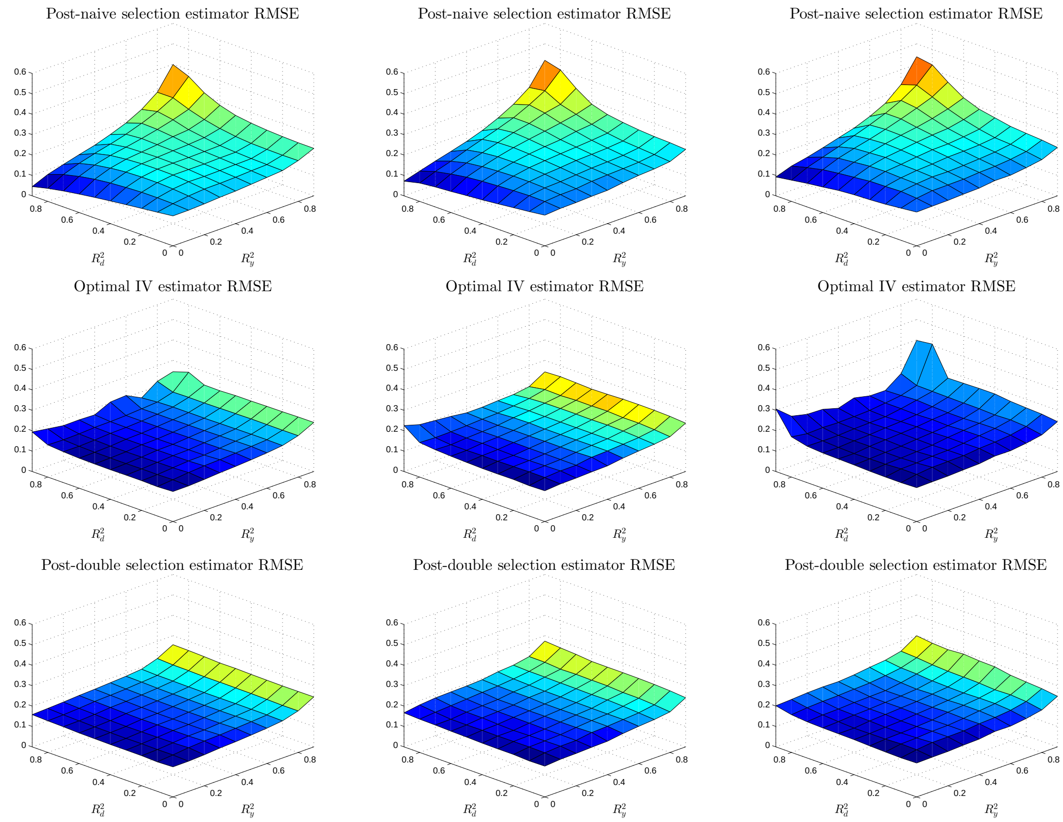


FIGURE 5. The figures display the RMSE of the post-naive selection estimator and the proposed estimators based on optimal IV and post-double selection. The left column refers to $\alpha_0 = 0$, the middle column refers to $\alpha_0 = 0.25$, and the right column refers to $\alpha_0 = 0.5$. There are a total of 100 different designs for each value of α_0. The results are based on 1000 replications for each design.

This permits the use of Lasso or Post-Lasso to estimate θ_0 which in turn can be used to construct an estimate of z_{0i}. Naturally, if the function m_0 satisfies different structured properties that could motivate different estimators (for example, we can use ridge estimators if the m_0 is "dense" with respect to x).

Our technical results establish that, uniformly over $\{\alpha : \sqrt{n}|\alpha - \alpha_0| \leqslant C\}$,

$$\mathbb{E}_n[\{y_i - G(d_i\alpha + x_i'\widehat{\beta})\}z_{0i}] - \mathbb{E}_n[\{y_i - G(d_i\alpha + x_i'\beta_0)\}z_{0i}] = o_{\mathrm{P}}(n^{-1/2}), \tag{5.17}$$

for the estimators $\widehat{\beta}$ proposed in this work. This does not require $\widehat{\beta}$ to converge at root-n rate to β_0 (which generically is not achievable in the current setting) though we do impose the sparsity condition $s^2 \log p/n \to 0$ (or stronger) to guarantee that $\|\widehat{\beta} - \beta\| = o_{\mathrm{P}}(n^{-1/4})$. Equation (5.17) implies that the empirical estimating equations behave as if β_0 was used instead of $\widehat{\beta}$. Hence, for estimation, we can use the instrumental logistic regression estimator, namely $\check{\alpha}$ as a minimizer of the statistic

$$nL_n(\alpha) = \|\sqrt{n}\mathbb{E}_n[\{y_i - G(d_i\alpha + x_i'\widehat{\beta})\}z_{0i}]\|^2 \ \ / \ \ \mathbb{E}_n[\{y_i - G(d_i\alpha + x_i'\widehat{\beta})\}^2 z_{0i}^2].$$

From (5.16) we have that $\theta_0 = \bar{\mathrm{E}}[f_i^2 x_i x_i']^- \bar{\mathrm{E}}[f_i^2 d_i x_i]$, where A^- denotes a generalized inverse of A. Letting $\widehat{\varepsilon}_i(\alpha) = y_i - G(d_i\alpha + x_i'\widehat{\beta})$ and

$$z_{0i} = f_i \tilde{v}_i / w_i = (f_i^2/w_i) d_i - (f_i^2/w_i) x_i' \bar{\mathrm{E}}[f_i^2 x_i x_i']^- \bar{\mathrm{E}}[f_i^2 d_i x_i],$$

$nL_n(\alpha)$ can be rewritten as a (perhaps) familiar version of Neyman's $C(\alpha)$ statistic

$$nL_n(\alpha) = \frac{\|\sqrt{n}\{\mathbb{E}_n[\widehat{\varepsilon}_i(\alpha)(f_i^2/w_i)d_i] - \mathbb{E}_n[\widehat{\varepsilon}_i(\alpha)(f_i^2/w_i)x_i']\bar{\mathrm{E}}[f_i^2 x_i x_i']^- \bar{\mathrm{E}}[f_i^2 d_i x_i]\}\|^2}{\mathbb{E}_n[\widehat{\varepsilon}_i^2(\alpha) z_{0i}^2]}.$$

Thus, our IV estimator minimizes a Neyman's $C(\alpha)$ statistic for testing point hypotheses about α. Hence our construction builds on the classical ideas of Neyman for dealing with (hard-to-estimate) nuisance parameters.

An estimator $\check{\alpha}$ that minimizes the criterion nL_n up to a $o_{\mathrm{P}}(1)$ term satisfies

$$\tilde{\Sigma}_n^{-1}\sqrt{n}(\check{\alpha} - \alpha_0) \rightsquigarrow N(0,1), \quad \tilde{\Sigma}_n^2 = \bar{\mathrm{E}}[w_i d_i z_{0i}]^{-2} \bar{\mathrm{E}}[w_i z_{0i}^2].$$

It is not difficult to check that using $f_i = \sqrt{w_i}$ leads to the smallest possible value $\bar{\mathrm{E}}[v_i^2]^{-1}$ of $\tilde{\Sigma}_n^2$. Therefore, $z_{0i} = v_i/\sqrt{w_i}$ is the optimal instrument among all instruments that can be derived by the preceding approach. Using the optimal instrument translates into more precise estimators, smaller confidence regions, and better power for testing based on either $\check{\alpha}$ or nL_n.

5.3. **Relation to Minimax Efficiency.** Next we consider a connection to the (local) minimax efficiency analysis from the semiparametric literature, where we follow the discussion in [6]. Our model is a special case of a partially linear logistic model, and [11] derives an efficient score function for the latter:

$$S_i = \{y_i - G(d_i\alpha_0 + x_i'\beta_0)\}\{d_i - m_0^*(x_i)\},$$

where

$$m_0^*(x_i) = \frac{\mathrm{E}[w_i d_i | x_i]}{\mathrm{E}[w_i | x_i]}.$$

We note that $m_0^*(x_i)$ is $m_0(x_i)$ in (5.14) induced by the weight $f_i = \sqrt{w_i}$. Thus, the efficient score function can be reexpressed as:

$$S_i = \{y_i - G(d_i\alpha_0 + x_i'\beta_0)\} v_i/\sqrt{w_i},$$

where v_i is defined via (5.14) . Using this score leads to the same estimating equations as those constructed above using Neymanization (with an optimal instrument). It follows that the estimator based on the instrument $z_{0i} = v_i/\sqrt{w_i}$ is efficient in the local minimax sense (see Theorem 18.4 in [11]), and inference about α_0 based on this estimator provides best minimax power against local alternatives (see Theorem 18.12 in [11]).

The preceding claim is formal provided that the least favorable submodels are permitted as deviations within the overall set of potential models $\mathcal{Q}_n$ (defined similarly to Corollary 1). Specifically, given a law Q_n, there should be a suitable neighborhood $\mathcal{Q}_n^\delta$ of Q_n such that $Q_n \in \mathcal{Q}_n^\delta \subset \mathcal{Q}_n$. For that, we assume $m_0^*(x_i) = x_i'\theta_0$ and consider a collection of models indexed by $t = (t_1, t_2)$ satisfying:

$$\mathrm{E}[y_i \mid d_i, x_i] = G(d_i\{\alpha_0 + t_1\} + x_i'\{\beta_0 + t_2\theta_0\}), \quad \|t\| \leqslant \delta, \tag{5.18}$$

$$\sqrt{w_i} d_i = \sqrt{w_i} x_i'\theta_0 + v_i, \quad \mathrm{E}[\sqrt{w_i} v_i | x_i] = 0, \tag{5.19}$$

where $\|\beta_0\|_0 + \|\theta_0\|_0 \leqslant s$ and Condition L as in Section 3 hold. By construction, the model associated with $t = 0$ generates precisely the model Q_n. As t varies within a δ-ball, we generate the set of models $\mathcal{Q}_n^\delta$ that contains the least favorable deviations, and which still belong to $\mathcal{Q}_n$. As shown in [11], S_i is the efficient score for such parametric submodel so we cannot have a better regular estimator than the estimator whose influence function is $\Sigma_n S_i$. Because the set of models $\mathcal{Q}_n$ contains $\mathcal{Q}_n^\delta$, all the formal conclusions about (local minimax) optimality of the proposed estimators hold from theorems cited above (using subsequence arguments to handle models changing with n).

Appendix A. Generic Instrumental Logistic Regression with Estimated Data

Let $(d, x) \in \mathcal{D} \times \mathcal{X}$. In this section for $\tilde{h} = (\tilde{\beta}, \tilde{z})$, where $\tilde{z}$ is a function on $(d, x) \mapsto \tilde{z}(d, x)$ we write

$$\psi_{\tilde{\alpha},\tilde{h}}(y_i, d_i, x_i) = \psi_{\tilde{\alpha},\tilde{\beta},\tilde{z}}(y_i, d_i, x_i) = \{y_i - G(x_i'\tilde{\beta} + d_i\alpha)\}\tilde{z}(d_i, x_i).$$

For a fixed $\tilde{\alpha} \in \mathbb{R}$, $\tilde{\beta} \in \mathbb{R}^p$, $\tilde{z} : \mathcal{D} \times \mathcal{X} \to \mathbb{R}$, and $\tilde{h} = (\tilde{\beta}, \tilde{z})$, we define

$$\Gamma(\tilde{\alpha}, \tilde{h}) := \bar{\mathrm{E}}[\psi_{\tilde{\alpha},\tilde{h}}(y_i, d_i, x_i)]$$

For notational convenience we let $\tilde{z}_i = \tilde{z}(d_i, x_i)$, $h_0 = (\beta_0, z_0)$ and $\widehat{h} = (\widehat{\beta}, \widehat{z})$. The partial derivative of Γ with respect to α at $(\tilde{\alpha}, \tilde{h})$ is denoted by $\Gamma_1(\tilde{\alpha}, \tilde{h})$ and the directional derivative with respect to $[\widehat{h} - h_0]$ at $(\tilde{\alpha}, \tilde{h})$ is denoted as

$$\Gamma_2(\tilde{\alpha}, \tilde{h})[\widehat{h} - h_0] = \lim_{t \to 0} \frac{\Gamma(\tilde{\alpha}, \tilde{h} + t[\widehat{h} - h_0]) - \Gamma(\tilde{\alpha}, \tilde{h})}{t}.$$

We assume that the estimated vector $\widehat{\beta}$ and the estimated function $\widehat{z}$ satisfy the following condition.

Condition *ILOG*. For some sequences $\delta_n \to 0$ and $\Delta_n \to 0$ with probability at least $1 - \Delta_n$:
(i) $\{\alpha : |\alpha - \alpha_0| \leqslant n^{-1/2}/\delta_n\} \subset \mathcal{A}$, where $\mathcal{A}$ is a (possibly random) compact interval;
(ii) $\bar{\mathrm{E}}[w_i z_{0i} \mid x_i] = 0$, $|\bar{\mathrm{E}}[w_i d_i z_{0i}]| \geqslant c > 0$, $\mathrm{E}[z_{0i}^4 + d_i^4] \leqslant C$, $\max_{i \leqslant n} \mathrm{E}[|z_{0i}|] \leqslant C$,
(iii) the estimated quantities $\widehat{h} = (\widehat{\beta}, \widehat{z})$

$$\begin{gathered} \max_{i \leqslant n} \{1 + \mathrm{E}[|\widehat{z}_i - z_{0i}|]\}^{1/2} \|x_i'(\widehat{\beta} - \beta_0)\|_{2,n} \leqslant \delta_n n^{-1/4}, \\ \{\bar{\mathrm{E}}[(\widehat{z}_i - z_{0i})^2]\}^{1/2} \leqslant \delta_n, \quad \|x_i'(\widehat{\beta} - \beta_0)\|_{2,n} \cdot \{\bar{\mathrm{E}}[(\widehat{z}_i - z_{0i})^2]\}^{1/2} \leqslant \delta_n n^{-1/2}, \end{gathered} \tag{A.20}$$

$$\sup_{\alpha \in \mathcal{A}} \left| (\mathbb{E}_n - \bar{\mathrm{E}}) \left[\psi_{\alpha,\widehat{h}}(y_i, d_i, x_i) - \psi_{\alpha,h_0}(y_i, d_i, x_i) \right] \right| \leqslant \delta_n \; n^{-1/2} \tag{A.21}$$

$$|\check{\alpha} - \alpha_0| \leqslant \delta_n \quad \text{and} \quad \left| \; \mathbb{E}_n[\psi_{\check{\alpha},\widehat{h}}(y_i, d_i, x_i)] \; \right| \leqslant \delta_n \; n^{-1/2}. \tag{A.22}$$

(iv) $\|(1 \vee |d_i|)(\widehat{z}_i - z_{0i})\|_{2,n} \leqslant \delta_n$ and $\|\{x_i'(\widehat{\beta} - \beta_0)\}^2\|_{2,n} \leqslant \delta_n$.

Lemma 1. *Under Condition ILOG(i,ii,iii) we have*

$$\{\mathrm{E}[w_i z_{0i}^2]\}^{-1/2} \bar{\mathrm{E}}[w_i d_i z_{0i}] \sqrt{n}(\check{\alpha} - \alpha_0) = \{\mathrm{E}[w_i z_{0i}^2]\}^{-1/2} \mathbb{G}_n(\psi_{\alpha_0,h_0}(y_i, d_i, x_i)) + o_P(1)$$

$$\textit{and} \quad \{\bar{\mathrm{E}}[w_i d_i z_{0i}]^{-1} \bar{\mathrm{E}}[w_i z_{0i}^2] \bar{\mathrm{E}}[w_i d_i z_{0i}]^{-1}\}^{-1/2} \sqrt{n}(\check{\alpha} - \alpha_0) \rightsquigarrow N(0, 1).$$

Moreover, if additionally ILOG(iv) holds, we have

$$n L_n(\alpha_0) \rightsquigarrow \chi^2(1)$$

and the variance estimator is consistent, namely

$$\mathbb{E}_n[\widehat{w}_i d_i \widehat{z}_i]^{-1}\mathbb{E}_n[\{y_i - G(x_i'\widehat{\beta} + d_i\check{\alpha})\}^2 \widehat{z}_i^2]\mathbb{E}_n[\widehat{w}_i d_i \widehat{z}_i]^{-1} = \bar{\mathrm{E}}[w_i d_i z_{0i}]^{-1}\bar{\mathrm{E}}[w_i z_{0i}^2]\bar{\mathrm{E}}[w_i d_i z_{0i}]^{-1} + o_P(1).$$

Proof of Lemma 1. Steps 1-4 we use ILOG(i-iii). In Steps 5 and 6 we will also use ILOG(iv).

Step 1. (Main Step for Normality) We have

$$\begin{aligned}\overbrace{\mathbb{E}_n[\psi_{\check{\alpha},\widehat{h}}(y_i,d_i,x_i)]}^{(0)} &= \mathbb{E}_n[\psi_{\alpha_0,h_0}(y_i,d_i,x_i)] + \mathbb{E}_n[\psi_{\check{\alpha},\widehat{h}}(y_i,d_i,x_i) - \psi_{\alpha_0,h_0}(y_i,d_i,x_i)]\\ &= \underbrace{\mathbb{E}_n[\psi_{\alpha_0,h_0}(y_i,d_i,x_i)]}_{(I)} + \underbrace{\Gamma(\check{\alpha},\widehat{h})}_{(II)} + \underbrace{n^{-1/2}\mathbb{G}_n(\psi_{\check{\alpha},\widehat{h}} - \psi_{\check{\alpha},h_0})}_{(III)} + \underbrace{n^{-1/2}\mathbb{G}_n(\psi_{\check{\alpha},h_0} - \psi_{\alpha_0,h_0})}_{(IV)}.\end{aligned}$$

By Condition ILOG(iii), (A.22), with probability at least $1-\Delta_n$ we have $|(0)| \lesssim \delta_n n^{-1/2}$.

By Step 2 below we have $|(II) + \bar{\mathrm{E}}[w_i d_i z_{0i}](\check{\alpha} - \alpha_0)| \lesssim_P \delta_n n^{-1/2} + \delta_n|\check{\alpha} - \alpha_0|$.

By Condition ILOG(iii), (A.21), with probability at least $1-\Delta_n$ we have $|(III)| \lesssim \delta_n n^{-1/2}$.

