Journal of Business & Economic Statistics

Taylor & Francis
Taylor & Francis Group

Journal of Business & Economic Statistics

ISSN: 0735-0015 (Print) 1537-2707 (Online) Journal homepage: www.tandfonline.com/journals/ubes20

Post-Selection Inference for Generalized Linear Models With Many Controls

Alexandre Belloni, Victor Chernozhukov & Ying Wei

To cite this article: Alexandre Belloni, Victor Chernozhukov & Ying Wei (2016) Post-Selection Inference for Generalized Linear Models With Many Controls, Journal of Business & Economic Statistics, 34:4, 606-619, DOI: 10.1080/07350015.2016.1166116

To link to this article: https://doi.org/10.1080/07350015.2016.1166116

Supplementary materials for this article are available online. Please go to http://tandfonline.com/r/JBES

Post-Selection Inference for Generalized Linear Models With Many Controls

Alexandre Belloni
The Fuqua School of Business, Duke University, Durham, NC 27708 (abn5@duke.edu)

Victor Chernozhukov
Dept. of Economics, Massachusetts Institute of Technology, Cambridge, MA 02139 (vchern@mit.edu)

Ying Wei
Department of Biostatistics, Columbia University, New York, NY 10032
(yw2148@columbia.edu)

This article considers generalized linear models in the presence of many controls. We lay out a general methodology to estimate an effect of interest based on the construction of an instrument that immunizes against model selection mistakes and apply it to the case of logistic binary choice model. More specifically 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, potentially exceeds the sample size n using sparsity assumptions. The sparsity assumption means that there is a subset of $s < n$ controls, which suffices to accurately approximate the nuisance part of the regression function. Importantly, the estimators and these resulting confidence regions are valid uniformly over s-sparse models satisfying $s^2 \log^2 p = o(n)$ and other technical conditions. 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. Under suitable conditions, the estimators are semi-parametrically efficient in the sense of attaining the semi-parametric efficiency bounds for the class of models in this article.

KEY WORDS: Double selection; Instruments; Model selection; Neymanization; Optimality; Sparsity; Uniformly valid inference.

1. INTRODUCTION

The literature on high-dimensional generalized linear models has experienced rapid development (van de Geer 2008; Negahban et al. 2012). As in the case of linear mean regression models, a striking result of this literature is to achieve consistency when the total number of covariates p is 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 like Lasso (Tibshirani 1996; Bickel, Ritov, and Tsybakov 2009). 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 Bunea (2008), Bach (2010), and Kwemou (2012). Group logistic regressions were studied in Meier, der Geer, and Bühlmann (2008) and Kwemou (2012) to exploit addition sparsity patterns. Ising models were considered in Ravikumar, Wainwright, and Lafferty (2010) and connections with robust 1-bit recovery were derived in Plan and Vershynin (2013). These works also derive rates of convergence for the coefficients, prediction error consistency, and variable selection consistency under various conditions.

This article attacks the problem of estimation and inference on a regression coefficient of interest in generalized linear models allowing for the total number of covariates p to be much larger than the sample size n. 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. We discuss the model framework for generalized linear models and then provide estimators to the logistic regression case to illustrate the results.

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 nonzero coefficients are sufficiently separated from zero. Although the

Journal of Business & Economic Statistics
October 2016, Vol. 34, No. 4
DOI: 10.1080/07350015.2016.1166116
Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/r/jbes.

separation condition is commonly used and might be appealing in some technometric applications (e.g., signal processing), it is often unrealistic and not credible in econometric, biometric, and many other applications. Even if applicable, it might not lead to accurate approximations of the 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 that inevitably occur in many applications (mistakes are very likely to occur when some coefficients are at the range of $O(n^{-1/2})$, which is typically not distinguishable from zero).

Our work contributes to a growing literature that avoids imposing separation conditions. In the context of instrumental regression, Belloni, Chernozhukov, and Hansen (2010) and Belloni et al. (2012) provided uniformly valid estimation and inference methods for instrumental variable models, using either post-selection or ℓ_1-regularization methods to estimate "optimal instruments." They provided a $\sqrt{n}$-consistent, semi-parametrically efficient estimator of the main low-dimensional structural parameter. In the context of the linear mean regression model, Belloni, Chernozhukov, and Hansen (2013, 2014) proposed a "double selection" approach to constructing uniformly valid estimation and inference methods, and Zhang and Zhang (2014) used one-step corrections to ℓ_1-regularized estimators. In either case a $\sqrt{n}$-consistent, semi-parametrically efficient estimator of the low-dimensional regression parameter of interest is provided. In the case of linear quantile regression models, Belloni, Chernozhukov, and Kato (2013, 2015) provided uniformly $\sqrt{n}$-consistent estimators and uniformly valid inference methods for least absolute deviations and quantile regressions. In an independent and contemporaneous work, van de Geer et al. (2014) proposed 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 article is an independent proposal, and relies instead on either optimal instrument strategy or the double selection strategy, which is related to Neyman's approach to dealing with nuisance parameters.

The aforementioned works as well as the current approach deviate substantially from the traditional approach of performing inference based upon perfect model selection results. Leeb and Pötscher (2005), Leeb and Pötscher (2008), Pötscher (2009), and Pötscher and Leeb (2009) have shown that such traditional/naive inference approach is not robust to violations of the separation condition, which bounds the magnitude of the nonzero 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 when the separation condition is violated. We shall confirm the failure of such naive post-selection procedures in 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 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 Belloni, Chernozhukov, and Hansen (2014) and Belloni, Chernozhukov, and Kato (2013, 2015).

We construct our estimators and confidence regions via three steps. The first step uses post-model selection methods to estimate the nuisance part of the regression—the part of the regression function associated to controls (i.e., nonmain 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 lead to our proposed estimators and confidence regions. The framework allows for different methods to be used on each step leading to different estimators for generalized linear models. For the case of logit link function, we propose one estimator based upon instrumental logistic regression with optimal instrument and another estimator 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. 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 (e.g., 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). One of the main results is to establish $\sqrt{n}$-consistency and asymptotic normality of estimators for generalized linear modes under high-level conditions on nuisance parameters.

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. As mentioned earlier, 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. 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. In the estimation of optimal instruments, which we use as an ingredient in building the optimal estimating equation to create immunization property, we rely on post-selection least squares estimator with data-dependent weights. The presence of data-dependent weights 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.

We organize the remainder of the article as follows. In Section 2, we present the framework for generalized linear models and the proposed estimators specialized to the logistic link function case. In Section 3 we provide the statements of our main results on the uniform validity of the estimators and confidence regions. We present primitive conditions for the logistic case and high-level conditions for generalized linear models. Section 4 contains a Monte Carlo experiment. We present the proofs of these results in online Appendix A. In online Appendix B we collect results on Lasso and Post-Lasso with estimated weights (Appendix B.1) as well as results on ℓ_1-penalized Logistic regression and post-model selection Logistic regression (Appendix B.2). In the online Appendix C 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 $1 \leqslant i \leqslant n$, that is, 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 nonzero 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 support $(\delta) = \{j \in \{1, \ldots, p\} : \delta_j \neq 0\}$. 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. GENERIC SETUP AND METHOD

Consider a generalized linear regression model, where the outcome of interest y_i relates to a scalar main regressor d_i (e.g., a treatment or a policy variable) and p-dimensional controls x_i through a link function G, namely for $i = 1, \ldots, n$

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

Here α_0 is the main target parameter, and $x_i'\beta_0$ is the nuisance regression function. The vector β_0 is a high-dimensional parameter that is assumed to be sparse, namely $\|\beta_0\|_0 \leqslant s$. We require s to be small relative to n in the sense that will be specified below, in particular

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

is required. In many settings this condition allows for the estimation of the nuisance function at the rate of $o(n^{-1/4})$.

