Finite Sample Inference for the Maximum Score Estimand*

Adam M. Rosen
Duke University and CeMMAP†

Takuya Ura
University of California, Davis‡

July 15, 2024

Abstract

We provide a finite sample inference method for the structural parameters of the semiparametric binary response model under a conditional median restriction originally studied by Manski (1975, 1985). This is achieved by exploiting distributional properties of observable outcomes conditional on the observed sequence of exogenous variables. Moment inequalities conditional on the size n sequence of exogenous covariates are constructed, and the proposed test statistic is a monotone function of violations of the corresponding sample moment inequalities. The critical value used for inference is provided by the appropriate quantile of a known function of n independent Bernoulli random variables, and does not require the use of a cube root asymptotic approximation employing a point estimator of the target parameter. Simulation studies demonstrate favorable finite sample performance of the test in comparison to several existing approaches. Empirical use is illustrated with an application to the classical setting of transportation choice.

JEL classification: C12 C14.

Keywords: Finite sample inference, Maximum score estimation, Moment inequalities, Partial identification.

*This is a revised version of the May 2020 CeMMAP working paper CWP 22/20. We thank Tim Armstrong, Federico Bugni, Bryan Graham, Jiaying Gu, Michal Kolesár, Sokbae Lee, Jia Li, Chuck Manski, Matt Masten, Ilya Molchanov, Francesca Molinari, Ulrich Müller, Whitney Newey, Pepe Montiel Olea, Thomas Russell, Chris Sims, and seminar and conference participants at Columbia, Duke, National University of Singapore, Penn State, Princeton, Singapore Management University, Vanderbilt, Yale, the 2019 Triangle Econometrics Conference, the 2019 Southern Economic Association Conference, the 2019 California Econometrics Conference, the 2019 Young Econometricians Conference, the cemmap/WISE Workshop on Advances in Econometrics, a cemmap/Turing Institute Economic Data Science Workshop, and the University of Tokyo Workshop on Advances in Econometrics for helpful comments and discussion. We are grateful to Gurobi Optimization for their free academic license. Kuong (Lucas) Do, Cheuk Fai Ng, and especially Xinyue Bei provided excellent research assistance at various stages of this project. Financial support from the Economic and Social Research Council ESRC Large Research Grant ES/P008909/1 to the Centre for Microdata Methods and Practice is gratefully acknowledged. Takuya Ura acknowledges financial support from Small Grants in Aid of Research at UC Davis.

†Address: Adam Rosen, Department of Economics, Duke University, 213 Social Sciences Box 90097, Durham, NC 27708; Email: adam.rosen@duke.edu

‡Address: Takuya Ura, Department of Economics, University of California, Davis, One Shields Avenue, Davis, CA 95616; Email: takura@ucdavis.edu

1 Introduction

Point identification of parameters relies on exogenous variables exhibiting sufficient variation. Precisely what properties constitute sufficient variation depends on the model employed, for example taking the form of a rank condition in parametric models with linear index restrictions. Many semiparametric and nonparametric models rely on additional support conditions to ensure point identification. Hypothesis tests and confidence intervals in turn rely on the asymptotic distribution of test statistics in order to achieve adequate asymptotic performance. The hope is that such asymptotic characterizations provide a suitable approximation to the finite sample distribution of these test statistics.

The support conditions that ensure point identification in semiparametic models often require that at least one of the exogenous variables is continuously distributed, and, moreover that it has full support on the real line, so that it can take values of arbitrarily large magnitude with positive probability.[1] In contrast to rank conditions, if the support of exogenous variables in the population were the finite set of points observed in an actual data set, such conditions could not possibly be satisfied and the parameter of interest would not necessarily be point identified. One may then wonder whether a test statistic whose asymptotic distribution presupposes point identification in fact provides a reasonable approximation to its finite sample distribution.

We examine this issue in the context of Manski's (1985) semiparametric binary response model, a central model in the literature on semiparametric econometrics.[2] Like all semiparametric models, the model features a parametric component $\beta \in \mathbb{R}^K$ and a nonparametric component, in this case, a distribution-free specification of unobservable heterogeneity. We introduce the concept of the *set of conditionally observationally equivalent parameters*, denoted $\mathcal{B}_n^*$, as the set of parameter vectors b that satisfy the observable implications of the model *conditional* on a size n sequence of observable exogenous covariate vectors $\mathcal{X}_n \equiv (X_1, ..., X_n)$.[3] This is the set of parameter vectors that would comprise the identified set if the support of exogenous variables were in fact their observed support in the finite sample. Thus, the difference between $\mathcal{B}_n^*$ and the identified set for β is that the former is based on the conditional distribution of the outcome variables given $X = x$ for values of

[1] Formally these support conditions on the distribution of a single exogenous variable are required to hold conditional on any value of the other exogenous variables.

[2] This model is a binary outcome version of the model introduced by Manski (1975), cited by Powell (1994) as the earliest example of semiparametric analysis of limited dependent variable models. Estimation and inference on the parameters of this model – the maximum score estimand – have featured prominently in the literature on semiparametric estimation.

[3] In some earlier drafts of this paper this set of conditionally observationally equivalent parameters was called the "finite sample identified set".

$x \in \mathcal{X}_n$, while the latter is based on the conditional distribution for values of $x \in \mathcal{S}_X$, the support of exogenous variables in the population. As we illustrate, the set $\mathcal{B}_n^*$ is useful for understanding both the possibilities and limitations of finite sample inference conditional on $\mathcal{X}_n$. While our focus in this paper is on the semiparametric binary response model, the set of conditionally observationally equivalent parameters could be explicitly defined and potentially used for studying finite sample inference using other models too.