To control (IV) note that

$$|\psi_{\alpha,h_0}(y_i,d_i,x_i) - \psi_{\alpha_0,h_0}(y_i,d_i,x_i)| \leqslant |G(d_i\alpha + x_i'\beta_0) - G(d_i\alpha_0 + x_i'\beta_0)| \cdot |z_{0i}| \leqslant |\alpha - \alpha_0| \cdot |d_i z_{0i}|.$$

By Condition ILOG(iii), (A.22), we have $|\check{\alpha} - \alpha_0| \leqslant \delta_n$ so that $\bar{\mathrm{E}}[\{\psi_{\alpha,h_0}(y_i,d_i,x_i) - \psi_{\alpha_0,h_0}(y_i,d_i,x_i)\}^2] \leqslant |\alpha - \alpha_0|^2\bar{\mathrm{E}}[d_i^2 z_{0i}^2]$ and using a version of Theorem 2.14.1 in [30] we have

$$\sup_{|\alpha-\alpha_0|\leqslant\delta_n} \mathbb{E}_n[\{\psi_{\alpha,h_0}(y_i,d_i,x_i) - \psi_{\alpha_0,h_0}(y_i,d_i,x_i)\}^2] \leqslant \delta_n^2\mathbb{E}_n[d_i^2 z_{0i}^2] \lesssim_P \delta_n^2\bar{\mathrm{E}}[d_i^2 z_{0i}^2]$$

from concentration of measure and Condition ILOG(ii). Because of these relations and the maximal inequality in Lemma 5, we have

$$\begin{aligned}|(IV)| &\lesssim_P \sup_{|\alpha-\alpha_0|\leqslant\delta_n}\left|n^{-1/2}\mathbb{G}_n(\psi_{\alpha,h_0} - \psi_{\alpha_0,h_0})\right|\\ &\lesssim_P n^{-1/2}\sup_{|\alpha-\alpha_0|\leqslant\delta_n}|\alpha - \alpha_0|\bar{\mathrm{E}}[d_i^2 z_{0i}^2] \lesssim \delta_n n^{-1/2}\end{aligned} \tag{A.23}$$

Combining the bounds for (0), (II)-(IV) above we have

$$\bar{\mathrm{E}}[w_i d_i z_{0i}](\check{\alpha} - \alpha_0) = \mathbb{E}_n[\psi_{\alpha_0,h_0}(y_i,d_i,x_i)] + O_P(\delta_n n^{-1/2}) + O_P(\delta_n)|\check{\alpha} - \alpha_0|$$

which establish the first assertion since $|\bar{\mathrm{E}}[w_i d_i z_{0i}]| \geqslant c > 0$ is bounded away from zero. The second assertion follows since $\bar{\mathrm{E}}[\psi_{\alpha_0,h_0}(y_i,d_i,x_i)] = 0$ and $\bar{\mathrm{E}}[w_i z_{0i}^2] \leqslant C$, by the Lyapunov CLT, we have

$$(I) = \mathbb{E}_n[\psi_{\alpha_0,h_0}(y_i,d_i,x_i)] \rightsquigarrow N(0, \bar{\mathrm{E}}[w_i z_{0i}^2]).$$

Step 2. (Bounding $\Gamma(\alpha,\widehat{h})$ for $|\alpha - \alpha_0| \leqslant \delta_n$ which covers (II)) We have

$$\begin{aligned}\Gamma(\alpha,\widehat{h}) &= \Gamma(\alpha,h_0) + \Gamma(\alpha,\widehat{h}) - \Gamma(\alpha,h_0)\\ &= \Gamma(\alpha,h_0) + \{\Gamma(\alpha,\widehat{h}) - \Gamma(\alpha,h_0) - \Gamma_2(\alpha,h_0)[\widehat{h} - h_0]\} + \Gamma_2(\alpha,h_0)[\widehat{h} - h_0].\end{aligned} \tag{A.24}$$

Because $\Gamma(\alpha_0,h_0) = 0$, by Taylor expansion there is some $\tilde{\alpha} \in [\alpha_0,\alpha]$ such that

$$\Gamma(\alpha,h_0) = \Gamma_1(\tilde{\alpha},h_0)(\alpha - \alpha_0) = \{\Gamma_1(\alpha_0,h_0) + \eta_n\}(\alpha - \alpha_0)$$

where $|\eta_n| \leqslant \delta_n \bar{\mathrm{E}}[|d_i^2 z_{0i}|]$ by relation (A.31) in Step 4.

Combining the argument above with relations (A.26), (A.27) and (A.29) in Step 3 below we have

$$\begin{aligned}\Gamma(\alpha,\widehat{h}) &= \Gamma_2(\alpha_0,h_0)[\widehat{h}-h_0] + \Gamma(\alpha_0,h_0) + \{\Gamma_1(\alpha_0,h_0) + O(\delta_n \bar{\mathrm{E}}[|d_i^2 z_{0i}|])\}(\alpha-\alpha_0) + O(\delta_n n^{-1/2})\\ &= \Gamma_1(\alpha_0,h_0)(\alpha-\alpha_0) + O(\delta_n|\alpha-\alpha_0|\bar{\mathrm{E}}[|d_i^2 z_{0i}|] + \delta_n n^{-1/2})\end{aligned} \tag{A.25}$$

Step 3. (Relations for Γ_2) The directional derivative Γ_2 with respect the direction $\widehat{h}-h_0$ at a point $\tilde{h} = (\tilde{\beta},\tilde{z})$ is given by

$$\begin{aligned}\Gamma_2(\alpha,\tilde{h})[\widehat{h}-h_0] &= -\bar{\mathrm{E}}[G'(d_i\alpha + x_i'\tilde{\beta})\tilde{z}_i x_i'\{\widehat{\beta}-\beta_0\}] +\\ &\quad + \bar{\mathrm{E}}[\{G(d_i\alpha_0 + x_i'\beta_0) - G(d_i\alpha + x_i'\tilde{\beta})\}\{\widehat{z}_i - z_{0i}\}].\end{aligned}$$

Note that when Γ_2 is evaluated at (α_0,h_0) we have

$$\Gamma_2(\alpha_0,h_0)[\widehat{h}-h_0] = -\bar{\mathrm{E}}[w_i z_{0i} x_i'(\widehat{\beta}-\beta_0)] = 0 \tag{A.26}$$

because of the orthogonality condition $\bar{\mathrm{E}}[w_i z_{0i} \mid x_i] = 0$ in Condition ILOG(ii), and by definition $w_i = G(x_i'\beta_0 + \alpha_0 d_i)\{1 - G(x_i'\beta_0 + \alpha_0 d_i)\}$. In addition, the expression for Γ_2 leads to the following bound

$$\begin{aligned}&\left|\Gamma_2(\alpha,h_0)[\widehat{h}-h_0] - \Gamma_2(\alpha_0,h_0)[\widehat{h}-h_0]\right| \leqslant\\ &\leqslant \bar{\mathrm{E}}[|\alpha-\alpha_0|\ |d_i z_{0i}|\ |x_i'\{\widehat{\beta}-\beta_0\}|] + \bar{\mathrm{E}}[|(\alpha-\alpha_0)d_i|\ |\widehat{z}_i - z_{0i}|]\\ &\leqslant |\alpha-\alpha_0| \cdot \|x_i'\{\widehat{\beta}-\beta_0\}\|_{2,n}\{\bar{\mathrm{E}}[z_{0i}^2 d_i^2]\}^{1/2} + |\alpha-\alpha_0| \cdot \{\bar{\mathrm{E}}[(\widehat{z}_i - z_{0i})^2]\}^{1/2}\{\bar{\mathrm{E}}[d_i^2]\}^{1/2}\\ &\lesssim |\alpha-\alpha_0|\delta_n.\end{aligned} \tag{A.27}$$

To bound the second derivative, recall that for $G(t) = \exp(t)/\{1+\exp(t)\}$, we have $G'(t) = G(t)[1-G(t)]$, $G''(t) = G(t)[1-G(t)][1-2G(t)]$, are all less than 1 in absolute value. The second directional derivative Γ_{22} at $\tilde{h} = (\tilde{\beta},\tilde{z})$ with respect to the direction $\widehat{h}-h_0$ can be bounded by

$$\begin{aligned}&\left|\Gamma_{22}(\alpha,\tilde{h})[\widehat{h}-h_0,\widehat{h}-h_0]\right| = \Big|-\bar{\mathrm{E}}[G''(x_i'\tilde{\beta}+\alpha d_i)\tilde{z}_i\{x_i'(\widehat{\beta}_0-\beta_0)\}^2]\\ &-2\bar{\mathrm{E}}[G'(x_i'\tilde{\beta}+d_i\alpha)\{x_i'(\widehat{\beta}-\beta_0)\}\{\widehat{z}_i - z_i\}]\Big|\\ &\leqslant \{\max_{i\leqslant n}\mathrm{E}[|\tilde{z}_i|]\}\|x_i'(\widehat{\beta}-\beta_0)\|_{2,n}^2 + 2\|x_i'(\widehat{\beta}-\beta_0)\|_{2,n}\{\bar{\mathrm{E}}[(\widehat{z}_i - z_{0i})^2]\}^{1/2}\end{aligned} \tag{A.28}$$

In turn, since $\tilde{h} \in [h_0,\widehat{h}]$, $|\tilde{z}(d_i,x_i)| \leqslant |z_0(d_i,x_i)| + |\widehat{z}(d_i,x_i) - z_0(d_i,x_i)|$, we have that

$$\begin{aligned}\left|\Gamma(\alpha,\widehat{h}) - \Gamma(\alpha,h_0)\right. &\left. - \Gamma_2(\alpha,h_0)\left[\widehat{h}-h_0\right]\right| \leqslant \sup_{\tilde{h}\in[h_0,\widehat{h}]}\left|\Gamma_{2,2}(\alpha,\tilde{h})\left[\widehat{h}-h_0,\widehat{h}-h_0\right]\right|\\ &\leqslant \left(\max_{i\leqslant n}\mathrm{E}[|z_{0i}|] + \mathrm{E}[|\widehat{z}_i - z_{0i}|]\right)\|x_i'\{\widehat{\beta}-\beta_0\}\|_{2,n}^2 +\\ &+ \|x_i'\{\widehat{\beta}-\beta_0\}\|_{2,n}\{\bar{\mathrm{E}}[(\widehat{z}_i - z_{0i})^2]\}^{1/2}\\ &\lesssim_P \delta_n n^{-1/2}\end{aligned} \tag{A.29}$$

where $\max_{i\leqslant n}\mathrm{E}[|z_{0i}|]$ is uniformly bounded by ILOG(ii) and the last relation is assumed in Condition ILOG(iii).

Step 4. (Relations for Γ_1) By definition of Γ, its derivative with respect to α at $(\alpha,\tilde{h})$ is

$$\Gamma_1(\alpha,\tilde{h}) = -\bar{\mathrm{E}}[G'(x_i'\tilde{\beta}+\alpha d_i)\tilde{z}_i d_i].$$

Therefore, when the function above is evaluated at $\alpha = \alpha_0$ and $\tilde{h} = h_0 = (\beta_0, z_0)$, since for $G'(x_i'\beta_0 + \alpha_0 d_i) = w_i$, we have

$$\Gamma_1(\alpha_0, h_0) = -\bar{\mathrm{E}}[w_i d_i z_{0i}]. \tag{A.30}$$

Moreover, Γ_1 also satisfies

$$\begin{aligned} |\Gamma_1(\alpha, h_0) - \Gamma_1(\alpha_0, h_0)| &= \left|\bar{\mathrm{E}}[G'(x_i'\beta_0 + \alpha d_i) z_{0i} d_i] - \bar{\mathrm{E}}[G'(x_i'\beta_0 + \alpha_0 d_i) z_{0i} d_i]\right| \\ &\leqslant |\alpha - \alpha_0| \bar{\mathrm{E}}[|d_i^2 z_{0i}|]. \end{aligned} \tag{A.31}$$

Step 5. (Estimation of Variance) First note that

$$\begin{aligned} &|\mathbb{E}_n[\widehat{w}_i d_i \widehat{z}_i] - \bar{\mathrm{E}}[w_i d_i z_{0i}]| \\ &= |\mathbb{E}_n[\widehat{w}_i d_i \widehat{z}_i] - \mathbb{E}_n[w_i d_i z_{0i}]| + |\mathbb{E}_n[w_i d_i z_{0i}] - \bar{\mathrm{E}}[w_i d_i z_{0i}]| \\ &\leqslant |\mathbb{E}_n[(\widehat{w}_i - w_i) d_i \widehat{z}_i]| + |\mathbb{E}_n[w_i d_i (\widehat{z}_i - z_{0i})]| + |\mathbb{E}_n[w_i d_i z_{0i}] - \bar{\mathrm{E}}[w_i d_i z_{0i}]| \\ &\leqslant |\mathbb{E}_n[(\widehat{w}_i - w_i) d_i (\widehat{z}_i - z_{0i})]| + |\mathbb{E}_n[(\widehat{w}_i - w_i) d_i z_{0i}]| \\ &+ \|w_i d_i\|_{2,n} \|\widehat{z}_i - z_{0i}\|_{2,n} + |\ \mathbb{E}_n[w_i d_i z_{0i}] - \bar{\mathrm{E}}[w_i d_i z_{0i}]| \\ &\lesssim_P \|(\widehat{w}_i - w_i) d_i\|_{2,n} \|\widehat{z}_i - z_{0i}\|_{2,n} + \|\widehat{w}_i - w_i\|_{2,n} \|d_i z_{0i}\|_{2,n} \\ &+ \|w_i d_i\|_{2,n} \|\widehat{z}_i - z_{0i}\|_{2,n} + |\mathbb{E}_n[w_i d_i z_{0i}] - \bar{\mathrm{E}}[w_i d_i z_{0i}]| \\ &\lesssim_P \delta_n \end{aligned} \tag{A.32}$$

because $0 \leqslant w_i, \widehat{w}_i \leqslant 1$, $\bar{\mathrm{E}}[d_i^4] \leqslant C$, $\bar{\mathrm{E}}[z_{0i}^4] \leqslant C$ by Condition ILOG(ii) and Conditions ILOG(iii) and (iv).

Next we proceed to control the other term of the variance. Since $|\psi_{\check{\alpha},\widehat{h}}(y_i, d_i, x_i) - \psi_{\alpha_0,\widehat{h}}(y_i, d_i, x_i)| \leqslant |d_i(\check{\alpha} - \alpha_0)\widehat{z}_i|$ and $|\psi_{\alpha_0,\widehat{h}}(y_i, d_i, x_i) - \psi_{\alpha_0,h_0}(y_i, d_i, x_i)| \leqslant |\widehat{z}_i - z_{0i}| + |x_i'\{\widehat{\beta} - \beta_0\} z_{0i}|$ we have

$$\begin{aligned} &|\ \|\psi_{\check{\alpha},\widehat{h}}(y_i, d_i, x_i)\|_{2,n} - \|\psi_{\alpha_0,h_0}(y_i, d_i, x_i)\|_{2,n}| \\ &\leqslant \|d_i(\check{\alpha} - \alpha_0)\widehat{z}_i\|_{2,n} + \|\widehat{z}_i - z_{0i}\|_{2,n} + \|x_i'\{\widehat{\beta} - \beta_0\} z_{0i}\|_{2,n} \\ &\leqslant |\check{\alpha} - \alpha_0| \|d_i z_{0i}\|_{2,n} + |\check{\alpha} - \alpha_0| \|d_i(\widehat{z}_i - z_{0i})\|_{2,n} \\ &+ \|\widehat{z}_i - z_{0i}\|_{2,n} + \|(x_i'\{\widehat{\beta} - \beta_0\})^2\|_{2,n}^{1/2} \|z_{0i}^2\|_{2,n}^{1/2} \\ &\lesssim_P \delta_n \end{aligned} \tag{A.33}$$

by ILOG(ii) and ILOG(iv). Also, by Condition ILOG(ii), $\bar{\mathrm{E}}[w_i z_{0i}^2] \leqslant \bar{\mathrm{E}}[z_{0i}^2] \leqslant C$ we have $|\mathbb{E}_n[\psi^2_{\alpha_0,h_0}(y_i, d_i, x_i)] - \bar{\mathrm{E}}[\psi^2_{\alpha_0,h_0}(y_i, d_i, x_i)]| \lesssim_P \delta_n$.

Step 6. (Main Step for χ^2) Note that the denominator of $L_n(\alpha_0)$ was analyzed in relation (A.33) of Step 5. Next consider the numerator of $L_n(\alpha_0)$. Since $\Gamma(\alpha_0, h_0) = \bar{\mathrm{E}}[\psi_{\alpha_0,h_0}(y_i, d_i, x_i)] = 0$, we have

$$\mathbb{E}_n[\psi_{\alpha_0,\widehat{h}}(y_i, d_i, x_i)] = (\mathbb{E}_n - \bar{\mathrm{E}})[\psi_{\alpha_0,\widehat{h}}(y_i, d_i, x_i) - \psi_{\alpha_0,h_0}(y_i, d_i, x_i)] + \Gamma(\alpha_0, \widehat{h}) + \mathbb{E}_n[\psi_{\alpha_0,h_0}(y_i, d_i, x_i)].$$

By Condition ILOG(iii) and (A.25) with $\alpha = \alpha_0$, it follows that

$$|(\mathbb{E}_n - \bar{\mathrm{E}})[\psi_{\alpha_0,\widehat{h}}(y_i, d_i, x_i) - \psi_{\alpha_0,h_0}(y_i, d_i, x_i)]| \leqslant \delta_n n^{-1/2} \quad \text{and} \quad |\Gamma(\alpha_0, \widehat{h})| \lesssim_P \delta_n n^{-1/2}.$$

Therefore, using that $nA_n^2 = nB_n^2 + n(A_n - B_n)^2 + 2nB_n(A_n - B_n)$, for $A_n = \mathbb{E}_n[\psi_{\alpha_0,\widehat{h}}(y_i,d_i,x_i)]$ and $B_n = \mathbb{E}_n[\psi_{\alpha_0,h_0}(y_i,d_i,x_i)] \lesssim_P \{\bar{\mathrm{E}}[w_i z_{0i}^2]\}^{1/2} n^{-1/2}$ we have

$$\begin{aligned} nL_n(\alpha_0) &= \frac{n|\mathbb{E}_n[\psi_{\alpha_0,\widehat{h}}(y_i,d_i,x_i)]|^2}{\mathbb{E}_n[\psi^2_{\alpha_0,\widehat{h}}(y_i,d_i,x_i)]} = \frac{n|\mathbb{E}_n[\psi_{\alpha_0,h_0}(y_i,d_i,x_i)]|^2 + O_P(\delta_n)}{\bar{\mathrm{E}}[w_i z_{0i}^2] + O_P(\delta_n)} \\ &= \frac{n|\mathbb{E}_n[\psi_{\alpha_0,h_0}(y_i,d_i,x_i)]|^2}{\bar{\mathrm{E}}[w_i z_{0i}^2]} + O_P(\delta_n) \end{aligned}$$

since $\bar{\mathrm{E}}[w_i z_{0i}^2]$ is bounded away from zero because $c \leqslant |\bar{\mathrm{E}}[w_i d_i z_{0i}]| \leqslant \{\bar{\mathrm{E}}[w_i d_i^2]\bar{\mathrm{E}}[w_i z_{0i}^2]\}^{1/2}$ and $\bar{\mathrm{E}}[w_i d_i^2]$ is bounded above uniformly. The result then follows since $\sqrt{n}\mathbb{E}_n[\psi_{\alpha_0,h_0}(y_i,d_i,x_i)] \rightsquigarrow N(0, \bar{\mathrm{E}}[w_i z_{0i}^2])$ and $\mathrm{E}[\psi^2_{\alpha_0,h_0}(y_i,d_i,x_i) \mid x_i, d_i] = w_i z_{0i}^2$.

■

Appendix B. Proofs of Theorems

Proof of Theorem 1. We will verify Condition ILOG and the result follows by Lemma 1. We will use the (optimal) instrument $z_{0i} = v_i/\sqrt{w_i}$. The assumptions on the conditional variance w_i and the moment conditions on d_i and v_i in Condition L imply Condition ILOG(ii).

Under the assumption on the weights stated in Condition L(ii) and the sparse eigenvalue bounds stated in Condition L(iii), we have that $\kappa_{\mathbf{c}}$ defined in (C.41) is bounded away from zero with probability $1-2\Delta_n$ for n sufficiently large, see [9].