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 as those are unavoidable if coefficients are near zero. Our proposed methods rely (implicitly or explicitly) on an instrument $z_{0i} = z_0(d_i, x_i)$ such that:

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

$$\frac{\partial}{\partial\alpha}\mathrm{E}[\{y_i - G(d_i\alpha + x_i'\beta_0)\}z_{0i}]\bigg|_{\alpha=\alpha_0} \neq 0, \tag{2.3}$$

$$\frac{\partial}{\partial\beta}\mathrm{E}[\{y_i - G(d_i\alpha_0 + x_i'\beta)\}z_{0i}]\bigg|_{\beta=\beta_0} = 0. \tag{2.4}$$

The first and second relations provide an estimating equation for α_0. Relation (2.4) is key in our analysis and states that the estimating equation (2.2) is insensitive with respect to first-order perturbations of the nuisance function $x_i'\beta_0$. We call this orthogonality condition "immunity." Such immunization ideas can be traced to Neyman's approach to dealing with nuisance parameters, as we discuss in Section 5.2. (Note also that because of (2.1), there is also immunity with respect to perturbations on z_0.)

Our methods proceed in three 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:

$$f_id_i = f_ix_i'\theta_0 + v_i, \quad \text{with } \mathrm{E}\,[f_iv_ix_i] = 0, \tag{2.5}$$

where

$$f_i := w_i/\sigma_i, \quad w_i := G'(d_i\alpha_0 + x_i'\beta_0), \quad \sigma_i^2 := \mathrm{var}(y_i|d_i, x_i), \tag{2.6}$$

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

$$z_{0i} := v_i/\sigma_i. \tag{2.7}$$

Here too we shall impose a sparsity condition in (2.5), 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 article extending to this case; see Remark 2.2.

Table 1. Estimators and confidence regions based on optimal instrument

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

$$(\widehat{\alpha}, \widehat{\beta}) \in \arg\min_{\alpha,\beta} \ \mathbb{E}_n[\Lambda_i(\alpha, \beta)] + \tfrac{\lambda_1}{n}\|(\alpha, \beta)\|_1$$
$$(\widetilde{\alpha}, \widetilde{\beta}) \in \arg\min_{\alpha,\beta} \ \mathbb{E}_n[\Lambda_i(\alpha, \beta)] \ : \ \text{support}(\beta) \subseteq \text{support}(\widehat{\beta})$$

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

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

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

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

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

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

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

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

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

NOTE: The algorithm has three steps: (1) initial estimation of the regression function via post-selection logistic regression, (2) estimation of instruments that are orthogonal to the weighted controls via a weighted post-selection least squares, and (3) estimation of α_0 based on the nuisance estimates obtained in Steps 1 and 2. Without loss of generality 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 Appendix B; see (B.5). 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 $\widehat{\Sigma}_{2n}^2 = \mathbb{E}_n[\widehat{v}_i^2]$.

The weights $f_i = w_i/\sigma_i$'s are used to achieve the orthogonality condition (2.4):

$$\frac{\partial}{\partial\beta}\text{E}[\{y_i - G(d_i\alpha_0 + x_i'\beta)\}z_{0i}]\bigg|_{\beta=\beta_0} = \text{E}[w_i z_{0i} x_i] = \text{E}\,[f_i v_i x_i] = 0; \qquad (2.8)$$

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 Equation (2.2). The orthogonality condition (2.4) is therefore a critical ingredient in achieving asymptotic uniform validity of the coverage of confidence regions. Among all instruments that provide such immunization, the instrument given in (2.7) minimizes the asymptotic variance of the asymptotically normal and $\sqrt{n}$-consistent estimator based on the estimating Equation (2.2). Other valid (but sub-optimal) choices of instruments are discussed in Remark 2.1. We will establish results for generalized linear models under high-level conditions.

2.1 Logistic Case and Specific Estimators

Next we apply the above principle to the case of logistic regression and propose specific implementations of estimators. In this case the link function G is given by the logistic link function

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

and the following simplification occurs: w_i in (2.6) equals the conditional variance of the outcome σ_i^2, namely

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

and $\quad f_i = \sqrt{w_i},$

so that the decomposition (2.5) and the optimal instrument (2.7) become

$$\sqrt{w_i}d_i = \sqrt{w_i}x_i'\theta_0 + v_i, \quad \text{E}\left[\sqrt{w_i}v_i x_i\right] = 0$$

and $\quad z_{0i} = v_i/\sigma_i = d_i - x_i'\theta_0. \qquad (2.9)$

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

$$\Lambda(\alpha, \beta) = \mathbb{E}_n[\Lambda_i(\alpha, \beta)] = \mathbb{E}_n[\log\{1 + \exp(d_i\alpha + x_i'\beta)\} - y_i(d_i\alpha + x_i'\beta)]. \qquad (2.10)$$

Table 1 displays an estimator 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 estimated weights constructed based on Step 1. Note that Step 2 is used to construct the optimal instrument. Step 3 uses 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 the first two steps instead of penalized estimators was motivated by a better finite sample performance in our experiments. We

Table 2. Estimators and confidence region based on double selection

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

$$(\widehat{\alpha}, \widehat{\beta}) \in \arg\min_{\alpha,\beta} \ \mathbb{E}_n[\Lambda_i(\alpha, \beta)] + \tfrac{\lambda_1}{n}\|(\alpha, \beta)\|_1$$
$$(\widetilde{\alpha}, \widetilde{\beta}) \in \arg\min_{\alpha,\beta} \ \mathbb{E}_n[\Lambda_i(\alpha, \beta)] \ : \ \text{support}(\beta) \subseteq \text{support}(\widehat{\beta})$$

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

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

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 Steps 1 and 2:

$$(\check{\alpha}, \check{\beta}) \in \arg\min_{\alpha,\beta} \ \mathbb{E}_n[\Lambda_i(\alpha, \beta)\widehat{f}_i/\widehat{\sigma}_i] \ : \ \text{support}(\beta) \subseteq \text{support}(\widehat{\beta}) \cup \text{support}(\widehat{\theta})$$

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

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

NOTE: 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. Without loss of generality 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 Appendix B, see (B.5). 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 $\check{x}_i = x_{i,\text{support}(\check{\beta})}$.

also provide two confidence regions for α_0 in Table 1. 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 a second estimator, which builds upon the idea of the double selection method proposed in Belloni, Chernozhukov, and Hansen (2014) for partial linear mean regression models. The method replaces Step 3 in Table 1 with a (weighted) logistic regression of the outcome on the main regressor as well as the union of controls selected in two selection steps—Steps 1 and 2. (Note that the algorithm is stated for any generalized linear model in which case Step 3 is a weighted regression where the weights are given by $\widehat{f}_i/\widehat{\sigma}_i$, which equals to 1 in the case of a logistic link function.) 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.

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

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

and is nontrivial, namely $\bar{\mathrm{E}}[w_i d_i z_{0i}] \neq 0$. A valid, nontrivial 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 $z_{0i} := v_i/\sigma_i$. Estimation of this instrument requires that in Step 2 a Lasso method is applied in the weighted Equation (2.5). 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 online Appendix B. Thus, estimation of the optimal instruments poses an interesting problem in its own right. 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, 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.

Remark 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 they perform better in an extensive set of Monte Carlo experiments.