Step 1 relies on Post-Lasso-Logistic. To apply Lemma 2 to obtain rates and sparsity bounds, we first verify the side condition $q_{\Delta_{\mathbf{c}}} > 3(1+\frac{1}{c})\lambda\sqrt{s}/(n\kappa_{\mathbf{c}})$. Without loss of generality assume that T contains the treatment d in its support. Thus for $\tilde{x}_i = (d_i, x_i')'$, $\delta = (\delta_d, \delta_x')'$, we have

$$\begin{aligned} \inf_{\delta\in\Delta_{\mathbf{c}}} \tfrac{\|\tilde{x}_i'\delta\|_{2,n}^3}{\mathbb{E}_n[|\tilde{x}_i'\delta|^3]} &\geqslant \inf_{\delta\in\Delta_{\mathbf{c}}} \tfrac{\|\tilde{x}_i'\delta\|_{2,n}^2\|\delta_T\|\kappa_{\mathbf{c}}}{4\mathbb{E}_n[|x_i'\delta_x|^3]+4\mathbb{E}_n[|d_i\delta_d|^3]} \geqslant \inf_{\delta\in\Delta_{\mathbf{c}}} \tfrac{\|\tilde{x}_i'\delta\|_{2,n}^2\|\delta_T\|\kappa_{\mathbf{c}}}{4K_x\|\delta_x\|_1\mathbb{E}_n[|x_i'\delta_x|^2]+4|\delta_d|^3\mathbb{E}_n[|d_i|^3]} \\ &\geqslant \inf_{\delta\in\Delta_{\mathbf{c}}} \tfrac{\|\tilde{x}_i'\delta\|_{2,n}^2\|\delta_T\|\kappa_{\mathbf{c}}}{4K_x\|\delta_x\|_1\{\|\tilde{x}_i'\delta\|_{2,n}+\|\delta_d d_i\|_{2,n}\}^2+4|\delta_d|^2\mathbb{E}_n[|d_i|^3]\|\delta_T\|_1} \\ &\geqslant \inf_{\delta\in\Delta_{\mathbf{c}}} \tfrac{\|\tilde{x}_i'\delta\|_{2,n}^2\|\delta_T\|_1\kappa_{\mathbf{c}}/\sqrt{s}}{8K_x(1+\mathbf{c})\|\delta_T\|_1\|\tilde{x}_i'\delta\|_{2,n}^2+8K_x(1+\mathbf{c})\|\delta_T\|_1|\delta_d|^2\{\|d_i\|_{2,n}^2+\mathbb{E}_n[|d_i|^3]\}} \\ &\geqslant \tfrac{\kappa_{\mathbf{c}}/\sqrt{s}}{8K_x(1+\mathbf{c})\{1+\|d_i\|_{2,n}^2/\kappa_{\mathbf{c}}^2+\mathbb{E}_n[|d_i|^3]/\kappa_{\mathbf{c}}^2\}} \gtrsim_P \tfrac{1}{\sqrt{s}K_x} \end{aligned}$$

by $\bar{\mathrm{E}}[d_i^4] \leqslant C$ and $\kappa_{\mathbf{c}}$ bounded away from zero by Condition L. Also by Condition L we have $\min_{i\leqslant n} w_i > c > 0$ with probability $1-\Delta_n$. Therefore, since $\lambda \lesssim \sqrt{n\log(p\vee n)}$, we have

$$\frac{n\kappa_{\mathbf{c}}}{\lambda\sqrt{s}+\sqrt{sn\log(p\vee n)}} \inf_{\delta\in\Delta_{\mathbf{c}}} \frac{\|\sqrt{w_i}\tilde{x}_i'\delta\|_{2,n}^3}{\mathbb{E}_n[w_i|\tilde{x}_i'\delta|^3]} \gtrsim_P \frac{\sqrt{n}}{K_x s\log(p\vee n)} \to_P \infty$$

under $K_x^2 s^2 \log^2(p\vee n) \leqslant \delta_n n$.

To apply Lemma 3 we need to verify the side condition

$$q_{A_{\widehat{s}+s}}/6 > \sqrt{\widehat{s}+s}\|\nabla\Lambda(\eta_0)\|_\infty/\sqrt{\phi_{\min}(\widehat{s}+s)},$$

where $\widehat{s} \lesssim_P s$ by Lemma 2. Similarly to the previous argument, for $\tilde{x}_i = (d_i, x_i')'$ and $\delta = (\delta_d, \delta_x')'$,

$$\begin{array}{rl} \inf_{\|\delta\|_0 \leqslant s+Cs} \frac{\|\tilde{x}_i'\delta\|_{2,n}^3}{\mathbb{E}_n[|\tilde{x}_i'\delta|^3]} & \geqslant \inf_{\|\delta\|_0 \leqslant s+Cs} \frac{\{\phi_{\min}(s+Cs)\}^{3/2}\|\delta\|^3}{4\mathbb{E}_n[|x_i'\delta_x|^3]+4|\delta_d|^3\mathbb{E}_n[|d_i|^3]} \\ & \geqslant \inf_{\|\delta\|_0 \leqslant s+Cs} \frac{\{\phi_{\min}(s+Cs)\}^{3/2}\|\delta\|^3}{4K_x\|\delta_x\|_1\phi_{\max}(s+Cs)\|\delta_x\|^2+4\|\delta\|^3\mathbb{E}_n[|d_i|^3]} \\ & \geqslant \frac{\{\phi_{\min}(s+Cs)\}^{3/2}}{4K_x\sqrt{s+Cs}\phi_{\max}(s+Cs)+4\mathbb{E}_n[|d_i|^3]} \gtrsim_P \frac{1}{K_x\sqrt{s}} \end{array}$$

by Condition L(iii) and $\mathbb{E}_n[|d_i|^3] \lesssim \bar{\mathrm{E}}[|d_i|^3] \leqslant C$ with probability $1-o(1)$. Therefore, by $K_x^2 s^2 \log^2(p\vee n) \leqslant \delta_n n$ and $\lambda \lesssim \sqrt{n\log(p\vee n)}$, we have

$$\frac{n\sqrt{\phi_{\min}(s+Cs)}}{\lambda\sqrt{s}+\sqrt{sn\log(p\vee n)}} \inf_{\|\delta\|_0 \leqslant s+Cs} \frac{\|\sqrt{w_i}\tilde{x}_i'\delta\|_{2,n}^3}{\mathbb{E}_n[w_i|\tilde{x}_i'\delta|^3]} \gtrsim_P \frac{\sqrt{n}}{K_x s\log(p\vee n)} \to \infty,$$

Therefore $|\widetilde{\alpha} - \alpha_0| \lesssim_P \sqrt{s\log(p\vee n)/n}$ so that $\mathcal{A} = \{\alpha : |\alpha - \widetilde{\alpha}| \leqslant C\log^{-1} n\} \supseteq \{\alpha : |\alpha - \alpha_0| \leqslant n^{-1/2}/\delta_n\}$ under $s\log(p\vee n)\log^2 n \leqslant \delta_n n$ which is required in ILOG(i). This also ensures the initial rate required for $\check{\alpha}$ in ILOG(iii) since $\check{\alpha} \in \mathcal{A}$.

Step 2 relies on Post-Lasso with estimated weights. Condition WL(i) and (ii) is assumed by Conditions L and a suitable choice of the confidence level γ to satisfy the growth condition. Condition WL(iii) follows from Lemma 4 applied twice with $\zeta_i = v_i$ and $\zeta_i = d_i$ under the condition that $K_x^4 \log p \leqslant \delta_n n$.

The first part of Condition WL(iv), since G is 1-Lipschitz and $0 \leqslant \widehat{w}_i \leqslant 1$, follows from

$$\begin{array}{rl} \max_{j\leqslant p}\mathbb{E}_n[\{\sqrt{\widehat{w}_i}-\sqrt{w_i}\}^2 x_{ij}^2 v_i^2] & \leqslant \max_{i\leqslant n}\frac{|\sqrt{\widehat{w}_i}-\sqrt{w_i}|^2}{w_i}\max_{j\leqslant p}\mathbb{E}_n[w_i x_{ij}^2 v_i^2] \\ & \leqslant \frac{K_x^2\|\widetilde{\beta}-\beta_0\|_1^2+|\widetilde{\alpha}-\alpha_0|^2\max_{i\leqslant n}|d_i|}{\min_{i\leqslant n} w_i}\max_{j\leqslant p}\mathbb{E}_n[w_i x_{ij}^2 v_i^2] \\ & \lesssim_P \delta_n \end{array}$$

since $\max_{j\leqslant p}\mathbb{E}_n[w_i x_{ij}^2 v_i^2] \leqslant \max_{j\leqslant p}(\mathbb{E}_n - \bar{\mathrm{E}})[w_i x_{ij}^2 v_i^2] + \max_{j\leqslant p}\bar{\mathrm{E}}[w_i x_{ij}^2 v_i^2] \lesssim_P 1$ (by Lemma 4 with $\zeta_i = \sqrt{w_i}v_i$ under the condition that $K_x^4\log p \leqslant \delta_n n$), $\min_{i\leqslant n} w_i \geqslant c > 0$ with probability $1-\Delta_n$, $\max_{i\leqslant n} d_i^2 \lesssim_P n^{1/2}$ and the rate of $(\widetilde{\beta}, \widetilde{\alpha})$.

In this case $\widehat{f}_i = \widehat{w}_i/\widehat{\sigma}_i = \sqrt{\widehat{w}_i}$ so that $\widehat{c}_f^2 = \mathbb{E}_n[(\widehat{w}_i - w_i)^2 v_i^2/\sqrt{w_i}]$. To bound $\widehat{c}_f$ note that $\widehat{w}_i + w_i \leqslant 1$ and $|\widehat{w}_i - w_i| \leqslant |x_i'(\widetilde{\beta}-\beta_0)| + |d_i(\widetilde{\alpha}-\alpha_0)|$. Then since with probability $1-\Delta_n$ $\min_{i\leqslant n} w_i \geqslant c > 0$, with the same probability

$$\begin{array}{rl} \widehat{c}_f^2 & = \mathbb{E}_n[(\widehat{w}_i - w_i)^2 v_i^2/\sqrt{w_i}] \leqslant \frac{2}{\sqrt{c}}\mathbb{E}_n[\{v_i x_i'(\widetilde{\beta}-\beta_0)\}^2] + \frac{2}{\sqrt{c}}|\widetilde{\alpha}-\alpha_0|^2\mathbb{E}_n[d_i^2 v_i^2] \\ & \leqslant \frac{2}{\sqrt{c}}(\mathbb{E}_n - \bar{\mathrm{E}})[\{v_i x_i'(\widetilde{\beta}-\beta_0)\}^2] + \frac{2}{\sqrt{c}}\bar{\mathrm{E}}[\{v_i x_i'(\widetilde{\beta}-\beta_0)\}^2] + \frac{2}{\sqrt{c}}|\widetilde{\alpha}-\alpha_0|^2\{\mathbb{E}_n[d_i^4]\}^{1/2}\{\mathbb{E}_n[v_i^4]\}^{1/2} \end{array} \quad \text{(B.34)}$$

Recall that $\|\widetilde{\beta}\|_0 \lesssim_P s$, $\|\widetilde{\beta}-\beta_0\| \lesssim_P \sqrt{s\log p/n}$, $|\widetilde{\alpha}-\alpha_0| \lesssim_P \sqrt{s\log p/n}$. Conditional on $\{x_i, i = 1,\ldots,n\}$, we will apply Lemma 6 with $X_i = v_i x_i$. In that case, we have $K = \{\mathrm{E}[\max_{i\leqslant n}\|X_i\|_\infty^2]\}^{1/2} \leqslant K_x\{\bar{\mathrm{E}}[\max_{i\leqslant n} v_i^2]\}^{1/2} \lesssim n^{1/8}K_x$ (since $\bar{\mathrm{E}}[v_i^8] \leqslant C$), and $\bar{\mathrm{E}}[(\delta' X_i)^2] \leqslant \mathbb{E}_n[(x_i'\delta)^2]\max_{i\leqslant n}\mathrm{E}[v_i^2 \mid x_i] \leqslant C\phi_{\max}(\|\delta\|_0)\|\delta\|^2$.

$$\begin{array}{rl} (\mathbb{E}_n - \bar{\mathrm{E}})[\{v_i x_i'(\widetilde{\beta}-\beta_0)\}^2] & \leqslant \|\widetilde{\beta}-\beta_0\|^2 \sup_{\|\delta\|_0\leqslant 2Cs, \|\delta\|=1}\left|\mathbb{E}_n\left[(\delta' X_i)^2 - \mathrm{E}[(\delta' X_i)^2]\right]\right| \\ & \lesssim_P \|\widetilde{\beta}-\beta_0\|^2\left\{\frac{K_x^2 n^{1/4} s\log^3 n\log(p\vee n)}{n} + \sqrt{\frac{K_x^2 n^{1/4} s\log^3 n\log(p\vee n)}{n}}\phi_{\max}(2Cs)\right\} \\ & \lesssim \frac{s\log p}{n}\left\{\frac{K_x\log^3 n}{n^{1/4}}\frac{K_x s\log(p\vee n)}{n^{1/2}} + \sqrt{\frac{K_x\log^3 n}{n^{1/4}}\frac{K_x s\log(p\vee n)}{n^{1/2}}}\phi_{\max}(2Cs)\right\} \end{array}$$

under the conditions $K_x^4 \leqslant \delta_n n^{4/q}$, $q > 4$, and $K_x^2 s^2 \log^2(p \vee n) \leqslant \delta_n n$, and $\phi_{\max}(s/\delta_n)$ being bounded from above with probability $1 - \Delta_n$ by Condition L(iii).

Also conditional on $\{x_i, i = 1, \ldots, n\}$, since $\max_{i \leqslant n} \mathrm{E}[v_i^2 \mid x_i] \leqslant C$, we have

$$\bar{\mathrm{E}}[\{v_i x_i'(\widetilde{\beta} - \beta_0)\}^2] \leqslant \mathbb{E}_n[\{x_i'(\widetilde{\beta} - \beta_0)\}^2] \max_{i \leqslant n} \mathrm{E}[v_i^2 \mid x_i] \leqslant C \mathbb{E}_n[\{x_i'(\widetilde{\beta} - \beta_0)\}^2] \lesssim_P s \log p/n.$$

Since $\mathbb{E}_n[d_i^4] \lesssim_P \bar{\mathrm{E}}[d_i^4] \leqslant C$ and $\mathbb{E}_n[v_i^4] \lesssim_P \bar{\mathrm{E}}[v_i^4] \leqslant C$, the last term in (B.34) satisfies

$$|\widetilde{\alpha} - \alpha_0|^2 \{\mathbb{E}_n[d_i^4]\}^{1/2} \{\mathbb{E}_n[v_i^4]\}^{1/2} \lesssim_P s \log p/n.$$

Therefore, by Theorem 3, we have $\|x_i'(\widetilde{\theta} - \theta_0)\|_{2,n} \lesssim_P \sqrt{s \log(p \vee n)}/\sqrt{n}$ and $\|\widetilde{\theta}\|_0 \lesssim_P Cs$.

The choice of instrument is $z_{0i} = v_i/\sqrt{w_i} = d_i - x_i'\theta_0$ and $\widehat{z}_i = d_i - x_i'\widetilde{\theta}$ so that

$$\widehat{z}_i - z_{0i} = x_i'\{\theta_0 - \widetilde{\theta}\} \tag{B.35}$$

The rates established above for $(\widetilde{\alpha}, \widetilde{\beta}, \widetilde{\theta})$ imply (A.20) in ILOG(ii) since by Condition L(ii)

$$\begin{aligned}
\{1 + \max_{i \leqslant n} |x_i'(\widetilde{\theta} - \theta_0)|^{1/2}\} \|x_i'(\widetilde{\beta} - \beta_0)\|_{2,n} &\lesssim_P \left(1 + \sqrt{K_x s \sqrt{\log(p \vee n)/n}}\right) \sqrt{\tfrac{s \log p}{n}} \\
&\lesssim n^{-1/4} \left(\sqrt{\tfrac{s \log p}{n^{1/2}}} + \sqrt{\tfrac{K_x s \log p}{n^{1/2}} \tfrac{s \log^{1/2}(p \vee n)}{n^{1/2}}} \right) \lesssim \delta_n n^{-1/4} \\
\|x_i'(\widetilde{\beta} - \beta_0)\|_{2,n} \|x_i'(\widetilde{\theta} - \theta_0)\|_{2,n} &\lesssim \sqrt{\tfrac{s \log p}{n}} \sqrt{\tfrac{s \log(p \vee n)}{n}} \leqslant n^{-1/2} \tfrac{s \log(p \vee n)}{n^{1/2}} \lesssim \delta_n n^{-1/2} \\
|\widetilde{\alpha} - \alpha_0| \; \|x_i'(\widetilde{\theta} - \theta_0)\|_{2,n} &\lesssim \delta_n n^{-1/2}
\end{aligned}$$

Moreover, Condition ILOG(iv) holds since

$$\begin{aligned}
\|\{x_i'(\widetilde{\theta} - \theta_0)\}^2\|_{2,n} &\leqslant K_x \|\widetilde{\theta} - \theta_0\|_1 \|x_i'(\widetilde{\theta} - \theta_0)\|_{2,n} \lesssim_P \sqrt{\tfrac{K_x^2 s^2 \log(p \vee n)}{n}} \sqrt{\tfrac{s \log(p \vee n)}{n}} = o_P(1) \\
\|\{x_i'(\widetilde{\beta} - \beta_0)\}^2\|_{2,n} &= o_P(1)
\end{aligned}$$

and $\|(1 \vee |d_i|)(\widehat{z}_i - z_{0i})\|_{2,n} \leqslant (1 + \|d_i^2\|_{2,n}^{1/2}) \|\{x_i'(\widetilde{\theta} - \theta_0)\}^2\|_{2,n}^{1/2} = o_P(1)$.

Next we verify Condition ILOG(iii). Let $\widehat{\varphi}_i(\alpha) = y_i - G(x_i'\widehat{\beta} + d_i\alpha)$, $\varphi_i(\alpha) = y_i - G(x_i'\beta_0 + d_i\alpha)$. Note that

$$\sup_{\alpha \in \mathcal{A}} \left| (\mathbb{E}_n - \bar{\mathrm{E}}) \left[\widehat{\varphi}_i(\alpha) \widehat{z}_i - \varphi_i(\alpha) z_{0i} \right] \right| \leqslant \sup_{\alpha \in \mathcal{A}} \left| (\mathbb{E}_n - \bar{\mathrm{E}}) \left[\{\widehat{\varphi}_i(\alpha) - \varphi_i(\alpha)\} (\widehat{z}_i - z_{0i}) \right] \right| + \tag{B.36}$$

$$+ \sup_{\alpha \in \mathcal{A}} \left| (\mathbb{E}_n - \bar{\mathrm{E}}) \left[\varphi_i(\alpha) (\widehat{z}_i - z_{0i}) \right] \right| + \tag{B.37}$$

$$+ \sup_{\alpha \in \mathcal{A}} \left| (\mathbb{E}_n - \bar{\mathrm{E}}) \left[\{\widehat{\varphi}_i(\alpha) - \varphi_i(\alpha)\} z_{0i} \right] \right|. \tag{B.38}$$

To bound (B.36), since $|\widehat{\varphi}_i(\alpha) - \varphi_i(\alpha)| \leqslant |x_i'(\widetilde{\beta} - \beta_0)|$, $|\widehat{w}_i| \leqslant 1$, we use Cauchy-Schwartz to obtain

$$\text{(B.36)} \quad \leqslant \|x_i'(\widetilde{\beta} - \beta_0)\|_{2,n} \{2 \|x_i'(\widetilde{\theta} - \theta_0)\|_{2,n}\} \lesssim_P \delta_n n^{-1/2}.$$

To bound (B.37) we consider

$$\begin{aligned}
\text{(B.37)} \quad &\leqslant \sup_{\alpha \in \mathcal{A}} \left| (\mathbb{E}_n - \bar{\mathrm{E}}) \left[\{\varphi_i(\alpha) - \varphi_i(\alpha_0)\} x_i'(\theta_0 - \widetilde{\theta}) \right] \right| + \left| (\mathbb{E}_n - \bar{\mathrm{E}}) \left[\varphi_i(\alpha_0) x_i'(\theta_0 - \widetilde{\theta}) \right] \right| \\
&\lesssim_P \sup_{\alpha \in \mathcal{A}, \|\delta\|_1 \leqslant Cs \sqrt{\frac{\log(p \vee n)}{n}}} \left| (\mathbb{E}_n - \bar{\mathrm{E}}) \left[\{\varphi_i(\alpha) - \varphi_i(\alpha_0)\} x_i'\delta \right] \right| + \sup_{\|\delta\|_1 \leqslant Cs \sqrt{\frac{\log(p \vee n)}{n}}} \left| (\mathbb{E}_n - \bar{\mathrm{E}}) \left[\varphi_i(\alpha_0) x_i'\delta \right] \right|.
\end{aligned}$$

Using Lemma 5 twice (one application of the lemma has $W_{ij} = d_i x_{ij}$, $\mathcal{T} = \{(\alpha - \alpha_0)\delta \in \mathbb{R}^p : \alpha \in \mathcal{A}, \|\delta\|_1 \leqslant Cs\sqrt{\log(p \vee n)/n}\}$ and $\xi_i = 1$, the other application is standard) we have

$$\text{(B.37)} \quad \lesssim_P \delta_n Cs \sqrt{\tfrac{\log(p \vee n)}{n}} \sqrt{\tfrac{\log p}{n}} + Cs \sqrt{\tfrac{\log(p \vee n)}{n}} \sqrt{\tfrac{\log p}{n}} \lesssim_P \delta_n n^{-1/2}$$

since by Lemma 4, $\max_{i\leqslant n} \mathrm{E}[d_i^2 \mid x_i] \leqslant C$ and $\max_{j\leqslant p} \mathbb{E}_n[x_{ij}^2] \leqslant 1$, we have

$$\max_{j\leqslant p} \mathbb{E}_n[d_i^2 x_{ij}^2] \leqslant \max_{j\leqslant p} (\mathbb{E}_n - \bar{\mathrm{E}})[d_i^2 x_{ij}^2] + C \max_{j\leqslant p} \mathbb{E}_n[x_{ij}^2] \lesssim \sqrt{\frac{\log p}{n}} K_x^2 \sqrt{\mathbb{E}_n[d_i^4]} + C \lesssim_P C$$

under $K_x^4 \log p \leqslant \delta_n n$ and $\bar{\mathrm{E}}[d_i^4] \leqslant C$.