Remark 2.3 (Data-Driven Choice of Penalty Parameters). The penalty parameters λ_1 and λ_2 as defined in the algorithms above are theoretically valid and are motivated by self-normalized moderate deviation theory. Other data-driven choices are possible but their theoretical validity is outside the scope of this article. For example, cross-validation typically underpenalize to reduce bias to obtain better estimates but tends to select a substantial larger number of variables. This suggests cross-validation to be more suitable for the algorithm based on optimal instrument than for the algorithm based on double selection. Another approach suggested in sec. 4.2 of Chernozhukov, Chetverikov, and Kato (2013) relies on new Gaussian approximation results and can be implemented via a multiplier bootstrap procedure.

3. MAIN THEORETICAL RESULTS

3.1 Logistic Regression Under Primitive Assumptions

In this section, we list and discuss primitive conditions that allow us to derive our results in the case of logistic regression. These conditions ensure good properties of ℓ_1-penalized methods and the associated post-selection estimators. Fix some sequences of constants, $\delta_n \to 0$, and $\Delta_n \to 0$, and constants $0 < c < C < \infty$.

Condition L. (i) Let $\{(y_i, d_i, x_i) : i =, \dots, n\}$ be independent random vectors that obey the models given by (2.1) and (2.5) with G being the logistic function. There exists $s = s_n$ such that $\|\beta_0\|_0 + \|\theta_0\|_0 \leqslant s$, $\|\beta_0\| + \|\theta_0\| \leqslant C$. (ii) The following moment conditions hold $\bar{\mathrm{E}}[\{(d, x')\xi\}^4] \leqslant C\|\xi\|^4$, $\bar{\mathrm{E}}[w_i\{(d, x')\xi\}^2] \geqslant c\|\xi\|^2$. We have that $\min_{j\leqslant p} \bar{\mathrm{E}}[w_i x_{ij}^2 v_i^2] \geqslant c > 0$ and $\max_{j\leqslant p} \bar{\mathrm{E}}[|\sqrt{w_i} x_{ij} v_i|^3]^{1/3} \log^{1/2}(p \vee n) \leqslant \delta_n n^{1/6}$. Furthermore, the conditional variance σ_i^2 satisfy $\min_{i\leqslant n} \sigma_i^2 \geqslant c > 0$ with probability $1 - \Delta_n$. (iii) For $K_q = \mathrm{E}[\max_{i\leqslant n} \|(d_i, z_{0i}, x')\|_\infty^q]^{1/q}$, we have $K_1^2 s^2 \log^2(p \vee n) \leqslant \delta_n n$ and $K_4^4 s \log(p \vee n) \log^3 n \leqslant \delta_n n$.

Condition L(i) assumes independence across i and the model described in Section 2 and sparsity conditions, which makes estimation possible even if $p > n$. Condition L(ii) assumes the conditional variance is bounded away from zero and imposes mild moment conditions. Condition L(iii) imposes growth requirements on the triple (s, p, n) as n grows. An important consequence of Condition L is to imply that submatrices of the design matrix are well behaved even though the design matrix cannot have rank p if $p > n$; see Rudelson and Vershynin (2008); Rudelson and Zhou (2011) for detailed discussion. This ensures that ℓ_1-penalized estimators are well behaved with suitable choices of penalty parameters under the stated sparsity assumptions.

Next we state the main inferential results of the article. 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 3.1 (Robust Estimation and Inference Based on the Optimal IV Estimator). Consider any triangular array of data $(y_i, d_i, x_i)_{i=1}^n$ that obeys Condition L for all $n \geqslant 1$. Then, the estimator $\check{\alpha}$ based on the optimal instrument, as defined in Table 1, 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, that is, $\widehat{\Sigma}_{1n}^2/\Sigma_n^2 = 1 + o_{\mathrm{P}}(1)$ and $\widehat{\Sigma}_{2n}^2/\Sigma_n^2 = 1 + o_{\mathrm{P}}(1)$.

Theorem 3.1 establishes that the IV estimator $\check{\alpha}$ is $\sqrt{n}$-consistent and asymptotically normal. Under suitable conditions the large-sample variance coinciding with the semi-parametric efficiency bound for the partially linear logistic regression model (see Section 5.3 for an additional discussion). 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 in Steps 1 and 2.

The following result derives similar properties for the double selection estimator described in Table 2.

Theorem 3.2 (Robust Estimation and Inference Based on Double Selection). Consider any triangular array of data $(y_i, d_i, x_i)_{i=1}^n$ that obeys Condition L for all $n \geqslant 1$. Then, the double selection estimator $\check{\alpha}$ as defined in Table 2 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, 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, that is, $\widehat{\Sigma}_{1n}^2/\Sigma_n^2 = 1 + o_{\mathrm{P}}(1)$ and $\widehat{\Sigma}_{2n}^2/\Sigma_n^2 = 1 + o_{\mathrm{P}}(1)$, where $\check{w}_i = G(d_i\check{\alpha} + x_i'\check{\beta})\{1 - G(d_i\check{\alpha} + x_i'\check{\beta})\}$ and $\check{x}_i = x_{i,\mathrm{support}(\check{\beta})}$.

Theorems 3.1 and 3.2 allow for the data-generating process to change with n, in particular allowing sequences of regression models, with coefficients never perfectly distinguishable from zero, that is, models where perfect model selection is not possible. In turn, the results achieved in Theorems 3.1 and 3.2 are uniformly valid over a large class of sparse models. In what follows, we formalize 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 sparse models where the stated above sparsity conditions, moment conditions, and growth conditions hold. This collection expressly permits models to have near zero coefficients, and thus does not impose the separation conditions. For $Q_n \in \mathcal{Q}_n$, let the notation P_{Q_n} mean that under P_{Q_n}, $\{(y_i, d_i, x_i')'\}_{i=1}^n$ is distributed according to Q_n.

Corollary 3.1 (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 the estimator $\check{\alpha}$, based either on optimal instrument or double selection, is $\sqrt{n}$-consistent and asymptotically normal uniformly over $\mathcal{Q}_n$, namely

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

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 3.2 (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, namely

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

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. These results motivate interesting questions on the construction of confidence regions for many parameters of interest that are simultaneously valid. We refer to Belloni et al. (2013, 2015) and Belloni, Chernozhukov, and Kato (2015) where simultaneous confidence regions are proposed in a variety of settings.

Remark 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 Belloni, Chernozhukov, and Hansen (2013, 2014). For example, if the model satisfies

$$\mathrm{E}[y_i \mid d_i, x_i] = G(\alpha_0 d_i + x_i'\beta_0 + r_{yi}), \tag{3.11}$$

$$f_i d_i = f_i x_i'\theta_0 + r_{di} + v_i, \quad \mathrm{E}[f_i v_i x_i] = 0, \tag{3.12}$$

where $\|\beta_0\|_0 \leqslant s$, $\|\theta_0\|_0 \leqslant s$, and the approximation errors r_{yi} and r_{di} are such that

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

and
$$|\bar{\mathrm{E}}[f_i v_i r_{yi}]| \leqslant \delta_n n^{-1/2}. \tag{3.13}$$

We can show that the results in Theorems 3.1 and 3.2 and Corollaries 3.1 and 3.2 continue to hold for this approximately sparse model. This means that the results are robust with respect to moderate violations of the sparsity assumption.

3.2 Generalized Linear Models Under High-Level Assumptions

In this section we establish $\sqrt{n}$-consistency and asymptotic normality for an estimator $\check{\alpha}$ of α_0 associated with a generalized linear model based on high-level conditions. These high-level conditions cover a variety of different estimators including the estimators described in Tables 1 and 2. In what follows note that the estimated instrument $\widehat{z}_i = \widehat{z}_i(d_i, x_i)$ and the expectations below are evaluated at the given estimates.

Condition IR. (i) The data $\{y_i, d_i, x_i\}$ independent across $i = 1, \ldots, n$, satisfies (2.1), $\sigma_i^2 = \mathrm{Var}(y_i \mid d_i, x_i)$, $w_i = G'(d_i\alpha_0 + x_i'\beta_0)$, and the link function G is such that $\sup_{t\in\mathbb{R}} |G(t)| \leqslant C$, $\sup_{t\in\mathbb{R}} |G'(t)| \leqslant C$, and $\sup_{t\in\mathbb{R}} |G''(t)| \leqslant C$. (ii) The following moment conditions hold $\mathrm{E}[w_i z_{0i} x_i] = 0$, $|\bar{\mathrm{E}}[w_i d_i z_{0i}]| \geqslant c > 0$, $\bar{\mathrm{E}}[\sigma_i^2 z_{0i}^2] \geqslant c > 0$, $\bar{\mathrm{E}}[z_{0i}^2 d_i^2] \leqslant C$, $\bar{\mathrm{E}}[\sigma_i^3 z_{0i}^3] \leqslant C$, and $\bar{\mathrm{E}}[(x_i'\xi)^4] \leqslant C$ for all $\|\xi\| = 1$. (iii) For some sequences $\delta_n \to 0$ and $\Delta_n \to 0$ with probability at least $1 - \Delta_n$, the estimates $(\check{\alpha}, \widehat{\beta}, \widehat{z})$ satisfy

$$\|\widehat{\beta}-\beta_0\| \leqslant \delta_n n^{-1/4}, \; \bar{\mathrm{E}}[(\tilde{z}_i - z_{0i})^2]\big|_{\tilde{z}=\widehat{z}} \leqslant \delta_n^2,$$

$$\|\widehat{\beta}-\beta_0\| \cdot \{\bar{\mathrm{E}}[(\tilde{z}_i - z_{0i})^2]\big|_{\tilde{z}=\widehat{z}}\}^{1/2} \leqslant \delta_n n^{-1/2}, \tag{3.14}$$

$$\sup_{\alpha:|\alpha-\alpha_0|\leqslant\delta_n} \Big|(\mathbb{E}_n - \bar{\mathrm{E}})\big[\{y_i - G(d_i\alpha + x_i'\widehat{\beta})\}\widehat{z}_i - \{y_i - G(d_i\alpha + x_i'\beta_0)\}z_{0i}\big]\Big| \leqslant \delta_n\, n^{-1/2} \tag{3.15}$$

$$|\check{\alpha}-\alpha_0| \leqslant \delta_n, \quad \big|\,\mathbb{E}_n[\{y_i - G(d_i\check{\alpha} + x_i'\widehat{\beta})\}\widehat{z}_i]\,\big| \leqslant \delta_n\, n^{-1/2}. \tag{3.16}$$

(iv) With probability $1 - \Delta_n$ we have $|\widehat{w}_i| \leqslant C$, $\|\widehat{w}_i - w_i\|_{2,n} \leqslant \delta_n$, $\|d_i(\widehat{w}_i - w_i)\|_{2,n} \leqslant \delta_n$, $\|d_i\widehat{z}_i\|_{2,n} \leqslant C$, $\|\widehat{z}_i - z_{0i}\|_{2,n} \leqslant \delta_n$, and $\|z_{0i}x_i'(\widehat{\beta}-\beta_0)\|_{2,n} \leqslant \delta_n$.

This set of high-level conditions allows us to cover several generalized models of interest. In particular Condition L and the choices of post-selection methods described in the previous section imply Condition IR. Next we formally state our main result for generalized linear models.

Theorem 3.3. Under Condition IR(i,ii,iii) we have

$$\{\bar{\mathrm{E}}[\sigma_i^2 z_{0i}^2]\}^{-1/2}\bar{\mathrm{E}}[w_i d_i z_{0i}]\sqrt{n}(\check{\alpha}-\alpha_0) = \frac{\{\bar{\mathrm{E}}[\sigma_i^2 z_{0i}^2]\}^{-1/2}}{\sqrt{n}}\sum_{i=1}^n\{y_i - G(d_i\alpha_0 + x_i'\beta_0)\}z_{0i} + o_{\mathrm{P}}(1)$$

and

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

Moreover, if Condition IR(iv) also holds, we have

$$nL_n(\alpha_0) \rightsquigarrow \chi^2(1)$$

and the variance estimator is consistent, namely

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

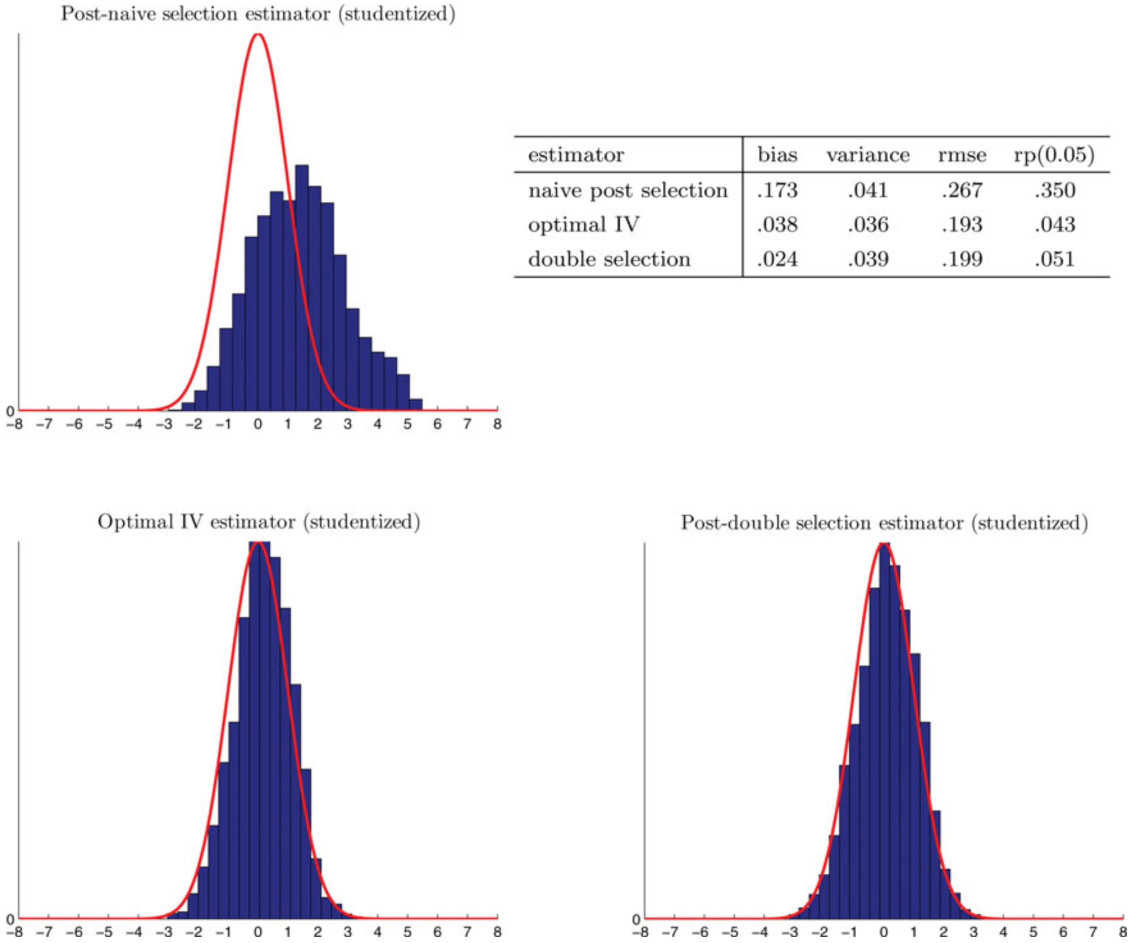
It is important to note that Theorem 3.3 is applicable to various estimation methods and we believe it will be of interest even in settings not based on sparsity assumptions.