Next we proceed to bound (B.38). We will consider the class of functions which pertains to $\{\widehat{\varphi}_i(\alpha) - \varphi_i(\alpha)\} z_{0i}$, namely for some C suitably large

$$\mathcal{F} = \{G(x_i'\beta + d_i\alpha) z_{0i} - G(x_i'\beta_0 + d_i\alpha) z_{0i} : \|\beta\|_0 \leqslant Cs,\ \|x_i'\{\beta - \beta_0\}\|_{2,n} \leqslant C\sqrt{s \log p / n}\}$$

Let $h_i(t, \alpha) = G(x_i'(\beta + t) + d_i\alpha) - G(x_i'\beta_0 + d_i\alpha)$ so that $|h_i(t,\alpha)| \leqslant |t'x_i|$ and $\xi_i = z_{0i}$. Therefore $\|\mathcal{T}\|_1 \lesssim s\sqrt{\log(p \vee n)/n}$ and note that $\max_{j\leqslant p} \mathbb{E}_n[x_{ij}^2 z_{0i}^2] \lesssim_P C$ under $K_x^4 \log p \leqslant \delta_n n$ and the moment conditions. By Lemma 5 we have

$$(B.38) \lesssim_P \sqrt{\frac{\log(p \vee n)}{n}} \|\mathcal{T}\|_1 \lesssim \frac{s \log(p \vee n)}{n} \lesssim o(1) n^{-1/2}$$

provided $s^2 \log^2(p \vee n) \leqslant \delta_n n$.

The last condition to be verified is the second condition in ILOG(iii). We will show that $\mathbb{E}_n[\widehat{\varphi}_i(\alpha)\widehat{z}_i]$ changes sign over $\alpha \in \mathcal{A}$ with high probability which by continuity of $\widehat{\varphi}_i(\cdot)$ implies that $\mathbb{E}_n[\widehat{\varphi}_i(\check{\alpha})\widehat{z}_i] =_P 0$. Note that for any $\alpha \in \mathcal{A}$

$$\begin{aligned} \mathbb{E}_n[\widehat{\varphi}_i(\alpha)\widehat{z}_i] &= \overbrace{(\mathbb{E}_n - \bar{\mathrm{E}})[\widehat{\varphi}_i(\alpha)\widehat{z}_i - \varphi_i(\alpha) z_{0i}]}^{(1)} + \overbrace{\bar{\mathrm{E}}[\widehat{\varphi}_i(\alpha)\widehat{z}_i] - \bar{\mathrm{E}}[\varphi_i(\alpha) z_{0i}]}^{(2)} + \\ &+ \underbrace{(\mathbb{E}_n - \bar{\mathrm{E}})[\varphi_i(\alpha) z_{0i}]}_{(3)} + \bar{\mathrm{E}}[\varphi_i(\alpha) z_{0i}]. \end{aligned}$$

Note that by the first part of ILOG(iii) established before, we have $(1) \lesssim_P \delta_n\ n^{-1/2}$. By the expansion (A.25), we have $(2) \leqslant \delta_n n^{-1/2} + \delta_n |\alpha - \alpha_0|$ from (A.26), (A.27) and (A.28). Moreover, we have by Lemma 5 and $\bar{\mathrm{E}}[d_i^2 z_{0i}^2] = O(1)$

$$(3) \leqslant \sup_{\alpha \in \mathcal{A}} |(\mathbb{E}_n - \bar{\mathrm{E}})[\{\varphi_i(\alpha) - \varphi_i(\alpha_0)\} z_{0i}]| + |(\mathbb{E}_n - \bar{\mathrm{E}})[\varphi_i(\alpha_0) z_{0i}]| \lesssim_P (\delta_n + 1) n^{-1/2}.$$

Therefore, since $\bar{\mathrm{E}}[\varphi_i(\alpha) z_{0i}] = (\alpha - \alpha_0)\bar{\mathrm{E}}[v_i^2] + O(|\alpha - \alpha_0|^2)$, we have

$$\begin{aligned} \mathbb{E}_n[\varphi_i(\alpha) z_{0i}] &= O_P(n^{-1/2} + \delta_n |\alpha - \alpha_0|) + \bar{\mathrm{E}}[\varphi_i(\alpha) z_{0i}] \\ &= O_P(n^{-1/2}) + (\alpha - \alpha_0)\{\bar{\mathrm{E}}[v_i^2] + O_P(\delta_n)\} + O(|\alpha - \alpha_0|^2). \end{aligned} \tag{B.39}$$

Since $\bar{\mathrm{E}}[v_i^2] \geqslant c$ and $\delta_n \to 0$, when we evaluate (B.39) on the extreme points $\alpha^k, k = 1, 2$, of $\mathcal{A}$, we obtain a positive value for one extreme and a negative value for the other extreme for n large enough since $|\alpha^k - \alpha_0| = C \log^{-1} n$. ∎

Proof of Theorem 2. Let $\widehat{T}^* = \mathrm{support}(\widehat{\theta}) \cup \mathrm{support}(\widehat{\beta})$. By the first order condition we have

$$\mathbb{E}_n[\{y_i - G(d_i\check{\alpha} + x_i'\check{\beta})\}(d_i,\ x_{i\widehat{T}^*}')'] = 0. \tag{B.40}$$

Next we will construct a suitable instrument to apply Lemma 1. Define

$$\widehat{\theta}^* \in \arg\min_{\theta} \|x_i'(\theta - \theta_0)\|_{2,n} \quad : \ \mathrm{support}(\theta) \subseteq \widehat{T}^*.$$

We use the optimal instrument $z_{0i} = v_i/\sqrt{w_i} = d_i - x_i'\theta_0$ and the estimated instrument $\widehat{z}_i = d_i - x_i'\widehat{\theta}^*$. Note that by (B.40), taking the linear combination $(1; -\widehat{\theta}^*)$ of the optimality condition we have

$$\mathbb{E}_n[\{y_i - G(d_i\check{\alpha} + x_i'\check{\beta})\}\widehat{z}_i] = 0.$$

Therefore $\check{\alpha}$ minimizes the criterion

$$L_n(\alpha) = \frac{|\mathbb{E}_n[\{y_i - G(d_i\alpha + x_i'\check{\beta})\}\widehat{z}_i]|^2}{\mathbb{E}_n[\{y_i - G(d_i\alpha + x_i'\check{\beta})\}^2\widehat{z}_i^2]},$$

induced by $\{(x_i'\check{\beta}, \widehat{z}_i) : i = 1, \ldots, n\}$, over $\alpha \in \mathbb{R}$.

Regarding Steps 1 and 2, rates of convergence for Lasso-Logistics, Post-Lasso-Logistics, Lasso with estimated weights and the associated sparsity bounds are established as in the proof of Theorem 1. Thus we have $\|\widehat{\theta}\|_0 \lesssim_P s$, $\|\widehat{\beta}\|_0 \lesssim_P s$, $\Lambda(\widehat{\alpha}, \widehat{\beta}) - \Lambda(\alpha_0, \beta_0) \lesssim_P s\log p/n$ and $\|x_i'(\widehat{\theta} - \theta_0)\|_{2,n} \lesssim_P \sqrt{s\log(p \vee n)/n}$.

Next we analyze Step 3. The sparsity results above implies that $\widehat{T}^* = \text{support}(\widehat{\theta}) \cup \text{support}(\widehat{\beta})$ satisfies $|\widehat{T}^*| \lesssim_P s$. Moreover, since $\text{support}(\widehat{\beta}) \subset \widehat{T}^*$ we have

$$\Lambda(\check{\alpha}, \check{\beta}) - \Lambda(\alpha_0, \beta_0) \leqslant \Lambda(\widehat{\alpha}, \widehat{\beta}) - \Lambda(\alpha_0, \beta_0) \lesssim_P s\log p/n.$$

Thus, by the condition on the sparse eigenvalue requirement in Condition L, Lemma 3 establishes a rate of convergence for post-model selection Logistic regression estimator $\|x_i'(\check{\beta} - \beta_0)\|_{2,n} \lesssim_P \sqrt{s\log p/n}$, $|\check{\alpha} - \alpha_0| \lesssim_P \sqrt{s\log p/n}$, and $\|\check{\beta} - \beta_0\|_1 \lesssim_P \sqrt{s}\|\check{\beta} - \beta_0\| \lesssim_P \sqrt{s}\|x_i'(\check{\beta} - \beta_0)\|_{2,n}/\{\phi_{\min}(C's)\}^{1/2} \lesssim_P s\sqrt{\log p/n}$. Moreover, since $\text{support}(\widehat{\theta}) \subset \widehat{T}^*$ we have $\|x_i'(\widehat{\theta}^* - \theta_0)\|_{2,n} \leqslant \|x_i'(\widehat{\theta} - \theta_0)\|_{2,n} \lesssim_P \sqrt{s\log(p \vee n)/n}$ and $\|\widehat{\theta}^* - \theta_0\|_1 \lesssim_P \sqrt{s}\|\widehat{\theta}^* - \theta_0\| \lesssim_P \sqrt{s}\|x_i'(\widehat{\theta}^* - \theta_0)\|_{2,n}/\{\phi_{\min}(C's)\}^{1/2} \lesssim_P s\sqrt{\log(p \vee n)/n}$.

The remaining assumptions in Condition ILOG can be verified as in the proof of Theorem 1.

Next we show the validity of the calculation of $\widehat{\Sigma}_{2n}^2 = \{\mathbb{E}_n[\check{w}_i(d_i, x_{i\widehat{T}^*}')'(d_i, x_{i\widehat{T}^*}')]\}_{11}^{-1}$. Since $\min_{i\leqslant n} w_i > c$ with probability $1 - \Delta_n$ and sparse eigenvalues of size s/δ_n are bounded away from zero and from above with probability $1 - \Delta_n$ by Condition L, and $\max_{i\leqslant n}|\check{w}_i - w_i| = o_P(1)$ by the rates above, we have

$$\{\mathbb{E}_n[\check{w}_i(d_i, x_{i\widehat{T}^*}')'(d_i, x_{i\widehat{T}^*}')]\}_{11}^{-1} = \{\mathbb{E}_n[w_i(d_i, x_{i\widehat{T}^*}')'(d_i, x_{i\widehat{T}^*}')]\}_{11}^{-1} + o_P(1).$$

Next note that

$$\widetilde{\Sigma}_{2n} = \{\mathbb{E}_n[w_i(d_i, x_{i\widehat{T}^*}')'(d_i, x_{i\widehat{T}^*}')]\}_{11}^{-1} = \{\mathbb{E}_n[w_id_i^2] - \mathbb{E}_n[w_id_ix_{i\widehat{T}^*}']\{\mathbb{E}_n[w_ix_{i\widehat{T}^*}x_{i\widehat{T}^*}']\}^{-1}\mathbb{E}_n[w_ix_{i\widehat{T}^*}d_i]\}^{-1}.$$

Note that $\check{\theta}[\widehat{T}^*] = \{\mathbb{E}_n[w_ix_{i\widehat{T}^*}x_{i\widehat{T}^*}']\}^{-1}\mathbb{E}_n[w_ix_{i\widehat{T}^*}d_i]$ is the least squares estimator of regressing $\sqrt{w_i}d_i$ on $\sqrt{w_i}x_{i\widehat{T}^*}$. We let $\check{\theta}$ denote the corresponding p-dimensional (sparse) vector. Therefore, using that $\sqrt{w_i}x_i'\theta_0 = \sqrt{w_i}d_i - v_i$ we have

$$\begin{aligned}
\widetilde{\Sigma}_{2n}^{-2} &= \mathbb{E}_n[w_id_i^2] - \mathbb{E}_n[w_id_ix_i'\check{\theta}] \\
&= \mathbb{E}_n[w_id_i^2] - \mathbb{E}_n[\sqrt{w_i}d_i\sqrt{w_i}x_i'\theta_0] - \mathbb{E}_n[\sqrt{w_i}d_i\sqrt{w_i}x_i'(\check{\theta} - \theta_0)] \\
&= \mathbb{E}_n[\sqrt{w_i}d_iv_i] - \mathbb{E}_n[\sqrt{w_i}d_i\sqrt{w_i}x_i'(\check{\theta} - \theta_0)] \\
&= \mathbb{E}_n[v_i^2] + \mathbb{E}_n[\sqrt{w_i}v_ix_i'\theta_0] - \mathbb{E}_n[\sqrt{w_i}d_i\sqrt{w_i}x_i'(\check{\theta} - \theta_0)]
\end{aligned}$$

We have that $|\mathbb{E}_n[\sqrt{w_i}v_ix_i'\theta_0]| = o_P(\delta_n)$ since $\bar{\mathrm{E}}[\sqrt{w_i}v_ix_i'\theta_0] = 0$ and $\bar{\mathrm{E}}[(\sqrt{w_i}v_ix_i'\theta_0)^2] \leqslant \bar{\mathrm{E}}[w_iv_i^2d_i^2] \leqslant \{\bar{\mathrm{E}}[v_i^4]\bar{\mathrm{E}}[d_i^4]\}^{1/2} \leqslant C$. Moreover, $|\mathbb{E}_n[\sqrt{w_i}d_i\sqrt{w_i}x_i'(\check{\theta} - \theta_0)]| \leqslant \|d_i\|_{2,n}\|\sqrt{w_i}x_i'(\check{\theta} - \theta_0)\|_{2,n} = o_P(\delta_n)$ since $|\widehat{T}^*| \lesssim_P s$ and $\text{support}(\widehat{\theta}_0) \subset \widehat{T}^*$. The result follows.

■

Appendix C. Auxiliary Results for Penalized and Post-Model Selection Estimators

In this section we state relevant theoretical results on the performance of the ℓ_1-penalized Logistic regression estimators, heteroscedastic Lasso with estimated weights estimators and the associated post-model selection estimators. The analysis of the latter builds upon the analysis of Lasso under heteroscedasticity of [2] and it was developed in [5]. The analysis of the former builds upon the work of [1] that established rates for ℓ_1-penalized Logistic regression exploiting self-concordance. The main design condition relies on the restricted eigenvalue proposed in [9], namely for $\tilde{x}_i = (d_i, x_i')'$

$$\kappa_{\mathbf{c}} = \inf_{\|\delta_{T^c}\|_1 \leqslant \mathbf{c}\|\delta_T\|_1} \|f_i \tilde{x}_i'\delta\|_{2,n} / \|\delta_T\|, \tag{C.41}$$

where $\mathbf{c} = (c+1)/(c-1)$ for the slack constant $c > 1$. In the original setting of [9] for least squares we have $f_i = 1$ and it is well known that $\kappa_{\mathbf{c}}$ is bounded away from zero if $\mathbf{c}$ is bounded for any subset $T \subset \{1, \ldots, p\}$ with $|T| \leqslant s$ if the sparse eigenvalues of order Cs are well behaved (bounded away from zero and from above uniformly) for suitably large constant C. When analyzing the logistic regression, the weights will be set to $f_i = \sqrt{w_i}$.

C.1. Results for Lasso and Post Lasso with Estimated Weights.

In this section we state results obtained in [5] for Post-Lasso estimators with estimated weights, namely the model

$$f_i d_i = f_i x_i'\theta_0 + v_i, \quad \mathrm{E}[f_i v_i \mid x_i] = 0 \tag{C.42}$$

where we observe $\{(d_i, x_i) : i = 1, \ldots, n\}$, i.n.i.d., and only an estimate $\widehat{f}_i$ of the weights f_i. The support $T_{\theta_0} = \text{support}(\theta_0)$ is unknown but a sparsity condition holds, namely $|T_{\theta_0}| \leqslant s$. Estimators for θ_0 and v_i can be computed based on Lasso or Post-Lasso, namely

$$\widehat{\theta} \in \arg\min_{\theta \in \mathbb{R}^p} \mathbb{E}_n[\widehat{f}_i^2 (d_i - x_i'\theta)^2] + \frac{\lambda}{n}\|\widehat{\Gamma}\theta\|_1, \quad \text{and} \quad \widehat{v}_i = \widehat{f}_i(d_i - x_i'\widehat{\theta}), \tag{C.43}$$

$$\widetilde{\theta} \in \arg\min_{\theta \in \mathbb{R}^p} \left\{ \mathbb{E}_n[\widehat{f}_i^2 (d_i - x_i'\theta)^2] \ : \ \theta_j = 0, \text{ if } \widehat{\theta}_j = 0 \right\}, \quad \text{and } \tilde{v}_i = \widehat{f}_i(d_i - x_i'\widetilde{\theta}). \tag{C.44}$$

where λ and $\widehat{\Gamma}$ are the associated penalty level and loadings. We will use penalty loadings $\widehat{\Gamma}$ that are diagonal matrices defined by the algorithm below.

Algorithm C.1 (Computation of $\widehat{\Gamma}$)**.**

Step 1. Compute the Post Lasso estimator $\widetilde{\theta}$ based on $\lambda = 2c'\sqrt{n}\Phi^{-1}(1-\gamma/2p)$, $c' > c > 1$ and the following penalty loadings, for $j = 1, \ldots, p$

$$\widehat{\Gamma}_{jj} = \sqrt{\mathbb{E}_n[\widehat{f}_i^2 x_{ij}^2 (\widehat{f}_i d_i - \overline{fd})^2]}, \quad \textit{where} \ \ \overline{fd} := \mathbb{E}_n[\widehat{f}_i d_i].$$

Step 2. Compute the residuals $\widehat{v}_i = \widehat{f}_i(d_i - x_i'\widetilde{\theta})$ and set $\widehat{\Gamma}$ as

$$\widehat{\Gamma}_{jj} = \sqrt{\mathbb{E}_n[\widehat{f}_i^2 x_{ij}^2 \widehat{v}_i^2]}, \ j = 1, \ldots, p. \tag{C.45}$$

[2] established the validity of using either of the choices in (C.45) in the case the weights f_i are known and equal to one and [5] considers the current case with estimated weights $\widehat{f}_i$. Next we provide sufficient high-level conditions to establish rates of convergence and sparsity bounds. As before the sequences Δ_n and δ_n go to zero, C is constant independent of n.