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 naive post-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:

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



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

Figure 1. The top right panel display bias, variance, RMSE, and rejection frequency for a 0.05-level test. The plots display the distribution of the naive post-selection estimator (top left panel) and the two proposed estimators: optimal instruments (bottom left panel) and double selection (bottom right).

where the coefficient vectors ν_y and ν_d are set to

$$\nu_y = (1, 1/2, 1/3, 1/4, 1/5, 0, 0, 0, 0, 0, 1, 1/2, 1/3, 1/4, 1/5, 0, 0, \ldots, 0)',$$

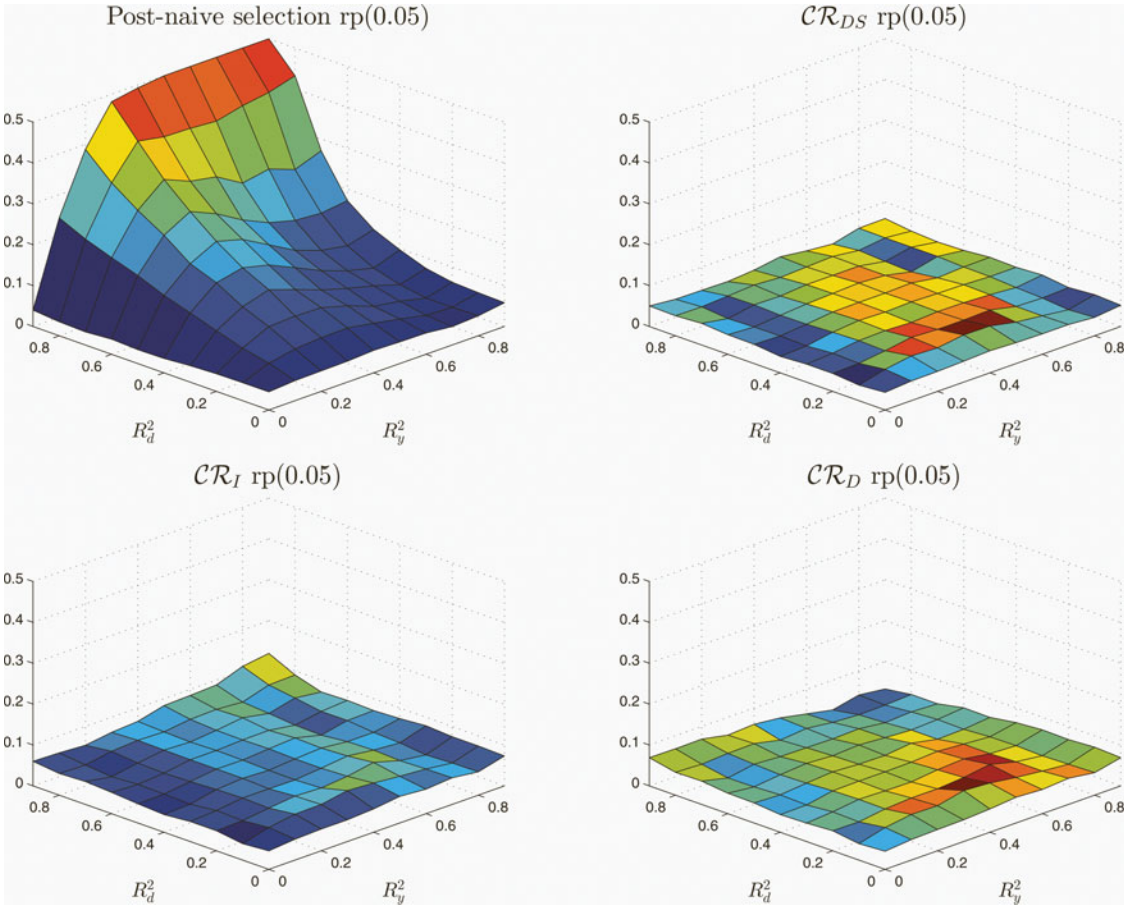
$$\nu_d = (1, 1/2, 1/3, 1/4, 1/5, 1/6, 1/7, 1/8, 1/9, 1/10, 0, 0, \ldots, 0)',$$

$x = (1, z')'$ consists of an intercept and covariates $z \sim N(0, \Theta)$, and the error $\tilde{v}$ is iid as $N(0, 1)$. The dimension p of the covariates x is 250, and the sample size n is 200. In this setting the coefficients feature a declining pattern, with the smallest nonzero coefficients being hard to differentiate from zero for 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 online supplementary materials where we also consider an approximately sparse model for which all 250 coefficients are nonzero. 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). For every replication, 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 nonzero 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. Figure 1 summarizes the performance and displays the distribution of the following estimators, which are centered by the true value α_0 and Studentized by their standard deviation:

1. Naive post-selection estimator—estimator of α_0 based on logistic regression after the naive selection using ℓ_1-penalized logistic regression;
2. Optimal instrument estimator—estimator of α_0 based on the instrumental logistic regression with the optimal instrument, as defined in Table 1;



Post-naive selection rp(0.05)

$\mathcal{CR}_{DS}$ rp(0.05)

$\mathcal{CR}_{I}$ rp(0.05)

$\mathcal{CR}_{D}$ rp(0.05)

Figure 2. The plots display the rejection frequencies at 0.05 level (rp(0.05)) of the confidence regions based on naive post-selection, 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.

3. Double selection estimator—estimator of α_0 based on the logistic regression after double selection, as defined in Table 2.

The optimal IV estimator and the 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 root mean squared errors, and confidence regions have rejection rates close to the nominal level of 0.05. 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 naive post-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 Leeb and Pötscher (2005, 2006); Leeb and Pötscher (2008) in the context of linear models.