Condition WL. For the model (C.42), normalize $\mathbb{E}_n[x_{ij}^2] = 1$, $j = 1, \dots, p$, suppose that
(i) $\|\theta_0\|_0 \leqslant s$ where $s \geqslant 1$, and the weights satisfy $0 < c \leqslant f_i \leqslant C$ uniformly in n with probability $1 - \Delta_n$,
(ii) $\bar{\mathrm{E}}[v_i^2] > c > 0$, $\max_{j \leqslant p} \dfrac{\{\bar{\mathrm{E}}[|f_i x_{ij} v_i|^3]\}^{1/3}}{\{\bar{\mathrm{E}}[|f_i x_{ij} v_i|^2]\}^{1/2}} \leqslant C$, $\Phi^{-1}(1-\gamma/2p) \leqslant \delta_n n^{1/3}$, $\gamma \leqslant \delta_n$
(iii) $\max_{j \leqslant p} |(\mathbb{E}_n - \bar{\mathrm{E}})[f_i^2 x_{ij}^2 v_i^2]| + \max_{j \leqslant p} |(\mathbb{E}_n - \bar{\mathrm{E}})[f_i^2 x_{ij}^2 f_i^2 d_i^2]| \leqslant \delta_n$, with probability $1 - \Delta_n$
(iv) the estimates $\widehat{f}_i, i = 1, \dots, n$, satisfy with probability $1 - \Delta_n$

$$\max_{j \leqslant p} \mathbb{E}_n[(\widehat{f}_i - f_i)^2 x_{ij}^2 v_i^2] \leqslant \delta_n, \quad \text{and} \quad \mathbb{E}_n\left[\frac{(\widehat{f}_i^2 - f_i^2)^2}{f_i} v_i^2\right] \leqslant \widehat{c}_f^2.$$

Condition WL(i) is a standard sparsity assumption and could be relaxed in different directions. Condition WL(ii) has mild moment conditions that are used to apply self-normalized moderate deviation theory to control heteroscedastic non-Gaussian errors similar to [2] where there are no estimated weights. Condition WL(ii) also has a condition on the dimension p relative to n and bounds how fast the confidence level $1-\gamma$ can converge to 1. Condition WL(iii) provides sufficient conditions for the uniform convergence of cross terms. Condition WL(iv) requires high-level rates of convergence for the estimate $\widehat{f}_i$. These estimates can be constructed with ℓ_1-penalized logistic regression estimators studied in Section C.2.

Comment C.1 (Control of $\widehat{c}_f$)**.** The quantity $\widehat{c}_f$ impacts directly the prediction rate and sparsity results which are needed for the post-model selection estimators. Bounds on $\widehat{c}_f$ will be dependent on regularities conditions. A simple bound is $\widehat{c}_f^2 \leqslant \mathbb{E}_n\left[(\widehat{f}_i^2 - f_i^2)^2\right] \max_{i \leqslant n} v_i^2/f_i$. In our analysis we pursued the use of matrix inequalities based on [26] which seems to lead to sharper results under typical conditions.

Next we present results on the performance of the estimators generated by Lasso and Post-Lasso with estimated weights.

Theorem 3 (Properties of Lasso and Post-Lasso with estimated Weights)**.** *Under Condition WL, setting $\lambda \geqslant 2c'\sqrt{n}\Phi^{-1}(1-\gamma/2p)$ for $c' > c > 1$, and using the penalty loadings $\widehat{\Gamma}$ defined in (C.45), there is an uniformly bounded* $\mathbf{c}$ *such that*

$$\|\widehat{f}_i x_i'(\widehat{\theta} - \theta_0)\|_{2,n} \lesssim_P \frac{\lambda\sqrt{s}}{n\kappa_{\mathbf{c}} \min_{i \leqslant n} \widehat{f}_i/f_i} + \widehat{c}_f$$

Moreover, provided that $\phi_{\max}(\{s + n^2\widehat{c}_f^2/\lambda^2\}/\delta_n) \leqslant C$, $\min_{i \leqslant n} \widehat{f}_i^2 \geqslant c/2$, the data-dependent model $\widehat{T}_{\theta_0}$ selected by a Lasso estimator satisfies with probability $1 - o(1)$:

$$\|\widetilde{\theta}\|_0 = |\widehat{T}_{\theta_0}| \lesssim s + \frac{n^2\widehat{c}_f^2}{\lambda^2} \tag{C.46}$$

and the Post-Lasso estimator obeys

$$\|x_i'(\widetilde{\theta} - \theta_0)\|_{2,n} \lesssim_P \frac{n\widehat{c}_f}{\lambda}\sqrt{\frac{\log p}{n}} + \sqrt{\frac{s\log(p \vee n)}{n}} + \frac{\lambda\sqrt{s}}{n\kappa_{\mathbf{c}}} \quad \textit{and}$$

$$\|\widetilde{\theta}-\theta_0\|_1 \lesssim_P \left\{\sqrt{s}+\frac{n\widehat{c}_f}{\lambda}\right\}\frac{\|x_i'(\widetilde{\theta}-\theta_0)\|_{2,n}}{\sqrt{\phi_{\min}(|\widehat{T}_{\theta_0}|)}}.$$

Theorem 3 above establishes the rate of convergence for Lasso and Post-Lasso with estimated weights. This leads to bounds on the error between estimated the instrumented instrument $\widehat{z}_i$ used in Table 1 with respect to the associated valid instrument $z_{0i} = v_i/\sqrt{w_i}$ since

$$\widehat{z}_i - z_{0i} = d_i - x_i'\widetilde{\theta} - \frac{v_i}{\sqrt{w_i}} = d_i - x_i'\widetilde{\theta} - \{d_i - x_i'\theta_0\} = x_i'(\theta_0-\widetilde{\theta}). \tag{C.47}$$

Sparsity properties of the Lasso estimator $\widehat{\theta}$ under estimated weights follows similarly to the standard Lasso analysis derived in [2]. By combining such sparsity properties and the rates in the prediction norm we can establish rates for the post-model selection estimator under estimated weights.

C.2. **ℓ_1-Penalized Logistic Regression.** Consider a data generating process such that

$$\mathrm{E}[y_i \mid \tilde{x}_i] = G(\tilde{x}_i'\eta_0)$$

which is independent across i ($i = 1,\dots,n$). Without loss of generality, we assume that $\|\eta_0\|_0 = s \geqslant 1$, $\mathbb{E}_n[\tilde{x}_{ij}^2] = 1$ for all $1 \leqslant j \leqslant p$. First we consider the estimation of η_0 via ℓ_1-penalized Logistic regression

$$\widehat{\eta} \in \arg\min_{\eta} \Lambda(\eta) + \frac{\lambda}{n}\|\eta\|_1. \tag{C.48}$$

Following a general principle used in ℓ_1-penalized estimators as discussed in [9, 1, 3, 18, 32], under the event that

$$\frac{\lambda}{n} \geqslant c\|\nabla\Lambda(\eta_0)\|_\infty = c\|\mathbb{E}_n[\{y_i - G(\tilde{x}_i'\eta_0)\}\tilde{x}_i]\|_\infty, \quad \text{where } c>1, \tag{C.49}$$

the estimator in (C.48) achieves good theoretical guarantees under mild design conditions. Although η_0 is unknown, we can set λ so that the event in (C.49) holds with high probability. In particular, Remark E.1 based on Lemma 12 shows that it suffices to set $\lambda = \frac{1.1}{2}\sqrt{n}\Phi^{-1}(1-\gamma/[2p])$ where we suggest $\gamma = 0.1/\log n$. Next we present results for the estimator (C.48). In what follows we consider (C.41) with $f_i = \sqrt{w_i}$.

Lemma 2 (Results for ℓ_1-Penalized Logistic Regression)**.** *Assume $\lambda/n \geqslant c\|\nabla\Lambda(\eta_0)\|_\infty$, $c>1$ and let $\mathbf{c} = (c+1)/(c-1)$. Then*

$$\|\sqrt{w_i}\tilde{x}_i'(\widehat{\eta}-\eta_0)\|_{2,n} \leqslant 3(1+\tfrac{1}{c})\frac{\lambda\sqrt{s}}{n\kappa_{\mathbf{c}}} \quad \textit{and} \quad \|\widehat{\eta}-\eta_0\|_1 \leqslant 3\frac{(1+c)(1+\mathbf{c})}{c}\frac{\lambda s}{n\kappa_{\mathbf{c}}^2}$$

provided that $\inf_{\delta\in\Delta_{\mathbf{c}}} \frac{\|\sqrt{w_i}\tilde{x}_i'\delta\|_{2,n}^3}{\|\sqrt{w_i}|\tilde{x}_i'\delta|^{3/2}\|_{2,n}^2} > 3(1+\frac{1}{c})\frac{\lambda\sqrt{s}}{n\kappa_{\mathbf{c}}}$. *Moreover, we have*

$$|\mathrm{support}(\widehat{\eta})| \leqslant 36s\mathbf{c}^2\frac{\min_{m\in\mathcal{M}}\phi_{\max}(m)}{\kappa_{\mathbf{c}}^2} \qquad \textit{and} \quad \Lambda(\widehat{\eta})-\Lambda(\eta_0) \leqslant 3(1+\tfrac{1}{c})\left(\frac{\lambda\sqrt{s}}{n\kappa_{\mathbf{c}}}\right)^2$$

where $\mathcal{M} = \{m\in\mathbf{N} : m > 72\mathbf{c}^2 s\phi_{\max}(m)/\kappa_{\mathbf{c}}^2\}$, *provided* $\max_{i\leqslant n}\|\tilde{x}_i\|_\infty\|\widehat{\eta}-\eta_0\|_1 \leqslant 1$.

The extra growth condition required for identification is mild. For instance we typically have $\lambda \lesssim \sqrt{\log(n\vee p)/n}$ and, if the weights w_i are bounded away from zero, for many designs of interest we have

$\inf_{\delta\in\Delta_{\mathbf{c}}} \|\tilde{x}_i'\delta\|_{2,n}^3/\mathbb{E}_n[|\tilde{x}_i'\delta|^3]$ bounded away from zero (see [3]). For more general designs and weights we have

$$\inf_{\delta\in\Delta_{\mathbf{c}}} \frac{\|\sqrt{w_i}\tilde{x}_i'\delta\|_{2,n}^3}{\mathbb{E}_n[w_i|\tilde{x}_i'\delta|^3]} \geqslant \inf_{\delta\in\Delta_{\mathbf{c}}} \frac{\|\sqrt{w_i}\tilde{x}_i'\delta\|_{2,n}}{\max_{i\leqslant n}\|\tilde{x}_i\|_\infty\|\delta\|_1} \geqslant \frac{\kappa_{\mathbf{c}}}{\sqrt{s}(1+\mathbf{c})\max_{i\leqslant n}\|\tilde{x}_i\|_\infty}$$

which implies the extra growth condition under $K_x^2 s^2 \log(p\vee n) \leqslant \delta_n \kappa_{\mathbf{c}}^2 n$. Under the condition that s/δ_n-sparse eigenvalues are bounded away from zero and from above, it follows that $s/\delta_n^{1/2}$ belongs to $\mathcal{M}$ for n large enough so that $|\text{support}(\widehat{\eta})| \lesssim s$ under the conditions above.

In order to alleviate the bias introduced by the ℓ_1-penalty, we can consider the associated post-model selection estimates. Let $\widehat{T}^*$ denote a subset of covariates (selected arbitrarily) and define the associated post-model selection estimator

$$\widetilde{\eta} \in \arg\min_{\eta} \left\{ \Lambda(\eta) : \eta_j = 0 \text{ if } j \notin \widehat{T}^* \right\}. \tag{C.50}$$

Typically $\widehat{T}^*$ can be taken as support$(\widehat{\eta})$. However, we can add additional variables through other procedures. (For example, in Step 1 we always include the treatment d_i; in Step 3 of the double selection procedure covariates selected in a different equation are included.) The following result characterizes the performance of the estimator in (C.50).

Lemma 3 (Estimation Error of Post-ℓ_1-penalized Logistic Regression)**.** *Let $\widehat{s}^* = |\widehat{T}^*|$. We have*

$$\|\sqrt{w_i}\tilde{x}_i'(\tilde{\eta}-\eta_0)\|_{2,n} \leqslant \frac{3\sqrt{\widehat{s}^*}\|\nabla\Lambda(\eta_0)\|_\infty}{\sqrt{\phi_{\min}(\widehat{s}+s)}} + 3\sqrt{\Lambda(\widetilde{\eta})-\Lambda(\eta_0)}$$

provided that

$$\inf_{\|\delta\|_0\leqslant \widehat{s}^*+s} \frac{\|\sqrt{w_i}\tilde{x}_i'\delta\|_{2,n}^3}{\|\sqrt{w_i}|\tilde{x}_i'\delta|^{3/2}\|_{2,n}^2} > 6\max\left\{ \sqrt{\widehat{s}^*+s}\frac{\|\nabla\Lambda(\eta_0)\|_\infty}{\sqrt{\phi_{\min}(\widehat{s}^*+s)}},\quad \sqrt{\Lambda(\widetilde{\eta})-\Lambda(\eta_0)} \right\}.$$

Lemma 3 provides the rate of convergence in the prediction norm for the post model selection estimator despite the possible imperfect model selection. The rates rely on the overall quality of the selected model and the overall number of components $\widehat{s}^*$. Once again, based on the results in Lemma 2, the extra growth condition required for identification is mild provided that support$(\widehat{\eta}) \subset \widehat{T}^*$ and $\widehat{s}^*$ is not much larger than s.

Comment C.2. In Step 1 of the algorithms, we use ℓ_1-penalized Logistic regression with $\tilde{x}_i = (d_i, x_i')'$, $\widehat{\delta} := \widehat{\eta} - \eta_0 = (\widehat{\alpha}-\alpha_0, \widehat{\beta}'-\beta_0')'$, and we are interested on rates for $\|x_i'(\widehat{\beta}-\beta_0)\|_{2,n}$ instead of $\|\tilde{x}_i'\widehat{\delta}\|_{2,n}$. However, it follows that

$$\|x_i'(\widehat{\beta}-\beta_0)\|_{2,n} \leqslant \|\tilde{x}_i'\widehat{\delta}\|_{2,n} + |\widehat{\alpha}-\alpha_0|\cdot\|d_i\|_{2,n}.$$

Since $s \geqslant 1$, without loss of generality we can assume the component associated with the treatment d_i belongs to T (at the cost of increasing the cardinality of T by one which will not affect the rate of convergence). Therefore we have that

$$|\widehat{\alpha}-\alpha_0| \leqslant \|\widehat{\delta}_T\| \leqslant \|\sqrt{w_i}\tilde{x}_i'\widehat{\delta}\|_{2,n}/\kappa_{\mathbf{c}}.$$

In most applications of interest $\|d_i\|_{2,n}$ and $1/\kappa_{\mathbf{c}}$ are bounded from above with high probability. Similarly, in Step 1 of Algorithm 1 we have that the Post-ℓ_1-Logistic estimator satisfies

$$\|x_i'(\widetilde{\beta}-\beta_0)\|_{2,n} \leqslant \|\tilde{x}_i'\widetilde{\delta}\|_{2,n}\left(1+\|d_i\|_{2,n}/\sqrt{\phi_{\min}(\widehat{s}+s)}\right).$$

Appendix D. Auxiliary Inequalities

Lemma 4. *Fix arbitrary vectors $x_1, \ldots, x_n \in \mathbb{R}^p$ with $\max_{i \leqslant n} \|x_i\|_\infty \leqslant K_x$. Let ζ_i $(i = 1, \ldots, n)$ be independent random variables such that $\bar{\mathrm{E}}[|\zeta_i|^q] < \infty$ for some $q \geqslant 4$. Then we have with probability $1 - 8\tau$*

$$\max_{1 \leqslant j \leqslant p} |(\mathbb{E}_n - \bar{\mathrm{E}})[x_{ij}^2 \zeta_i^2]| \leqslant 4\sqrt{\frac{\log(2p/\tau)}{n}} K_x^2 (\bar{\mathrm{E}}[|\zeta_i|^q]/\tau)^{4/q}$$

Proof. The result is derived in Lemma 2 of [6] which follows from a maximal inequality derived in [7]. ∎

Consider an empirical process $\mathbb{G}_n(f) = n^{-1/2} \sum_{i=1}^n \{f(Z_i) - \mathrm{E}[f(Z_i)]\}$ indexed by $\mathcal{F}$, a class of pointwise measurable functions (see [30] Chapter 2.3) and assume that $0 \in \mathcal{F}$. The random empirical measure for an underlying independent data sequence $\{Z_i, i = 1, \ldots, n\}$ is denoted by $\mathbb{P}_n$.

Lemma 5. *Let $|h_i(t)| \leqslant |t'W_i|$, $K/4 > \bar{\sigma}^2 := \sup_{t \in \mathcal{T}} \bar{\mathrm{E}}[h_i(t)^2 \xi_i^2]$, $\tilde{p} = \dim(W_i)$, and $\|\mathcal{T}\|_1 = \sup_{t \in \mathcal{T}} \|t\|_1$. We have*

$$\mathrm{E}\left[\sup_{t \in \mathcal{T}} |(\mathbb{E}_n - \bar{\mathrm{E}})[h_i(t)\xi_i]|\right] \leqslant 4\|\mathcal{T}\|_1 \mathrm{E}\left[\|\mathbb{E}_n[\varepsilon_i W_i \xi_i]\|_\infty\right] \quad \textit{and}$$

$$P\left(\sup_{t \in \mathcal{T}} |(\mathbb{E}_n - \bar{\mathrm{E}})[h_i(t)\xi_i]| > \frac{K\|\mathcal{T}\|_1 \sqrt{M}}{\sqrt{n}}\right) \leqslant 32\tilde{p} \exp\left(\frac{-K^2}{16}\right) + P\left(\max_{j \leqslant \tilde{p}} \mathbb{E}_n[W_{ij}^2 \xi_i^2] > M\right).$$

Proof. To establish the first relation, by symmetrization for expectation Lemma 6.3 in [13]

$$\mathrm{E}\left[\sup_{t \in \mathcal{T}} |(\mathbb{E}_n - \bar{\mathrm{E}})[h_i(t)\xi_i]|\right] \leqslant 2\mathrm{E}\left[\sup_{t \in \mathcal{T}} |\mathbb{E}_n[\varepsilon_i h_i(t)\xi_i]|\right]$$

and Contraction principle Lemma 4.12 in [13] we have

$$\mathrm{E}\left[\sup_{t \in \mathcal{T}} |(\mathbb{E}_n - \bar{\mathrm{E}})[h_i(t)\xi_i]|\right] \leqslant 4\mathrm{E}\left[\sup_{t \in \mathcal{T}} |\mathbb{E}_n[\varepsilon_i t' W_i \xi_i]|\right] \leqslant 4 \sup_{t \in \mathcal{T}} \|t\|_1 \mathrm{E}\left[\|\mathbb{E}_n[\varepsilon_i W_i \xi_i]\|_\infty\right].$$

By Lemma 2.3.7 in [31], symmetrization for probabilities, we have

$$P\left(\sup_{t \in \mathcal{T}} |\mathbb{G}_n(h_i(t)\xi_i)| > K\right) \leqslant \frac{2}{1 - (\bar{\sigma}^2/K^2)} P\left(\sup_{t \in \mathcal{T}} |\mathbb{G}_n(\varepsilon_i h_i(t)\xi_i)| > K/4\right)$$

since $\mathrm{var}(\mathbb{G}_n(h_i(t)\xi_i)) \leqslant \bar{\mathrm{E}}[h_i(t)^2 \xi_i^2] \leqslant \bar{\sigma}^2$. Conditional on $\{\xi, W_i\}$, also using Contraction principle Lemma 4.12 in [13] we have

$$\begin{aligned}
\mathrm{E}[\exp(\psi \sup_{t \in \mathcal{T}} |\mathbb{G}_n(\varepsilon_i h_i(t)\xi_i)|)] &\leqslant \mathrm{E}[\exp(4\psi \sup_{t \in \mathcal{T}} \|t\|_1 \|\mathbb{G}_n(\varepsilon_i W_i \xi_i)\|_\infty)] \\
&\leqslant \tilde{p} \cdot \max_{j \leqslant \tilde{p})} \mathrm{E}[\exp\{4\psi \sup_{t \in \mathcal{T}} \|t\|_1 |\mathbb{G}_n(\varepsilon_i W_{ij} \xi_i)|\}] \\
&\leqslant 2\tilde{p} \cdot \exp(8\psi^2 \sup_{t \in \mathcal{T}} \|t\|_1^2 \max_{j \leqslant \tilde{p}} \mathbb{E}_n[W_{ij}^2 \xi_i^2])
\end{aligned}$$

Since we have that $P(X > K) \leqslant \min_{\psi \geqslant 0} \exp(-\psi K) \mathrm{E}[\exp(\psi X)]$, by choosing the parameter ψ as $\psi = K/\{16\|\mathcal{T}\|_1^2 \max_{j \leqslant \tilde{p}} \mathbb{E}_n[W_{ij}^2 \xi_i^2]\}$ it follows

$$P_\varepsilon\left(\sup_{t \in \mathcal{T}} |\mathbb{G}_n(\varepsilon_i h_i(t)\xi_i)| > K \mid h_i, W_i, \xi_i\right) \leqslant 8\tilde{p} \exp(-K^2/\{16 \sup_{t \in \mathcal{T}} \|t\|_1^2 \max_{j \leqslant \tilde{p}} \mathbb{E}_n[W_{ij}^2 \xi_i^2]\})$$

The result follows by taking the expectation conditioned on $\{\max_{j \leqslant \tilde{p}} \mathbb{E}_n[W_{ij}^2 \xi_i^2] \leqslant M\}$. ∎

Lemma 6 (Essentially in Theorem 3.6 of [26]). *Let X_i, $i = 1, \ldots, n$, be independent random vectors in $\mathbb{R}^p$ be such that $\sqrt{\mathrm{E}[\max_{1 \leqslant i \leqslant n} \|X_i\|_\infty^2]} \leqslant K$. Let*

$$\delta_n := 2\left(\bar{C}K\sqrt{k}\log(1+k)\sqrt{\log(p \vee n)}\sqrt{\log n}\right)/\sqrt{n},$$

where $\bar{C}$ is the universal constant. Then,

$$\mathrm{E}\left[\sup_{\|\theta\|_0 \leqslant k, \|\theta\|=1} \left|\mathbb{E}_n\left[(\theta' X_i)^2 - \mathrm{E}[(\theta' X_i)^2]\right]\right|\right] \leqslant \delta_n^2 + \delta_n \sup_{\|\theta\|_0 \leqslant k, \|\theta\|=1} \sqrt{\bar{\mathrm{E}}[(\theta' X_i)^2]}.$$

References

[1] Francis Bach. Self-concordant analysis for logistic regression. *Eletronic Journal of Statistics*, 4:384–414, 2010.

[2] A. Belloni, D. Chen, V. Chernozhukov, and C. Hansen. Sparse models and methods for optimal instruments with an application to eminent domain. *Econometrica*, November 2012.

[3] A. Belloni and V. Chernozhukov. ℓ_1-penalized quantile regression for high dimensional sparse models. *Annals of Statistics*, 39(1):82–130, 2011.

[4] A. Belloni, V. Chernozhukov, and C. Hansen. Lasso methods for gaussian instrumental variables models. arXiv:[math.ST], http://arxiv.org/abs/1012.1297, 2010.

[5] A. Belloni, V. Chernozhukov, and K. Kato. Robust inference in high-dimensional approximately sparse quantile regression models. *working paper*, 2013.

[6] A. Belloni, V. Chernozhukov, and K. Kato. Uniform post model selection inference for LAD regression models. *ArXiv*, 2013.

[7] Alexandre Belloni, Victor Chernozhukov, and Christian Hansen. Inference on treatment effects after selection amongst high-dimensional controls. *ArXiv*, 2011.

[8] Alexandre Belloni, Victor Chernozhukov, and Christian Hansen. Inference methods for high-dimensional sparse econometric models. *Advances in Economics and Econometrics, 10th World Congress of Econometric Society*, Volume III, Econometrics, Edited by Daron Acemoglu, Manuel Arellano and Eddie Dekel:245–295, 2013.

[9] P. J. Bickel, Y. Ritov, and A. B. Tsybakov. Simultaneous analysis of Lasso and Dantzig selector. *Annals of Statistics*, 37(4):1705–1732, 2009.

[10] F. Bunea. Honest variable selection in linear and logistic regressionmodels via ℓ_1 and $\ell_1 + \ell_2$ penalization. *Electronic Journal of Statistics*, 2:11531194, 2008.

[11] Michael R. Kosorok. *Introduction to Empirical Processes and Semiparametric Inference.* Series in Statistics. Springer, Berlin, 2008.

[12] M. Kwemou. Non-asymptotic oracle inequalities for the lasso and group lasso in high dimensional logistic model. *arXiv preprint*, (arXiv:1206.0710), 2012.

[13] M. Ledoux and M. Talagrand. *Probability in Banach Spaces (Isoperimetry and processes).* Ergebnisse der Mathematik undihrer Grenzgebiete, Springer-Verlag, 1991.

[14] Hannes Leeb and Benedikt M. Pötscher. Model selection and inference: facts and fiction. *Economic Theory*, 21:21–59, 2005.

[15] Hannes Leeb and Benedikt M. Pötscher. Can one estimate the conditional distribution of post-model-selection estimator? *The Annals of Statistics*, 34(5):2554–2591, 2006.

[16] Hannes Leeb and Benedikt M. Pötscher. Sparse estimators and the oracle property, or the return of Hodges' estimator. *J. Econometrics*, 142(1):201–211, 2008.

[17] L. Meier, V. Van der Geer, and P. Bühlmann. The group lasso for logistic regression. *J. R. Statist. Soc.: Series B (Statist. Methodol.)*, 70(1):5371, 2008.

[18] Sahand N. Negahban, Pradeep Ravikumar, Martin J. Wainwright, and Bin Yu. A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. *Statistical Science*, 27(4):538–557, 2012.

[19] J. Neyman. Optimal asymptotic tests of composite statistical hypotheses. In U. Grenander, editor, *Probability and Statistics, the Harold Cramer Volume*. New York: John Wiley and Sons, Inc., 1959.

[20] J. Neyman. $c(\alpha)$ tests and their use. *Sankhya*, 41:1–21, 1979.

[21] Yaniv Plan and Roman Vershynin. Robust 1-bit compressed sensing and sparse logistic regression: A convex programming approach. *IEEE Transactions on Information Theory*, 59(1):482 – 494, Jan. 2013.

[22] Benedikt M. Pötscher. Confidence sets based on sparse estimators are necessarily large. *Sankhyā*, 71(1, Ser. A):1–18, 2009.

[23] Benedikt M. Pötscher and Hannes Leeb. On the distribution of penalized maximum likelihood estimators: the LASSO, SCAD, and thresholding. *J. Multivariate Anal.*, 100(9):2065–2082, 2009.

[24] P. Ravikumar, M. Wainwright, and J. Lafferty. High-dimensional ising model selection using -regularized logistic regression. *Ann. Statist.*, 38(2):12871319, 2010.

[25] M. Rudelson and S. Zhou. Reconstruction from anisotropic random measurements. *ArXiv:1106.1151*, 2011.

[26] Mark Rudelson and Roman Vershynin. On sparse reconstruction from fourier and gaussian measurements. *Communications on Pure and Applied Mathematics*, 61:10251045, 2008.

[27] R. Tibshirani. Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B*, 58:267–288, 1996.

[28] S. A. van de Geer. High-dimensional generalized linear models and the lasso. *Annals of Statistics*, 36(2):614–645, 2008.

[29] Sara Anna van de Geer, Peter Bühlmann, and Ya'acov Ritov. On asymptotically optimal confidence regions and tests for high-dimensional models. *ArXiv*, 2013.

[30] A. W. van der Vaart and J. A. Wellner. *Weak Convergence and Empirical Processes*. Springer Series in Statistics, 1996.

[31] Aad W. van der Vaart and Jon A. Wellner. Empirical process indexed by estimated functions. *IMS Lecture Notes-Monograph Series*, 55:234–252, 2007.

[32] Lie Wang. L_1 penalized lad estimator for high dimensional ilnear regression. *ArXiv*, 2012.

[33] Cun-Hui Zhang and Stephanie S. Zhang. Confidence intervals for low-dimensional parameters with high-dimensional data. *ArXiv.org*, (arXiv:1110.2563v1), 2011.

Supplementary Appendix for "Honest Confidence Regions for Logistic Regression with a Large Number of Controls"

Appendix E. Technical Results and Proofs for Logistic Regression

In this section our goal is to establish sparsity and rates of convergence of the Post-Lasso Logistic estimator. Both of these properties require us to also revisit the analysis of the ℓ_1-penalize logistic regression (Lasso-Logistic) estimator. In what follows we use a more compact notation, specifically $\eta = (\alpha, \beta)$, $\tilde{x}_i = (d_i, x_i')'$, $\eta_0 = (\alpha_0, \beta_0')'$. Thus the Lasso-Logistic estimator is defined as any vector $\widehat{\eta}$ such that

$$\widehat{\eta} \in \arg\min_{\eta} \Lambda(\eta) + \frac{\lambda}{n}\|\eta\|_1. \tag{E.51}$$

We will also consider the post-model selection Logistic estimator associated with a support $\widehat{T}^* \subset \{1, \ldots, p\}$ defined as

$$\widetilde{\eta} \in \arg\min_{\eta} \Lambda(\eta) \quad : \quad \text{support}(\eta) \subseteq \widehat{T}^*. \tag{E.52}$$

E.1. **Design conditions and Relations.** Next we collect relevant quantities associated with the design matrix $\mathbb{E}_n[\tilde{x}_i\tilde{x}_i']$ and the weighted counterpart $\mathbb{E}_n[w_i\tilde{x}_i\tilde{x}_i']$ where $w_i = G_i(1-G_i) \in [0,1]$, $G_i = G(\tilde{x}_i'\eta_0)$, $i = 1, \ldots, n$, is the conditional variance of the outcome variable y_i. The non-weighted quantities are well studied in the literature (namely restricted eigenvalue, minimum and maximal sparse eigenvalues).

Definition 1. For $T = \text{support}(\eta_0)$, $|T| \geqslant 1$, the (logistic) restricted eigenvalue is defined as

$$\kappa_{\mathbf{c}} := \min_{\|\delta_{T^c}\|_1 \leqslant \mathbf{c}\|\delta_T\|_1} \frac{\|\sqrt{w_i}\tilde{x}_i'\delta\|_{2,n}}{\|\delta_T\|}$$

Definition 2. For a subset $A \subset \mathbb{R}^p$ let the non-linear impact coefficient be defined as

$$\bar{q}_A = \inf_{\delta \in A} \mathbb{E}_n\left[w_i|\tilde{x}_i'\delta|^2\right]^{3/2} / \ \mathbb{E}_n\left[w_i|\tilde{x}_i'\delta|^3\right].$$

In this work we will apply this for $A = \Delta_{\mathbf{c}}$ and $A = \{\delta \in \mathbb{R}^p : \|\delta\|_0 \leqslant Cs\}$.

The definitions above differ from their counterpart in the analysis of ℓ_1-penalized least squares estimators by the weighting $0 \leqslant w_i \leqslant 1$. Thus it will be relevant to understand their relations through the quantities

$$\psi_{(r)}(\mathbf{c}) := \min_{\|\delta_{T^c}\|_1 \leqslant \mathbf{c}\|\delta_T\|_1} \frac{\|\sqrt{w_i}\tilde{x}_i'\delta\|_{2,n}}{\|\tilde{x}_i'\delta\|_{2,n}} \quad \text{and} \quad \psi_{(s)}(m) := \min_{1 \leqslant \|\delta\|_0 \leqslant m} \frac{\|\sqrt{w_i}\tilde{x}_i'\delta\|_{2,n}}{\|\tilde{x}_i'\delta\|_{2,n}}$$

Lemma 7 provides three relationships between the weighted versions and the non-weighted versions. Neither dominates the other. Most papers in the literature focus on the first pair of relations which entails assuming that $\min_{i \leqslant n} w_i$ is bounded away from zero uniformly in n. The second and third pairs of relations allow for better control in the presence of a few small weights. The second pair states that if the average harmonic mean of the weights is bounded the ratio between the weighted and non-weighted quantities is controlled by the intrinsic sparsity.

Lemma 7 (Relating weighted and non-weighted design quantities)**.** *Letting* $w_i = G_i(1-G_i)$ *we have the following inequalities* $\psi_{(r)}(\mathbf{c}) \geqslant \min_{i\leqslant n}\sqrt{w_i}$ *and* $\psi_{(s)}(m) \geqslant \min_{i\leqslant n}\sqrt{w_i}$*;*

$$\psi_{(r)}(\mathbf{c}) \geqslant \frac{\kappa^u_{\mathbf{c}}\{\mathbb{E}_n[1/w_i]\}^{-1/2}}{\sqrt{s}(1+\mathbf{c})\max_{i\leqslant n}\|\tilde{x}_i\|_\infty} \quad \textit{and} \quad \psi_{(s)}(m) \geqslant \frac{\sqrt{\phi_{\min}(m)}\{\mathbb{E}_n[1/w_i]\}^{-1/2}}{\sqrt{m}\max_{i\leqslant n}\|\tilde{x}_i\|_\infty}.$$

where $\kappa^u_{\mathbf{c}}$ *is the original (non-weighted) restricted eigenvalue. Moreover, for any* $\epsilon \in (0,1]$ *we have*

$$\psi_{(r)}(\mathbf{c}) \geqslant \sqrt{\epsilon}\kappa^u_{\mathbf{c}}\left\{1-\mathbb{E}_n[1\{w_i\leqslant \epsilon\}]\frac{s(1+\mathbf{c})^2\max_{i\leqslant n}\|\tilde{x}_i\|^2_\infty}{\kappa^{u2}_{\mathbf{c}}}\right\}^{1/2} \quad \textit{and}$$

$$\psi_{(s)}(m) \geqslant \sqrt{\epsilon}\sqrt{\phi_{\min}(m)}\left\{1-\mathbb{E}_n[1\{w_i\leqslant \epsilon\}]\frac{m\max_{i\leqslant n}\|\tilde{x}_i\|^2_\infty}{\phi_{\min}(m)}\right\}^{1/2}.$$

Proof. The first pair of bounds is trivial since $w_i \geqslant 0$. To show the second pair we have

$$\begin{aligned}\mathbb{E}_n[|\tilde{x}_i'\delta|^2] &= \mathbb{E}_n[\sqrt{w_i}|\tilde{x}_i'\delta|\cdot|\tilde{x}_i'\delta|/\sqrt{w_i}]\\ &\leqslant \{\mathbb{E}_n[w_i|\tilde{x}_i'\delta|^2]\}^{1/2}\cdot\{\mathbb{E}_n[|\tilde{x}_i'\delta|^2/w_i]\}^{1/2}\\ &\leqslant \{\mathbb{E}_n[w_i|\tilde{x}_i'\delta|^2]\}^{1/2}\cdot\{\mathbb{E}_n[1/w_i]\}^{1/2}\|\delta\|_1\max_{i\leqslant n}\|\tilde{x}_i\|_\infty\end{aligned}$$

Therefore, for $\vartheta_\delta = \|\tilde{x}_i'\delta\|_{2,n}/\|\delta\|_1$ we have

$$\begin{aligned}\frac{\|\sqrt{w_i}\tilde{x}_i'\delta\|_{2,n}}{\|\tilde{x}_i'\delta\|_{2,n}} &\geqslant \frac{\|\sqrt{w_i}\tilde{x}_i'\delta\|_{2,n}}{\|\sqrt{w_i}\tilde{x}_i'\delta\|^{1/2}_{2,n}\cdot\{\mathbb{E}_n[1/w_i]\}^{1/4}\|\delta\|_1^{1/2}\max_{i\leqslant n}\|\tilde{x}_i\|^{1/2}_\infty}\\ &= \frac{\|\sqrt{w_i}\tilde{x}_i'\delta\|^{1/2}_{2,n}}{\|\tilde{x}_i'\delta\|^{1/2}_{2,n}}\frac{\vartheta_\delta^{1/2}}{\max_{i\leqslant n}\|\tilde{x}_i\|^{1/2}_\infty}\frac{1}{\{\mathbb{E}_n[1/w_i]\}^{1/4}}\end{aligned}$$

By cancelling out $\|\sqrt{w_i}\tilde{x}_i'\delta\|^{1/2}_{2,n}/\|\tilde{x}_i'\delta\|^{1/2}_{2,n}$ and squaring both sides we have

$$\frac{\|\sqrt{w_i}\tilde{x}_i'\delta\|_{2,n}}{\|\tilde{x}_i'\delta\|_{2,n}} \geqslant \vartheta_\delta/\max_{i\leqslant n}\|\tilde{x}_i\|_\infty.$$

The result follows by noting that for $\delta \in \Delta_{\mathbf{c}}$ we have $\vartheta_\delta \geqslant \kappa^u_{\mathbf{c}}/\{(1+\mathbf{c})\sqrt{s}\}$ and for any non-zero δ with $\|\delta\|_0 \leqslant m$ we have $\vartheta_\delta \geqslant \sqrt{\phi_{\min}(m)}/\sqrt{m}$.

The third pair follows from noting that

$$\mathbb{E}_n[w_i|\tilde{x}_i'\delta|^2] = \mathbb{E}_n[w_i 1\{w_i > \epsilon\}|\tilde{x}_i'\delta|^2] + \mathbb{E}_n[w_i 1\{w_i \leqslant \epsilon\}|\tilde{x}_i'\delta|^2] \geqslant \epsilon\mathbb{E}_n[|\tilde{x}_i'\delta|^2] - \epsilon\mathbb{E}_n[1\{w_i\leqslant\epsilon\}|\tilde{x}_i'\delta|^2]$$

Moreover, by definition of ϑ_δ we have

$$\mathbb{E}_n[1\{w_i\leqslant\epsilon\}|\tilde{x}_i'\delta|^2] \leqslant \mathbb{E}_n[1\{w_i\leqslant\epsilon\}]\max_{i\leqslant n}\|\tilde{x}_i\|^2_\infty\|\delta\|_1^2 \leqslant \mathbb{E}_n[1\{w_i\leqslant\epsilon\}]\max_{i\leqslant n}\|\tilde{x}_i\|^2_\infty\frac{\|\tilde{x}_i'\delta\|^2_{2,n}}{\vartheta^2_\delta}.$$

The result follows. ∎

E.2. **Identification Lemmas.** In this section we collect new identification results for Logistic regression that might be of independent interest. We build upon the following technical lemma of [1] which is based on (modified) self-concordant functions. However we will apply it differently than in [1]. We exploit the separability of the objective function across observations and make use of the restricted non-linear impact coefficient [3]. In turn this allows us to weaken requirements of the analysis when compared to the literature.