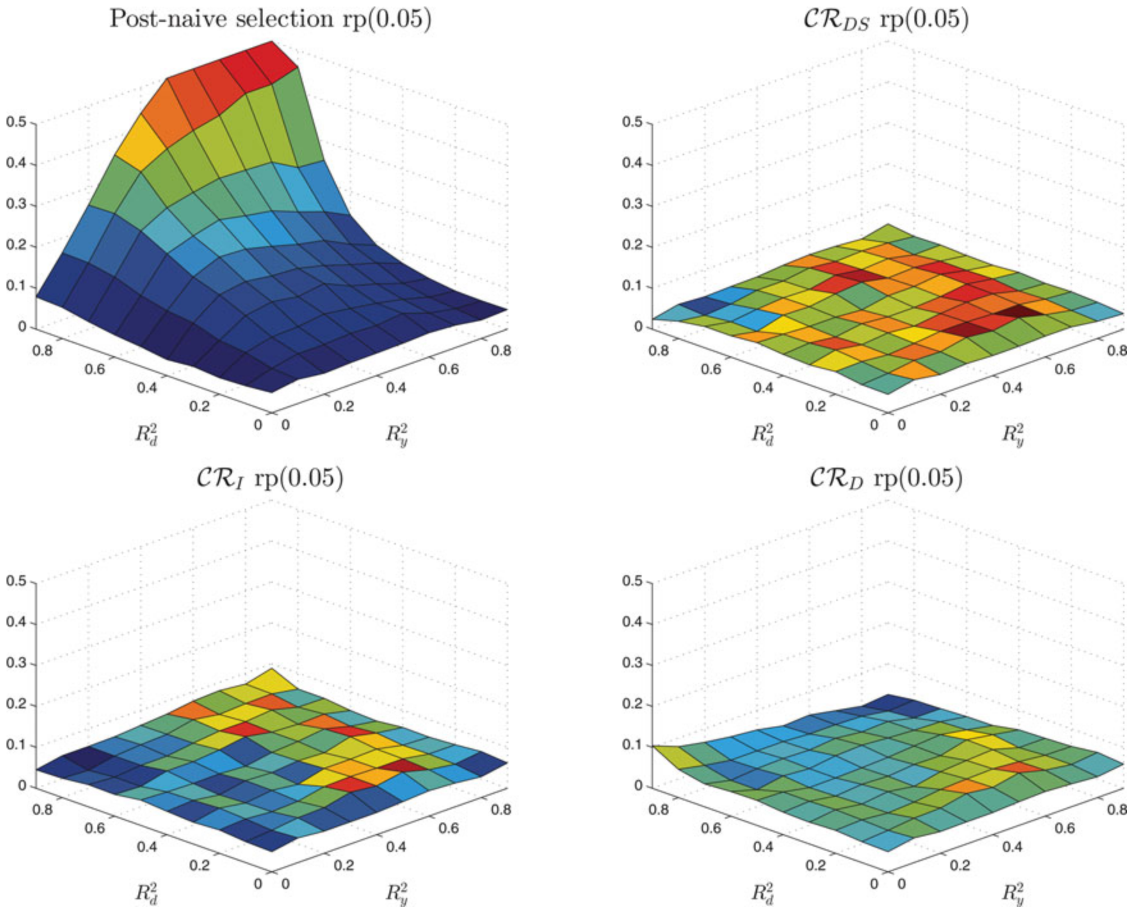
We now examine the performance more systematically by varying

$$(R_d^2, R_y^d) \in \{0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9\}^2$$
$$\text{and} \qquad \alpha_0 \in \{0, 0.25, 0.5\}. \qquad (4.17)$$

This gives us 300 different dgps. For each dgp we performed 1000 Monte Carlo simulations. Figures 2–4 display the rejection frequencies of confidence regions with (nominal) significance level of 0.05 and Figure 5 displays 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 3.1 and 3.2 are, and also confirm that the previous conclusions continue to hold across a wide set of dgp.

In Figures 2–4 we consider the rejection (noncoverage) frequencies of confidence regions based on: naive post-selection logistic estimator[1], optimal IV ($\mathcal{CR}_D$ and $\mathcal{CR}_I$), and double selection ($\mathcal{CR}_{DS}$). These figures illustrate the uniformity properties of the confidence regions based on the discussed estimators. The ideal figure would be a flat surface with the rejection

[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 naive post-selection logistic estimator.



Post-naive selection rp(0.05)

$\mathcal{CR}_{DS}$ rp(0.05)

$\mathcal{CR}_{I}$ rp(0.05)

$\mathcal{CR}_{D}$ rp(0.05)

Figure 3. The plots display the rejection frequencies at 0.05 level (rp(0.05)) of the confidence regions based on naive post-selection, 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.

frequency of the true value equal to the nominal level of 0.05. The confidence regions based on the naive post selection perform very poorly, and deviate strongly away from the ideal level of 0.05 throughout large parts of the model space (induced by (4.17)). 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 double selection estimator seems to outperform the estimator based on the explicit construction of the optimal instrument (e.g., the rejection rates and the RMSE for the case with $\alpha = 0.5$, where optimal instrument procedure tends to perform noticeably worse). Thus, based on the theoretical results and on the Monte Carlo results, we recommend the use of the double selection estimator over the optimal IV estimator and, certainly, over the naive post-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 (2.5) 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 $\widehat{T}^* = \text{support}(\widehat{\beta}) \cup \text{support}(\widehat{\theta})$ denote the variables selected in Step 1 and 2. By the first-order conditions of the double selection logistic regression of Step 3 in Table 2 we have

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

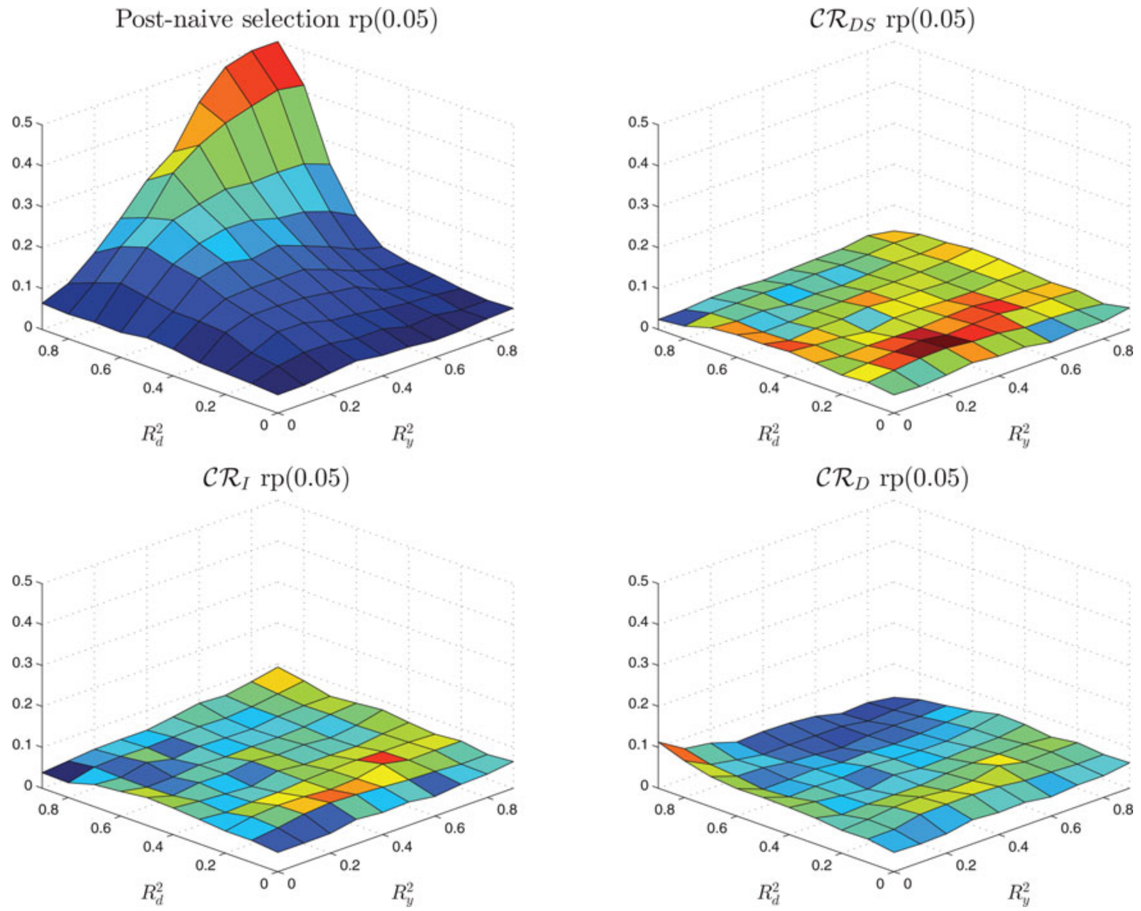
which creates an orthogonal relation to any linear combination of $(d_i,\ x_{i\widehat{T}^*}')'$. In particular, by taking the linear combination $(d_i,\ x_{i\widehat{T}^*}')(1, -\widetilde{\theta}')' = d_i - x_i'\widetilde{\theta} = \widehat{z}_i$, we have

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

Therefore, the 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 of the optimal instrument estimator that was *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 leading to a more robust performance.