Lemma 8 (Lemma 1 from [1])**.** *Let $g : \mathbb{R} \to \mathbb{R}$ be a convex three times differentiable function such that for all $t \in \mathbb{R}$, $|g'''(t)| \leqslant Mg''(t)$ for some $M \geqslant 0$. Then, for all $t \geqslant 0$ we have*

$$\frac{g''(0)}{M^2}\{\exp(-Mt)+Mt-1\} \leqslant g(t)-g(0)-g'(0)t \leqslant \frac{g''(0)}{M^2}\{\exp(Mt)+Mt-1\}.$$

Lemma 9. *For $t \geqslant 0$ we have $\exp(-t)+t-1 \geqslant \frac{1}{2}t^2-\frac{1}{6}t^3$.*

Proof of Lemma 9. For $t \geqslant 0$, consider the function $f(t)=\exp(-t)+t^3/6-t^2/2+t-1$. The statement is equivalent to $f(t) \geqslant 0$ for $t \geqslant 0$. It follows that $f(0)=0$, $f'(0)=0$, and $f''(t)=\exp(-t)+t-1 \geqslant 0$ so that f is convex. Therefore $f(t) \geqslant f(0)+tf'(0)=0$. ∎

Lemma 10 (Minoration Lemma)**.** *We have that*

$$\Lambda(\eta_0+\delta)-\Lambda(\eta_0)-\nabla\Lambda(\eta_0)'\delta \geqslant \left\{\tfrac{1}{3}\|\sqrt{w_i}\tilde{x}_i'\delta\|_{2,n}^2\right\} \wedge \left\{\frac{\bar{q}_A}{3}\|\sqrt{w_i}\tilde{x}_i'\delta\|_{2,n}\right\}$$

Proof. Step 1. (Minoration). Define the maximal radius over which the following criterion function can be bounded below by a suitable quadratic function

$$r_A = \sup_r \left\{ r \ : \ \begin{array}{c} \Lambda(\eta_0+\delta)-\Lambda(\eta_0)-\nabla\Lambda(\eta_0)'\delta \geqslant \frac{1}{3}\|\sqrt{w_i}\tilde{x}_i'\delta\|_{2,n}^2, \\ \text{for all } \delta \in A, \ \|\sqrt{w_i}\tilde{x}_i'\delta\|_{2,n} \leqslant r \end{array} \right\}.$$

Step 2 below shows that $r_A \geqslant \bar{q}_A$. By construction of r_A and the convexity of $\Lambda(\eta_0+\delta)-\Lambda(\eta_0)-\nabla\Lambda(\eta_0)'\delta$,

$$\begin{array}{l} \Lambda(\eta_0+\delta)-\Lambda(\eta_0)-\nabla\Lambda(\eta_0)'\delta \geqslant \\ \geqslant \frac{\|\sqrt{w_i}\tilde{x}_i'\delta\|_{2,n}^2}{3} \wedge \left\{ \frac{\|\sqrt{w_i}\tilde{x}_i'\delta\|_{2,n}}{r_A} \cdot \inf_{\tilde{\delta}\in A, \|\sqrt{w_i}\tilde{x}_i'\tilde{\delta}\|_{2,n}\geqslant r_A} \Lambda(\eta_0+\tilde{\delta})-\Lambda(\eta_0)-\nabla\Lambda(\eta_0)'\tilde{\delta} \right\} \\ \geqslant \frac{\|\sqrt{w_i}\tilde{x}_i'\delta\|_{2,n}^2}{3} \wedge \left\{ \frac{\|\sqrt{w_i}\tilde{x}_i'\delta\|_{2,n}}{r_A}\frac{r_A^2}{3} \right\} \geqslant \frac{\|\sqrt{w_i}\tilde{x}_i'\delta\|_{2,n}^2}{3} \wedge \left\{ \frac{\bar{q}_A}{3}\|\sqrt{w_i}\tilde{x}_i'\delta\|_{2,n} \right\}. \end{array}$$

Step 2. ($r_A \geqslant \bar{q}_A$) Defining $g_i(t)=\log\{1+\exp(\tilde{x}_i'\eta_0+t\tilde{x}_i'\delta)\}$ we have

$$\begin{array}{l} \Lambda(\eta_0+\delta)-\Lambda(\eta_0)-\nabla\Lambda(\eta_0)'\delta = \\ = \mathbb{E}_n\left[\log\{1+\exp(\tilde{x}_i'\{\eta_0+\delta\})\}-y_i\tilde{x}_i'(\eta_0+\delta)\right] \\ \quad -\mathbb{E}_n\left[\log\{1+\exp(\tilde{x}_i'\eta_0)-y_i\tilde{x}_i'\eta_0\}\right]-\mathbb{E}_n\left[(G_i-y_i)\tilde{x}_i'\delta\right] \\ = \mathbb{E}_n\left[\log\{1+\exp(\tilde{x}_i'\{\eta_0+\delta\})\}-\log\{1+\exp(\tilde{x}_i'\eta_0)\}-G_i\tilde{x}_i'\delta\right] \\ = \mathbb{E}_n[g_i(1)-g_i(0)-1\cdot g_i'(0)] \end{array}$$

Note that the function g_i is three times differentiable and satisfies, for $G_i(t) := \exp(\tilde{x}_i'\eta_0+t\tilde{x}_i'\delta)/\{1+\exp(\tilde{x}_i'\eta_0+t\tilde{x}_i'\delta)\}$,

$$g_i'(t)=(\tilde{x}_i'\delta)G_i(t), \quad g_i''(t)=(\tilde{x}_i'\delta)^2G_i(t)[1-G_i(t)], \quad g_i'''(t)=(\tilde{x}_i'\delta)^3G_i(t)[1-G_i(t)][1-2G_i(t)].$$

Thus $|g_i'''(t)| \leqslant |\tilde{x}_i'\delta| g_i''(t)$. Therefore, by Lemmas 8 and 9 we have

$$\begin{array}{rl} g_i(1)-g_i(0)-1\cdot g_i'(0) & \geqslant \frac{(\tilde{x}_i'\delta)^2 w_i}{(\tilde{x}_i'\delta)^2}\{\exp(-|\tilde{x}_i'\delta|)+|\tilde{x}_i'\delta|-1\} \\ & \geqslant w_i\left\{\frac{|\tilde{x}_i'\delta|^2}{2}-\frac{|\tilde{x}_i'\delta|^3}{6}\right\} \end{array}$$

Therefore we have

$$\Lambda(\eta_0+\delta)-\Lambda(\eta_0)-\nabla\Lambda(\eta_0)'\delta \quad \geqslant \tfrac{1}{2}\mathbb{E}_n\left[w_i|\tilde{x}_i'\delta|^2\right]-\tfrac{1}{6}\mathbb{E}_n\left[w_i|\tilde{x}_i'\delta|^3\right]$$

Note that for any $\delta \in A$ such that $\|\sqrt{w_i}\tilde{x}_i'\delta\|_{2,n} \leqslant \bar{q}_A$ we have

$$\|\tilde{x}_i'\delta\|_{2,n} \leqslant \bar{q}_A \leqslant \|\sqrt{w_i}\tilde{x}_i'\delta\|_{2,n}^3/\mathbb{E}_n\left[w_i|\tilde{x}_i'\delta|^3\right],$$

so that $\mathbb{E}_n[w_i|\tilde{x}_i'\delta|^3] \leqslant \mathbb{E}_n[w_i|\tilde{x}_i'\delta|^2]$. Therefore we have

$$\begin{array}{rl}\Lambda(\eta_0+\delta)-\Lambda(\eta_0)-\nabla\Lambda(\eta_0)'\delta & \geqslant \frac{1}{2}\mathbb{E}_n\left[w_i|\tilde{x}_i'\delta|^2\right]-\frac{1}{6}\mathbb{E}_n\left[w_i|\tilde{x}_i'\delta|^3\right]\\ & \geqslant \frac{1}{3}\mathbb{E}_n\left[w_i|\tilde{x}_i'\delta|^2\right]\end{array}$$

■

E.3. **Penalty Choice and Rate for ℓ_1-Penalized Logistic Regression.** Next we establish a simple (and known) bound for the choice of the penalty level λ within Lasso-Logistic under standard normalization. Refinements are possible under additional mild assumptions on the covariates.

Lemma 11 (Choice of Penalty, Hoeffding's Inequality)**.** *Assume that $\mathbb{E}_n[\tilde{x}_{ij}^2]=1$. Then, for any $\gamma \in (0,1)$ we have*

$$P\left(\|\nabla\Lambda(\eta_0)\|_\infty \leqslant \sqrt{2\log(2(p+1)/\gamma)/n}\right) \leqslant \gamma.$$

Proof. Let $G_i = \mathrm{E}[y_i \mid \tilde{x}_i] = \frac{\exp(\tilde{x}_i'\eta_0)}{1+\exp(\tilde{x}_i'\eta_0)}$, so that $\|\nabla\Lambda(\eta_0)\|_\infty = \|\mathbb{E}_n[(y_i-G_i)\,\tilde{x}_i]\|_\infty$. Then

$$P(\|\mathbb{E}_n[(y_i-G_i)\,\tilde{x}_i]\|_\infty \geqslant t) \leqslant (p+1)\max_{j\leqslant p} P(|\mathbb{E}_n[(y_i-G_i)\,\tilde{x}_{ij}]| \geqslant t) \leqslant 2(p+1)\exp(-t^2n/2).$$

■

Lemma 12 (Choice of Penalty, Self-Normalized Moderate Deviation Theory)**.** *Normalize the covariates so that $\mathbb{E}_n[\tilde{x}_{ij}^2]=1$, let $l_j=\sqrt{\mathbb{E}_n[w_i\tilde{x}_{ij}^2]}$, and $\widehat{l}_j=\sqrt{\mathbb{E}_n[\widehat{w}_i\tilde{x}_{ij}^2]}$. Assume that $K_{\tilde{x}}^2\log p \leqslant n\delta_n \min_j l_j^2$, $\Phi^{-1}(1-2p/\gamma) \leqslant \delta_n n^{1/3}$, and $\|\widehat{w}_i - w_i\|_{2,n}K_{\tilde{x}} \leqslant \delta_n \min_j l_j^2$. Then, setting $\widehat{\Gamma}=\mathrm{diag}(\widehat{l})$, for any $\gamma\in(0,1)$ and $\mu>0$, for n sufficiently large we have*

$$P\left(\|\widehat{\Gamma}^{-1}\nabla\Lambda(\eta_0)\|_\infty \leqslant \{1+\mu\}\Phi^{-1}(1-\gamma/[2p])/\sqrt{n}\right) \leqslant \gamma + o(1).$$

Proof. Let $\Gamma = \mathrm{diag}(l)$, $\tilde{l}_j = \sqrt{\mathbb{E}_n[(y_i-G_i)^2\tilde{x}_{ij}^2]}$, and $\widetilde{\Gamma}=\mathrm{diag}(\tilde{l})$. We have

$$\begin{array}{rl}\|\widehat{\Gamma}^{-1}\nabla\Lambda(\eta_0)\|_\infty & \leqslant \|\{\widehat{\Gamma}^{-1}-\Gamma^{-1}+\Gamma^{-1}-\widetilde{\Gamma}^{-1}\}\widetilde{\Gamma}\widetilde{\Gamma}^{-1}\nabla\Lambda(\eta_0)\|_\infty + \|\widetilde{\Gamma}^{-1}\nabla\Lambda(\eta_0)\|_\infty\\ & \leqslant \{\|\{\widehat{\Gamma}^{-1}-\Gamma^{-1}\}\widetilde{\Gamma}\|_\infty + \|\{\Gamma^{-1}-\widetilde{\Gamma}^{-1}\}\widetilde{\Gamma}\|_\infty\}\|\widetilde{\Gamma}^{-1}\nabla\Lambda(\eta_0)\|_\infty + \|\widetilde{\Gamma}^{-1}\nabla\Lambda(\eta_0)\|_\infty\\ & \leqslant \left\{\max_{j\leqslant p}\left|\frac{\tilde{l}_j}{l_j}\frac{l_j-\widehat{l}_j}{\widehat{l}_j}\right| + \max_{j\leqslant p}\left|\frac{\tilde{l}_j}{l_j}\frac{\tilde{l}_j-l_j}{\tilde{l}_j}\right| + 1\right\}\|\widetilde{\Gamma}^{-1}\nabla\Lambda(\eta_0)\|_\infty.\end{array}$$

Since w_i and $\widehat{w}_i$ are non-negative we have

$$\max_{j\leqslant p}|l_j-\widehat{l}_j| \quad \leqslant \max_{j\leqslant p}\sqrt{\mathbb{E}_n[|\widehat{w}_i-w_i|\tilde{x}_{ij}^2]} \leqslant \|\widehat{w}_i-w_i\|_{2,n}^{1/2}\max_{j\leqslant p}\{\mathbb{E}_n[\tilde{x}_{ij}^4]\}^{1/4}.$$

Also, since $\mathrm{E}[(y_i-G_i)^2 \mid \tilde{x}_i] = w_i$ and for positive number $|\sqrt{a}-\sqrt{b}| \leqslant \sqrt{|a-b|}$, we have

$$\begin{array}{rl}\max_{j\leqslant p}\left|\tilde{l}_j - l_j\right| & = \max_{j\leqslant p}\left|\sqrt{\mathbb{E}_n[(y_i-G_i)^2\tilde{x}_{ij}^2]}-\sqrt{\bar{\mathrm{E}}[w_i\tilde{x}_{ij}^2]}\right|\\ & \leqslant \sqrt{\max_{j\leqslant p}|(\mathbb{E}_n-\bar{\mathrm{E}})[(y_i-G_i)^2\tilde{x}_{ij}^2]|}\end{array}$$

By Lemma 4 we have

$$\max_{j\leqslant p} |(\mathbb{E}_n - \bar{\mathrm{E}})[(y_i - G_i)^2 \tilde{x}_{ij}^2]| \lesssim_P \sqrt{\frac{\log p}{n}} \max_{j\leqslant p} \{\mathbb{E}_n[\tilde{x}_{ij}^4]\}^{1/2}$$

Therefore for n large enough we have $\max_{j\leqslant p} \frac{|\widehat{l}_j - l_j|}{l_j} \vee \frac{|\tilde{l}_j - l_j|}{l_j} \leqslant \mu/16$ under the assumed growth conditions with probability $1 - o(1)$. In the same event we have

$$\|\widehat{\Gamma}^{-1}\nabla\Lambda(\eta_0)\|_\infty \leqslant \{1 + \mu/2\}\|\widetilde{\Gamma}^{-1}\nabla\Lambda(\eta_0)\|_\infty.$$

Finally, by self-normalized moderate deviation theory we have

$$P(\|\widetilde{\Gamma}^{-1}\nabla\Lambda(\eta_0)\|_\infty > t) \leqslant p \max_{j\leqslant p} P\left(\frac{\mathbb{E}_n[(y_i - G_i)\tilde{x}_{ij}]}{\sqrt{\mathbb{E}_n[(y_i - G_i)^2\tilde{x}_{ij}^2]}} > t \right) \leqslant 2p\Phi^{-1}(1 - \gamma/[2p])\{1 + O(\delta_n)\}$$

■

Comment E.1. Note that we can replace $(\widehat{w}_i)_{i=1}^n$ with $(\bar{w}_i)_{i=1}^n$ in Lemma 12 if $\bar{w}_i \geqslant w_i$ by construction. For instance $w_i \leqslant \bar{w}_i := 1/4$. Therefore it is valid to use $\lambda = \frac{c}{2}\sqrt{n}\Phi^{-1}(1 - \gamma/[2p])$ and $\widehat{l}_j = 1$ for $c > 1$.

Lemma 13. *Assume $\lambda/n \geqslant c\|\nabla\Lambda(\eta_0)\|_\infty$, $c > 1$ and let $\mathbf{c} = (c+1)/(c-1)$. Provided that $\bar{q}_{\Delta_{\mathbf{c}}} > 3(1 + \frac{1}{c})\lambda\sqrt{s}/(n\kappa_{\mathbf{c}})$*

$$\|\sqrt{w_i}\tilde{x}_i'(\widehat{\eta} - \eta_0)\|_{2,n} \leqslant 3(1 + \tfrac{1}{c})\frac{\lambda\sqrt{s}}{n\kappa_{\mathbf{c}}} \quad \textit{and} \quad \|\widehat{\eta} - \eta_0\|_1 \leqslant 3\frac{(1+c)(1+\mathbf{c})}{c}\frac{\lambda s}{n\kappa_{\mathbf{c}}^2}$$

Proof. Let $\delta = \widehat{\eta} - \eta_0$. By definition of $\widehat{\eta}$ in (E.51) we have $\Lambda(\widehat{\eta}) + \frac{\lambda}{n}\|\widehat{\eta}\|_1 \leqslant \Lambda(\eta_0) + \frac{\lambda}{n}\|\eta_0\|_1$. Thus,

$$\begin{array}{rl} \Lambda(\widehat{\eta}) - \Lambda(\eta_0) & \leqslant \frac{\lambda}{n}\|\eta_0\|_1 - \frac{\lambda}{n}\|\widehat{\eta}\|_1 \\ & \leqslant \frac{\lambda}{n}\|\delta_T\|_1 - \frac{\lambda}{n}\|\delta_{T^c}\|_1 \end{array}$$

However, by convexity of $\Lambda(\cdot)$ and Holder inequality we have

$$\begin{array}{rl} \Lambda(\widehat{\eta}) - \Lambda(\eta_0) & \geqslant -\|\nabla\Lambda(\eta_0)\|_\infty\|\delta\|_1 \\ & \geqslant -\frac{\lambda}{n}\frac{1}{c}\|\delta_T\|_1 - \frac{\lambda}{n}\frac{1}{c}\|\delta_{T^c}\|_1 \end{array}$$

Combining these relations we have $-\frac{\lambda}{n}\frac{1}{c}\|\delta_T\|_1 - \frac{\lambda}{n}\frac{1}{c}\|\delta_{T^c}\|_1 \leqslant \frac{\lambda}{n}\|\delta_T\|_1 - \frac{\lambda}{n}\|\delta_{T^c}\|_1$, which leads to $\|\delta_{T^c}\|_1 \leqslant \mathbf{c}\|\delta_T\|_1$.

By Lemma 10 with $A = \Delta_{\mathbf{c}}$ and the reasoning above we have

$$\begin{array}{rl} \frac{1}{3}\|\sqrt{w_i}\tilde{x}_i'\delta\|_{2,n}^2 \wedge \left\{\frac{\bar{q}_A}{3}\|\sqrt{w_i}\tilde{x}_i'\delta\|_{2,n}\right\} & \leqslant \Lambda(\widehat{\eta}) - \Lambda(\eta_0) - \nabla\Lambda(\eta_0)'\delta \\ & \leqslant \frac{\lambda}{n}\|\delta_T\|_1 - \frac{\lambda}{n}\|\delta_{T^c}\|_1 + \|\nabla\Lambda(\eta_0)\|_\infty\|\delta\|_1 \\ & \leqslant (1 + \frac{1}{c})\frac{\lambda}{n}\|\delta_T\|_1 \leqslant (1 + \frac{1}{c})\frac{\lambda\sqrt{s}}{n}\|\delta_T\| \\ & \leqslant (1 + \frac{1}{c})\frac{\lambda\sqrt{s}}{n}\|\sqrt{w_i}\tilde{x}_i'\delta\|_{2,n}/\kappa_{\mathbf{c}} \end{array}$$

Provided that $\bar{q}_A > 3(1 + \frac{1}{c})\lambda\sqrt{s}/(\kappa_{\mathbf{c}}n)$, so that the minimum on the LHS needs to be the quadratic term, we have

$$\|\sqrt{w_i}\tilde{x}_i'\delta\|_{2,n} \leqslant 3(1 + \tfrac{1}{c})\frac{\lambda\sqrt{s}}{n\kappa_{\mathbf{c}}}$$

■

E.4. **Sparsity of Lasso-Logistic.** We begin by establishing sparsity bounds which do not rely on large penalty choices nor on the irrepresentability condition[3] The (data-driven) sparsity is fundamental for the analysis of the rate of convergence of the Post-Lasso-Logistic estimator. The following lemma is useful.

Lemma 14. *The logistic link function satisfies* $|G(t+t_0)-G(t_0)| \leqslant G'(t_0)\{\exp(|t|)-1\}$. *If* $|t|\leqslant 1$ *we have* $\exp(|t|)-1 \leqslant 2|t|$.