Post-naive selection rp(0.05)

$\mathcal{CR}_{DS}$ rp(0.05)

$\mathcal{CR}_I$ rp(0.05)

$\mathcal{CR}_D$ rp(0.05)

Figure 4. The plots display the rejection frequencies at 0.05 level (rp(0.05)) of the confidence regions based on naive post-selection, 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.

5.2 Relation to Neyman's $C(\alpha)$ Test

Next we discuss connections between the proposed approach and Neyman's $C(\alpha)$ test (Neyman 1959, 1979) (here we draw on the discussion in Belloni, Chernozhukov, and Kato 2015). For the sake of exposition, we assume the instruments are known and iid observations. As stated in (2.2) and (2.4) we rely on instruments satisfying the two equations:

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

and

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

These conditions allows us to 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. Although there will be many instruments z_{0i} that can achieve the property stated above, the choice $z_{0i} = v_i/\sigma_i$ proposed in Section 2 is optimal as it minimizes the asymptotic variance of the resulting estimators.

Generally, valid (but not necessarily optimal) instruments can be constructed by modifying the weighted Equation (2.5) to

$$f_i d_i = f_i m_0(x_i) + \tilde{v}_i, \quad \mathrm{E}[f_i \tilde{v}_i | x_i] = 0, \tag{5.18}$$

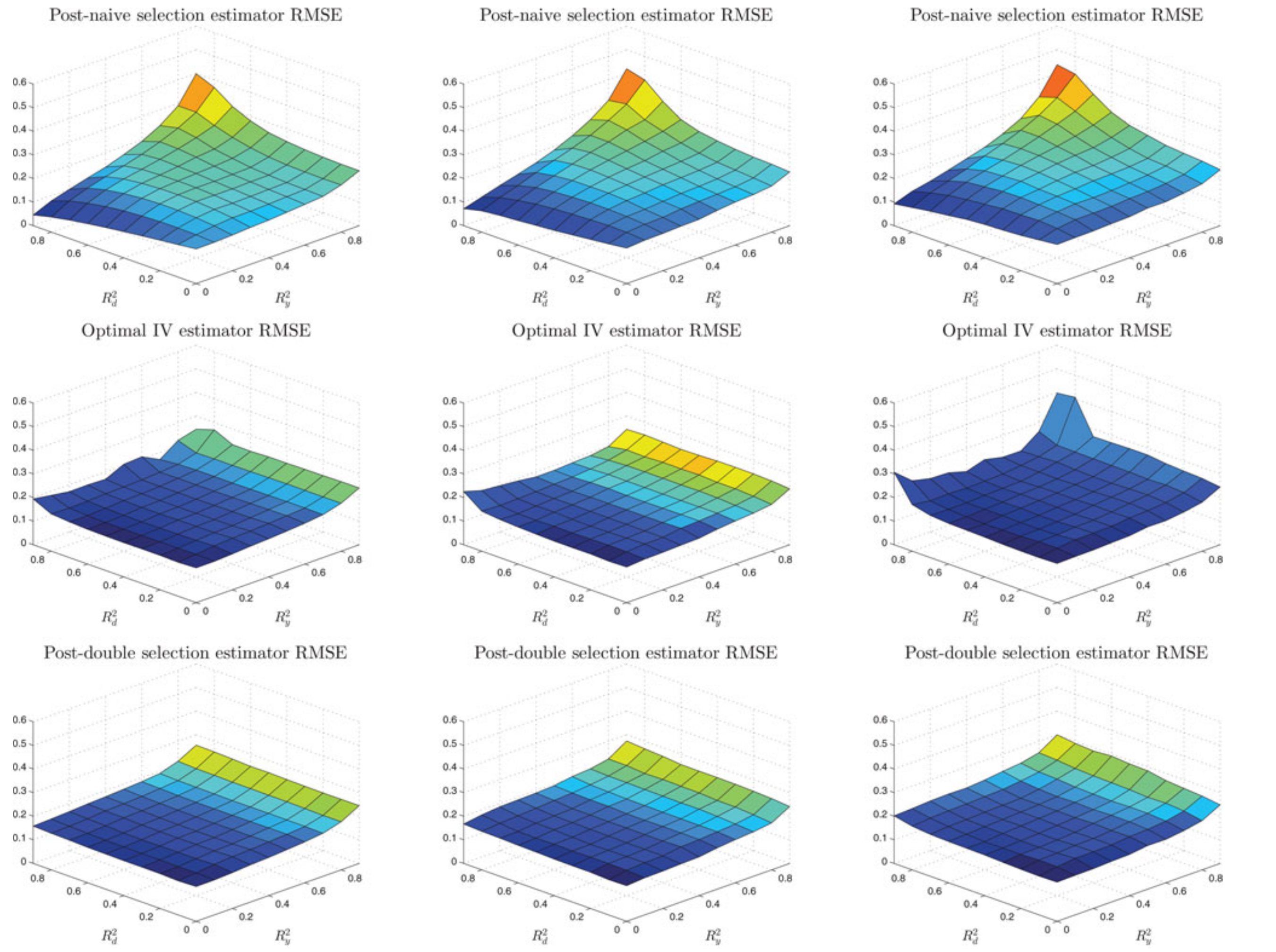
where $f_i = f(d_i, z_i)$ is a nonnegative weight, and setting the instrument as $z_{0i} := f_i\tilde{v}_i/w_i$. Because of the zero-mean condition in (5.18), and provided that $m_0 \in \mathcal{H}$, the function $m_0(x_i)$ in (5.18) is the solution of the following weighted least squares problem

$$\min_{h\in\mathcal{H}} \mathrm{E}\left[f_i^2\{d_i - h(x_i)\}^2\right], \tag{5.19}$$

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

$$f_i d_i = f_i x_i'\theta_0 + \tilde{v}_i, \quad \mathrm{E}[f_i \tilde{v}_i x_i] = 0. \tag{5.20}$$

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, such properties could motivate the use of different estimators (e.g., we can use ridge estimators if the m_0 is "dense" with respect to x).


Post-naive selection estimator RMSE

Post-naive selection estimator RMSE

Post-naive selection estimator RMSE

Optimal IV estimator RMSE

Optimal IV estimator RMSE

Optimal IV estimator RMSE

Post-double selection estimator RMSE

Post-double selection estimator RMSE

Post-double selection estimator RMSE

Figure 5. The plots display the RMSE of the naive post-selection estimator, optimal instrument estimator, and double selection estimator. 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.

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

$$\mathbb{E}_n[\{y_i - G(d_i\alpha + x_i'\widehat{\beta})\}z_{0i}] - \mathbb{E}_n[\{y_i - G(d_i\alpha + x_i'\beta_0)\}z_{0i}] = o_{\mathrm{P}}(n^{-1/2}), \quad (5.21)$$

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 $s^2\log^2(p \vee n)/n \to 0$ to guarantee that $\|\widehat{\beta} - \beta\| = o_{\mathrm{P}}(n^{-1/4})$. Equation (5.21) 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 (5.20) we have that $\theta_0 = \mathrm{E}[f_i^2 x_i x_i']^- \mathrm{E}[f_i^2 d_i x_i]$, where A^- denotes a generalized inverse of A. Letting $\widehat{\varepsilon}_i(\alpha) = y_i - G(d_i\alpha + x_i'\widehat{\beta})$ and

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

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

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

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

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

It is not difficult to check that using $f_i = w_i/\sigma_i$ leads to the smallest possible value $\mathrm{E}[v_i^2]^{-1}$ of $\tilde{\Sigma}_n^2$. Therefore, $z_{0i} = v_i/\sigma_i$ 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 for Logistic Model

Next we consider a connection to the (local) minimax efficiency analysis from the semiparametric literature, where we follow the discussion in Belloni, Chernozhukov, and Kato (2015). Our model is a special case of a partially linear logistic model, and Kosorok (2008) derived 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 (5.18) 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 (5.18). 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 Kosorok 2008), and inference about α_0 based on this estimator provides best minimax power against local alternatives (see theorem 18.12 in Kosorok 2008).