Proof. Note that $|G''(s)| \leqslant G'(s)$ for all s. So that $-1 \leqslant \frac{d}{ds}\log(G'(s)) = \frac{G''(s)}{G'(s)} \leqslant 1$. Suppose $s \geqslant 0$. Therefore

$$-s \leqslant \log(G'(s+t_0)) - \log(G'(t_0)) \leqslant s.$$

In turn this implies $G'(t_0)\exp(-s) \leqslant G'(s+t_0) \leqslant G'(t_0)\exp(s)$. Integrating one more time from 0 to t,

$$G'(t_0)\{1-\exp(-t)\} \leqslant G(t+t_0)-G(t_0) \leqslant G'(t_0)\{\exp(t)-1\}.$$

The first result follows by noting that $1-\exp(-t) \leqslant \exp(t)-1$. The second follows by verification. ∎

Lemma 15 (Sparsity)**.** *Consider* $\widehat{\eta}$ *as defined in (E.51), let* $\widehat{s} = |\mathrm{support}(\widehat{\eta})|$ *and suppose* $\lambda/n \geqslant c\|\nabla\Lambda(\eta_0)\|_\infty$. *Then*

$$\widehat{s} \leqslant \frac{c^2(n/\lambda)^2}{(c-1)^2}\phi_{\max}(\widehat{s})\|\tilde{x}_i'(\widehat{\eta}-\eta_0)\|_{2,n}^2.$$

Provided that $\bar{q}_{\Delta_{\mathbf{c}}} > 3(1+\frac{1}{c})\lambda\sqrt{s}/(n\kappa_{\mathbf{c}})$ *we have*

$$\widehat{s} \leqslant s\cdot\phi_{\max}(\widehat{s})\frac{9\mathbf{c}^2}{\{\psi_{(r)}(\mathbf{c})\}^2\kappa_{\mathbf{c}}^2}.$$

Moreover, if $\frac{3(1+c)(1+\mathbf{c})}{c}\frac{\lambda s}{n\kappa_{\mathbf{c}}^2}\max_{i\leqslant n}\|\tilde{x}_i\|_\infty \leqslant 1$ *we have*

$$\sqrt{\widehat{s}} \leqslant 6\mathbf{c}\frac{\sqrt{\phi_{\max}(\widehat{s})}}{\kappa_{\mathbf{c}}}\sqrt{s} \quad \textit{and} \quad \widehat{s} \leqslant s\cdot 36\mathbf{c}^2\min_{m\in\mathcal{M}}\frac{\phi_{\max}(m)}{\kappa_{\mathbf{c}}^2}$$

where $\mathcal{M} = \{m\in\mathbf{N} : m > 72s\mathbf{c}^2\phi_{\max}(m)/\kappa_{\mathbf{c}}^2\}$

Proof. Let $\widehat{T} = \mathrm{support}(\widehat{\eta})$, $\widehat{s} = |\widehat{T}|$, $\delta = \widehat{\eta}-\eta_0$, and $\widehat{G}_i = \exp(\tilde{x}_i'\widehat{\eta})/\{1+\exp(\tilde{x}_i'\widehat{\eta})\}$. For any $j\in\widehat{T}$ we have $|\nabla_j\Lambda(\widehat{\eta})| = |\mathbb{E}_n[(y_i-\widehat{G}_i)\tilde{x}_{ij}]| = \lambda/n$.

The first relation follows from

$$\begin{aligned}
\tfrac{\lambda}{n}\sqrt{\widehat{s}} &= \|\mathbb{E}_n[(y_i-\widehat{G}_i)\tilde{x}_{i\widehat{T}}]\|_2 \\
&\leqslant \|\mathbb{E}_n[(y_i-G_i)\tilde{x}_{i\widehat{T}}]\|_2 + \|\mathbb{E}_n[(\widehat{G}_i-G_i)\tilde{x}_{i\widehat{T}}]\|_2 \\
&\leqslant \sqrt{\widehat{s}}\|\mathbb{E}_n[(y_i-G_i)\tilde{x}_{i\widehat{T}}]\|_\infty + \|\mathbb{E}_n[\tilde{x}_i'\delta\tilde{x}_{i\widehat{T}}]\|_2 \\
&\leqslant \tfrac{\lambda}{cn}\sqrt{\widehat{s}} + \sqrt{\phi_{\max}(\widehat{s})}\|\tilde{x}_i'\delta\|_{2,n}
\end{aligned}$$

The second follows from the first, the definition of $\psi_{(r)}(\mathbf{c})$, and Lemma 13 so that

$$\widehat{s} \leqslant \frac{c^2(n/\lambda)^2}{(c-1)^2}\phi_{\max}(\widehat{s})\|\tilde{x}_i'\delta\|_{2,n}^2 \leqslant \frac{c^2(n/\lambda)^2}{(c-1)^2}\phi_{\max}(\widehat{s})\frac{\|\sqrt{w_i}\tilde{x}_i'\delta\|_{2,n}^2}{\psi_{(r)}(\mathbf{c})^2} \leqslant s\cdot\phi_{\max}(\widehat{s})\frac{9\mathbf{c}^2}{\psi_{(r)}(\mathbf{c})^2\kappa_{\mathbf{c}}^2}$$

[3]The irrepresentability condition is the assumption that $\|\mathbb{E}_n[\tilde{x}_{iT^c}\tilde{x}_{iT}](\mathbb{E}_n[\tilde{x}_{iT}\tilde{x}_{iT}])^{-1}\mathrm{sign}(\eta_{0T})\|_\infty < 1$.

The third relation follows from

$$\begin{aligned}\tfrac{\lambda}{n}\sqrt{\widehat{s}} &= \|\mathbb{E}_n[(y_i-\widehat{G}_i)\tilde{x}_{i\widehat{T}}]\|_2\\ &\leqslant \|\mathbb{E}_n[(y_i-G_i)\tilde{x}_{i\widehat{T}}]\|_2 + \|\mathbb{E}_n[(\widehat{G}_i-G_i)\tilde{x}_{i\widehat{T}}]\|_2\\ &\leqslant \sqrt{\widehat{s}}\|\mathbb{E}_n[(y_i-G_i)\tilde{x}_{i\widehat{T}}]\|_\infty + \sup_{\|\theta\|_0\leqslant|\widehat{T}|,\|\theta\|=1}\mathbb{E}_n[|\widehat{G}_i-G_i|\cdot|\tilde{x}_i'\theta|]\\ &\leqslant \tfrac{\lambda}{cn}\sqrt{\widehat{s}} + 2\sqrt{\phi_{\max}(\widehat{s})}\|\sqrt{w_i}\tilde{x}_i'\delta\|_{2,n}\end{aligned}$$

where we used Lemma 14 so that $|\widehat{G}_i-G_i|\leqslant w_i 2|\tilde{x}_i'\delta|$ since by Lemma 13 $\|\delta\|_1\leqslant 3\frac{(1+c)(1+\mathbf{c})}{c}\frac{\lambda s}{n\kappa_{\mathbf{c}}^2}$ so that $\max_{i\leqslant n}\|\tilde{x}_i\|_\infty\|\delta\|_1\leqslant 1$ by the assumed condition.

Therefore, by the $\|\cdot\|_{2,n}$ bound in Lemma 13 we have

$$(1-\tfrac{1}{c})\tfrac{\lambda}{n}\sqrt{\widehat{s}} \quad \leqslant 6\sqrt{\phi_{\max}(\widehat{s})}\tfrac{(1+c)}{c}\tfrac{\lambda\sqrt{s}}{n\kappa_{\mathbf{c}}}$$

which implies $\sqrt{\widehat{s}}\leqslant 6\mathbf{c}\frac{\sqrt{\phi_{\max}(\widehat{s})}}{\kappa_{\mathbf{c}}}\sqrt{s}$.

The last relation follows by the previous result and the fact that sparse eigenvalues are sublinear functions. ∎

E.5. **Post model selection Logistic regression rate.**

Lemma 16. *Consider $\widetilde{\eta}$ as defined in (E.52). Let $\widehat{s}^*:=|\widehat{T}^*|$. We have*

$$\|\sqrt{w_i}\tilde{x}_i'(\widetilde{\eta}-\eta_0)\|_{2,n}\leqslant\sqrt{3}\sqrt{\Lambda(\tilde{\eta})-\Lambda(\eta_0)}+3\sqrt{\widehat{s}^*+s}\|\nabla\Lambda(\eta_0)\|_\infty/\sqrt{\phi_{\min}(\widehat{s}^*+s)}$$

provided that $\bar{q}_A/6>\sqrt{\widehat{s}^*+s}\|\nabla\Lambda(\eta_0)\|_\infty/\sqrt{\phi_{\min}(\widehat{s}^*+s)}$ and $q_A/6>\sqrt{\Lambda(\tilde{\eta})-\Lambda(\eta_0)}$ for $A=\{\delta\in\mathbb{R}^p:\|\delta\|_0\leqslant\widehat{s}^*+s\}$.

Proof. Let $\tilde{\delta}=\tilde{\eta}-\eta_0$ and $\tilde{t}_{2,n}=\|\sqrt{w_i}\tilde{x}_i'\tilde{\delta}\|_{2,n}$. By Lemma 10 with $A=\{\delta\in\mathbb{R}^p:\|\delta\|_0\leqslant\widehat{s}^*+s\}$, we have

$$\begin{aligned}\tfrac{1}{3}\tilde{t}_{2,n}^2\wedge\left\{\tfrac{\bar{q}_A}{3}\tilde{t}_{2,n}\right\} &\leqslant \Lambda(\tilde{\eta})-\Lambda(\eta_0)-\nabla\Lambda(\eta_0)'\tilde{\delta}\\ &\leqslant \Lambda(\tilde{\eta})-\Lambda(\eta_0)+\|\nabla\Lambda(\eta_0)\|_\infty\|\tilde{\delta}\|_1\\ &\leqslant \Lambda(\tilde{\eta})-\Lambda(\eta_0)+\tilde{t}_{2,n}\sqrt{\widehat{s}^*+s}\|\nabla\Lambda(\eta_0)\|_\infty/\sqrt{\phi_{\min}(\widehat{s}^*+s)}\end{aligned}$$

Provided that $\bar{q}_A/6>\sqrt{\widehat{s}^*+s}\|\nabla\Lambda(\eta_0)\|_\infty/\sqrt{\phi_{\min}(\widehat{s}^*+s)}$ and $\bar{q}_A/6>\sqrt{\Lambda(\tilde{\eta})-\Lambda(\eta_0)}$, if the minimum on the LHS is the linear term, we have $\tilde{t}_{2,n}\leqslant\sqrt{\Lambda(\tilde{\eta})-\Lambda(\eta_0)}$ which implies the result. Otherwise, since for positive numbers $a^2\leqslant b+ac$ implies $a\leqslant\sqrt{b}+c$, we have

$$\tilde{t}_{2,n}\leqslant\sqrt{3}\sqrt{\Lambda(\tilde{\eta})-\Lambda(\eta_0)}+3\sqrt{\widehat{s}^*+s}\|\nabla\Lambda(\eta_0)\|_\infty/\sqrt{\phi_{\min}(\widehat{s}^*+s)}.$$

∎

Appendix F. Additional Monte Carlo

F.1. **Monte Carlo for Approximately Sparse Models.** In this section we provide further simulations to illustrate the performance of the proposed methods. In particular we illustrate the performance of the method when applied to approximately sparse models. We consider a similar design to the one used in Section 4 of the main text, namely

$$\mathrm{E}[y\mid d,x]=G(d\alpha_0+x'\{c_y\nu_y\}),\qquad d=x'\{c_d\nu_d\}+v.$$

However, the vectors ν_y and ν_d are set to

$$\nu_{yj} = 1/j^2, \nu_{dj} = 1/j^2, \tag{F.53}$$

so they are approximately sparse. Again we let $x = (1, z')'$ consists of an intercept and covariates $z \sim N(0, \Theta)$, and the error v is i.i.d. as $N(0, 1)$. The dimension p of the covariates x is 250, and the sample size n is 200. The regressors are correlated with $\Theta_{ij} = \rho^{|i-j|}$ and $\rho = 0.5$. As before the coefficient c_d is used to control the R^2 of the reduce form equation, c_y is set similarly and in every repetition, we draw new errors v_i's and controls x_i's. The figures display the results over 100 different designs where $\alpha_0 = 0.5$ and the values of c_y and c_d are set to achieve $R^2 = \{0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9\}$ for each equation. There were 1000 replications for each of the 100 designs.

Figure 6 reveals that the performance of the method for this approximately sparse design is very similar to the performance obtained with the sparse designs considered in Section 4. Again the double selection estimator arise as a more reliable estimator.

References

[1] Francis Bach. Self-concordant analysis for logistic regression. *Eletronic Journal of Statistics*, 4:384–414, 2010.

[2] A. Belloni, D. Chen, V. Chernozhukov, and C. Hansen. Sparse models and methods for optimal instruments with an application to eminent domain. *Econometrica*, November 2012.

[3] A. Belloni and V. Chernozhukov. ℓ_1-penalized quantile regression for high dimensional sparse models. *Annals of Statistics*, 39(1):82–130, 2011.

[4] A. Belloni, V. Chernozhukov, and C. Hansen. Lasso methods for gaussian instrumental variables models. arXiv:[math.ST], http://arxiv.org/abs/1012.1297, 2010.

[5] A. Belloni, V. Chernozhukov, and K. Kato. Robust inference in high-dimensional approximately sparse quantile regression models. *working paper*, 2013.

[6] A. Belloni, V. Chernozhukov, and K. Kato. Uniform post model selection inference for LAD regression models. *ArXiv*, 2013.

[7] Alexandre Belloni, Victor Chernozhukov, and Christian Hansen. Inference on treatment effects after selection amongst high-dimensional controls. *ArXiv*, 2011.

[8] Alexandre Belloni, Victor Chernozhukov, and Christian Hansen. Inference methods for high-dimensional sparse econometric models. *Advances in Economics and Econometrics, 10th World Congress of Econometric Society*, Volume III, Econometrics, Edited by Daron Acemoglu, Manuel Arellano and Eddie Dekel:245–295, 2013.

[9] P. J. Bickel, Y. Ritov, and A. B. Tsybakov. Simultaneous analysis of Lasso and Dantzig selector. *Annals of Statistics*, 37(4):1705–1732, 2009.

[10] F. Bunea. Honest variable selection in linear and logistic regressionmodels via ℓ_1 and $\ell_1 + \ell_2$ penalization. *Electronic Journal of Statistics*, 2:11531194, 2008.

[11] Michael R. Kosorok. *Introduction to Empirical Processes and Semiparametric Inference*. Series in Statistics. Springer, Berlin, 2008.

[12] M. Kwemou. Non-asymptotic oracle inequalities for the lasso and group lasso in high dimensional logistic model. *arXiv preprint*, (arXiv:1206.0710), 2012.

[13] M. Ledoux and M. Talagrand. *Probability in Banach Spaces (Isoperimetry and processes)*. Ergebnisse der Mathematik undihrer Grenzgebiete, Springer-Verlag, 1991.

[14] Hannes Leeb and Benedikt M. Pötscher. Model selection and inference: facts and fiction. *Economic Theory*, 21:21–59, 2005.

[15] Hannes Leeb and Benedikt M. Pötscher. Can one estimate the conditional distribution of post-model-selection estimator? *The Annals of Statistics*, 34(5):2554–2591, 2006.

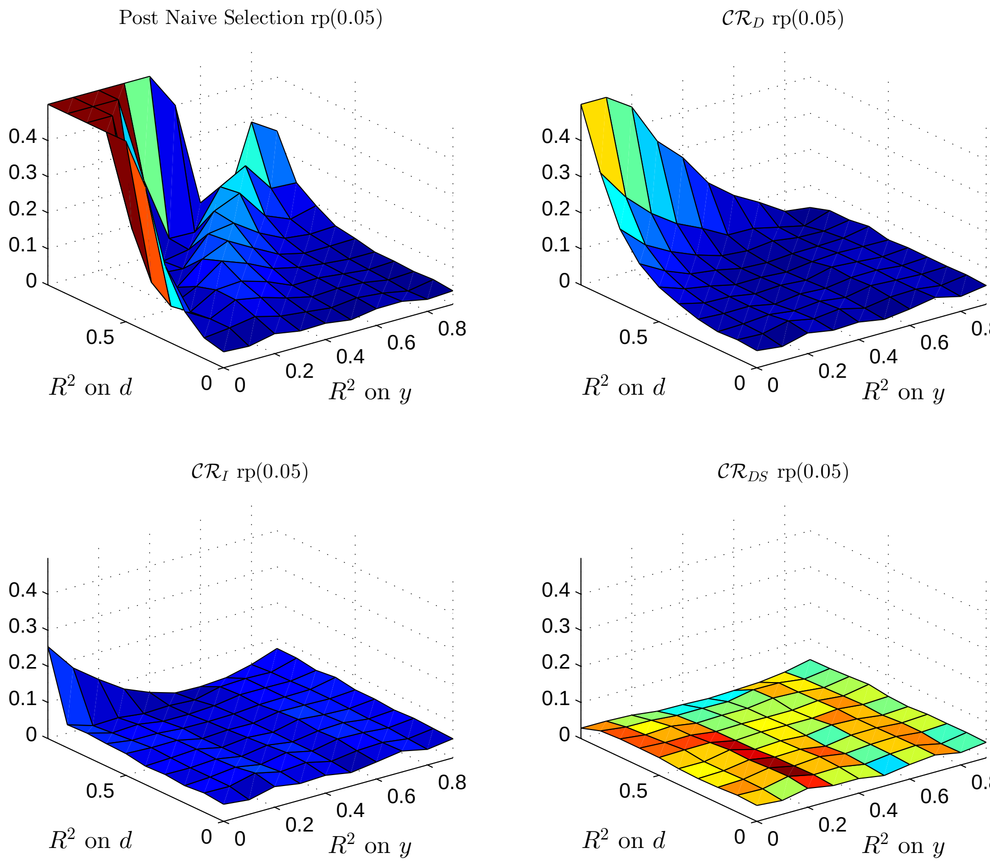


FIGURE 6. For the approximately sparse model defined by (F.53), the figures display the rp(0.05) of the naive post-model selection estimator and the proposed confidence regions based on optimal instrument ($\mathcal{CR}_D$ and $\mathcal{CR}_I$) and double selection ($\mathcal{CR}_{DS}$). There are a total of 100 different designs with $\alpha_0 = 0.5$. The results are based on 1000 replications for each design.

[16] Hannes Leeb and Benedikt M. Pötscher. Sparse estimators and the oracle property, or the return of Hodges' estimator. *J. Econometrics*, 142(1):201–211, 2008.

[17] L. Meier, V. Van der Geer, and P. Bühlmann. The group lasso for logistic regression. *J. R. Statist. Soc.: Series B (Statist. Methodol.)*, 70(1):5371, 2008.

[18] Sahand N. Negahban, Pradeep Ravikumar, Martin J. Wainwright, and Bin Yu. A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. *Statistical Science*, 27(4):538–557, 2012.

[19] J. Neyman. Optimal asymptotic tests of composite statistical hypotheses. In U. Grenander, editor, *Probability and Statistics, the Harold Cramer Volume*. New York: John Wiley and Sons, Inc., 1959.

[20] J. Neyman. $c(\alpha)$ tests and their use. *Sankhya*, 41:1–21, 1979.

[21] Yaniv Plan and Roman Vershynin. Robust 1-bit compressed sensing and sparse logistic regression: A convex programming approach. *IEEE Transactions on Information Theory*, 59(1):482 – 494, Jan. 2013.

[22] Benedikt M. Pötscher. Confidence sets based on sparse estimators are necessarily large. *Sankhyā*, 71(1, Ser. A):1–18, 2009.

[23] Benedikt M. Pötscher and Hannes Leeb. On the distribution of penalized maximum likelihood estimators: the LASSO, SCAD, and thresholding. *J. Multivariate Anal.*, 100(9):2065–2082, 2009.

[24] P. Ravikumar, M. Wainwright, and J. Lafferty. High-dimensional ising model selection using -regularized logistic regression. *Ann. Statist.*, 38(2):12871319, 2010.

[25] M. Rudelson and S. Zhou. Reconstruction from anisotropic random measurements. *ArXiv:1106.1151*, 2011.

[26] Mark Rudelson and Roman Vershynin. On sparse reconstruction from fourier and gaussian measurements. *Communications on Pure and Applied Mathematics*, 61:10251045, 2008.

[27] R. Tibshirani. Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B*, 58:267–288, 1996.

[28] S. A. van de Geer. High-dimensional generalized linear models and the lasso. *Annals of Statistics*, 36(2):614–645, 2008.

[29] Sara Anna van de Geer, Peter Bühlmann, and Ya'acov Ritov. On asymptotically optimal confidence regions and tests for high-dimensional models. *ArXiv*, 2013.

[30] A. W. van der Vaart and J. A. Wellner. *Weak Convergence and Empirical Processes*. Springer Series in Statistics, 1996.

[31] Aad W. van der Vaart and Jon A. Wellner. Empirical process indexed by estimated functions. *IMS Lecture Notes-Monograph Series*, 55:234–252, 2007.

[32] Lie Wang. L_1 penalized lad estimator for high dimensional ilnear regression. *ArXiv*, 2012.

[33] Cun-Hui Zhang and Stephanie S. Zhang. Confidence intervals for low-dimensional parameters with high-dimensional data. *ArXiv.org*, (arXiv:1110.2563v1), 2011.