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 3.1). 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

$$\mathrm{E}[y_i \mid d_i, x_i] = G(d_i\{\alpha_0 + t_1\} + x_i'\{\beta_0 + t_2\theta_0\}), \quad \|t\| \leqslant \delta, \tag{5.22}$$

$$\sqrt{w_i}d_i = \sqrt{w_i}x_i'\theta_0 + v_i, \quad \mathrm{E}[\sqrt{w_i}v_i|x_i] = 0, \tag{5.23}$$

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 Kosorok (2008), 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).

SUPPLEMENTARY MATERIALS

In the supplementary material, we provide detailed proofs of the main results in Appendix A. Appendix B collects results on Lasso and Post-Lasso with estimated weights as well as results on ℓ_1-penalized Logistic regression and post-model selection Logistic regression. In the online Appendix C we present auxiliary inequalities. Appendix D contains additional proofs while Appendix E contains additional Monte Carlo simulations for approximately sparse models.

[Received December 2014. Revised February 2016.]

REFERENCES

Bach, F. (2010), "Self-Concordant Analysis for Logistic Regression," *Electronic Journal of Statistics*, 4, 384–414. [606]

Belloni, A., Chen, D., Chernozhukov, V., and Hansen, C. (2012), "Sparse Models and Methods for Optimal Instruments With an Application to Eminent Domain," *Econometrica*, 80, 2369–2429. [607]

Belloni, A., Chernozhukov, V., Chetverikov, D., and Wei, Y. (2015), "Uniformly Valid Post-Regularization Confidence Regions for Many Functional Parameters in Z-Estimation Framework," *arXiv:1512.07619*. [612]

Belloni, A., Chernozhukov, V., Fernández-Val, I., and Hansen, C. (2013), "Program Evaluation With High-Dimensional Data," *arXiv preprint arXiv:1311.2645; Econometrica (forthcoming)*. [612]

Belloni, A., Chernozhukov, V., and Hansen, C. (2010), "LASSO Methods for Gaussian Instrumental Variables Models," ArXiv:[math.ST]. Available at *http://arxiv.org/abs/1012.1297*. [607]

——— (2013), "Inference Methods for High-Dimensional Sparse Econometric Models," in *Advances in Economics and Econometrics: 10th World Congress of Econometric Society* (Vol. III, Econometrics), eds. Daron Acemoglu, Manuel Arellano, and Eddie Dekel, New York: Cambridge University Press [607,612]

——— (2014), "Inference on Treatment Effects After Selection Among High-Dimensional Controls," *The Review of Economic Studies*, 81, 608–650. [607,610,612]

Belloni, A., Chernozhukov, V., and Kato, K. (2013), "Robust Inference in High-Dimensional Approximately Sparse Quantile Regression Models," *arXiv:1312.7186v1*. [607]

——— (2015), "Uniform Post Selection Inference for LAD Regression Models and Other Z-Estimators," *Biometrika*, 102, 77–94. [607,612,616,618]

Bickel, P. J., Ritov, Y., and Tsybakov, A. B. (2009), "Simultaneous Analysis of Lasso and Dantzig Selector," *Annals of Statistics*, 37, 1705–1732. [606]

Bunea, F. (2008), "Honest Variable Selection in Linear and Logistic Regression Models via ℓ_1 and $\ell_1 + \ell_2$ Penalization," *Electronic Journal of Statistics*, 2, 1153–1194. [606]

Chernozhukov, V., Chetverikov, D., and Kato, K. (2013), "Gaussian Approximations and Multiplier Bootstrap for Maxima of Sums of High-Dimensional Random Vectors," *The Annals of Statistics*, 41, 2786–2819. [611]

Kosorok, M. R. (2008), *Introduction to Empirical Processes and Semiparametric Inference*, Series in Statistics, Berlin: Springer. [618]

Kwemou, M. (2012), "Non-Asymptotic Oracle Inequalities for the Lasso and Group Lasso in High Dimensional Logistic Model," *arXiv preprint, arXiv:1206.0710*. [606]

Leeb, H., and Pötscher, B. M. (2005), "Model Selection and Inference: Facts and Fiction," *Economic Theory*, 21, 21–59. [607,614]

——— (2006), "Can One Estimate the Conditional Distribution of Post-Model-Selection Estimator?" *The Annals of Statistics*, 34, 2554–2591. [614]

——— (2008), "Sparse Estimators and the Oracle Property, or the Return of Hodges' Estimator," *Journal of Econometrics*, 142, 201–211. [607,614]

Meier, L., der Geer, V. V., and Bühlmann, P. (2008), "The Group Lasso for Logistic Regression," *Journal of the Royal Statistical Society,* Series B, 70, 53–71. [606]

Negahban, S. N., Ravikumar, P., Wainwright, M. J., and Yu, B. (2012), "A Unified Framework for High-Dimensional Analysis of M-Estimators With Decomposable Regularizers," *Statistical Science*, 27, 538–557. [606]

Neyman, J. (1959), "Optimal Asymptotic Tests of Composite Statistical Hypotheses," in *Probability and Statistics: the Harold Cramer Volume*, ed. U. Grenander, New York: Wiley. [616]

——— (1979), "$C(\alpha)$ Tests and Their Use," *Sankhya*, 41, 1–21. [616]

Plan, Y., and Vershynin, R. (2013), "Robust 1-Bit Compressed Sensing and Sparse Logistic Regression: A Convex Programming Approach," *IEEE Transactions on Information Theory*, 59, 482–494. [606]

Pötscher, B. M. (2009), "Confidence Sets Based on Sparse Estimators are Necessarily Large," *Sankhyā*, 71, 1–18. [607]

Pötscher, B. M., and Leeb, H. (2009), "On the Distribution of Penalized Maximum Likelihood Estimators: The LASSO, SCAD, and Thresholding," *Journal of Multivariate Analysis*, 100, 2065–2082. [607]

Ravikumar, P., Wainwright, M., and Lafferty, J. (2010), "High-Dimensional Ising Model Selection Using *l*1-Regularized Logistic Regression," *Annals of Statistics*, 38, 1287–1319. [606]

Rudelson, M., and Vershynin, R. (2008), "On Sparse Reconstruction From Fourier and Gaussian Measurements," *Communications on Pure and Applied Mathematics*, 61, 1025–1045. [611]

Rudelson, M., and Zhou, S. (2011), "Reconstruction From Anisotropic Random Measurements," *ArXiv:1106.1151*. [611]

Tibshirani, R. (1996), "Regression Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Society,* Series B, 58, 267–288. [606]

van de Geer, S., Bühlmann, P., Ritov, Y., and Dezeure, R. (2014), "On Asymptotically Optimal Confidence Regions and Tests for High-Dimensional Models," *Annals of Statistics*, 42, 1166–1202. [607]

van de Geer, S. A. (2008), "High-Dimensional Generalized Linear Models and the Lasso," *Annals of Statistics*, 36, 614–645. [606]

Zhang, C.-H., and Zhang, S. S. (2014), "Confidence Intervals for Low-Dimensional Parameters With High-Dimensional Data," *Journal of the Royal Statistical Society,* Series B, 76, 217–242. [607]