The model features a binary outcome determined by the linear index threshold-crossing specification

$$Y = 1\{X\beta + U \geq 0\},$$

for observable variables $Y \in \{0, 1\}$ and X a row vector in $\mathbb{R}^K$, where the unobservable variable U is restricted to satisfy the zero conditional median restriction

$$\text{median}\,(U \mid X) = 0.$$

This semiparametric model is thus distribution-free with regard to unobservable U.[4] Full stochastic independence between U and X is not required, allowing for the conditional distribution of U given $X = x$ to vary with the conditioning value x, and thus accommodating general forms of heteroskedasticity. Under a rank condition and a large support condition on a continuous regressor Manski (1985) established point identification of β up to scale, as well as the large deviations convergence rate of the maximum score estimator. Several further analyses of the maximum score and similar estimators for this and closely related semiparametric binary response models have since been provided, and the literature on the *asymptotic* properties of the maximum score estimator is now vast.[5]

In contrast to prior approaches for inference on β that employ asymptotic distributional approximations, in this paper we develop a method for conducting *finite sample* inference on β. To do this we provide a conditional moment inequality characterization of the aforementioned set $\mathcal{B}_n^*$. Moment inequality characterizations of the model's implications have

[4] As noted in Manski (1985), his analysis easily generalizes to cover the restriction that the conditional τth quantile of U given X is 0, where $\tau \in (0, 1)$ is known. The analysis in this paper can be similarly generalized.

[5] Kim and Pollard (1990) showed that the convergence rate of the maximum score estimator is $n^{-1/3}$ and established its nonstandard asymptotic distribution after appropriate centering and scaling. Horowitz (1992) developed a smoothed maximum score estimator that converges faster than the $n^{-1/3}$ rate and is asymptotically normal under some additional smoothness assumptions. Additional papers that study large sample estimation and inference applicable in the maximum score context include Manski and Thompson (1986), Delgado, Rodríguez-Poo, and Wolf (2001), Abrevaya and Huang (2005), Léger and MacGibbon (2006), Komarova (2013), Blevins (2015), Jun, Pinkse, and Wan (2015, 2017), Chen and Lee (2018, 2019), Patra, Seijo, and Sen (2018), Seo and Otsu (2018), Cattaneo, Jansson, and Nagasawa (2020), Mukherjee, Banerjee, and Ritov (2021), and Khan, Komarova, and Nekipelov (2024).

been previously used by Komarova (2013), Blevins (2015), and Chen and Lee (2019). More recently Khan, Komarova, and Nekipelov (2024) introduce a random set quantile estimator for partially identifying discrete response models with discrete regressors and provide novel comparative analyses of the estimation properties of several estimators, including the maximum score estimator for binary response. None of these papers propose an inference method which is guaranteed to have finite sample validity, which is our focus here. As was the case in the analysis provided in these papers, we do not require that β is point identified. For instance, we do not require that any component of X is continuously distributed, much less with large support. Our proposed test is valid both when such conditions hold, and when they do not.

The approach taken here exploits the implication of the binary response model that conditional on $\mathcal{X}_n$, each outcome Y_i follows a Bernoulli distribution. In practice, the Bernoulli probabilities are unknown. Nonetheless, conditional on $\mathcal{X}_n$, the probability $\mathbb{P}\left(Y_i = 1 \mid \mathcal{X}_n\right) = \mathbb{P}\left(U_i \geq -X_i\beta \mid \mathcal{X}_n\right)$ is bounded from above or below by $1/2$ according to the sign of $X_i\beta$. Consequently, for any known nonnegative-valued function $g\left(\cdot\right)$, the finite sample distributions of $(2Y_i - 1)1\{X_i\beta \geq 0\}g\left(X_i\right)$ and $(1 - 2Y_i)1\{X_i\beta \leq 0\}g\left(X_i\right)$ conditional on $\mathcal{X}_n$ can be bounded from below. Intuitively this is achieved by the distribution of the test statistic under a least favorable configuration in which the unobservables U_i have large probability mass in the tails, which minimizes the contribution of β for the determination of Y_i. The test statistic $T_n(b)$ that we use to implement our test of the null hypothesis $H_0 : \beta = b$ is a supremum of weighted sample averages of $-(2Y_i - 1)1\{X_ib \geq 0\}g\left(X_i\right)$ and $-(1 - 2Y_i)1\{X_ib \leq 0\}g\left(X_i\right)$, where the supremum is taken over particular collections of functions $g\left(\cdot\right)$. The test statistic $T_n(b)$ is shown to be bounded from above by a function $T_n^*(b)$ of n independent Bernoulli random variables, such that the finite sample distribution of $T_n^*(b)$ given $\mathcal{X}_n$ is known. Then, under the null hypothesis $\beta = b$, we have

$$\mathbb{P}\left(T_n(b) > q_{1-\alpha}(b) \mid \mathcal{X}_n\right) \leq \alpha,$$

where $q_{1-\alpha}(b)$ is the conditional $1 - \alpha$ quantile of $T_n^*(b)$ given $\mathcal{X}_n$. We further show that if particular functions $g\left(\cdot\right)$ are used, the moment functions which $T_n(b)$ incorporates fully characterize the set of conditionally observationally equivalent parameters.

For any conditionally observationally equivalent parameter value $b \in \mathcal{B}^*$, we show that any test that achieves size control rejects the null hypotheses $H_0 : \beta = b$ with probability no greater than α conditional on $\mathcal{X}_n$. We subsequently establish a lower bound on the rejection probability of our test for values of $b \neq \beta$, and we show that this bound is increasing in the degree to which the hypothesized parameter vector b violates the inequalities that

characterize the set of conditionally observationally equivalent parameters. Thus values of b that are sufficiently far from the set of conditionally observationally equivalent parameters by this measure are guaranteed to be rejected with probability exceeding the size of the test.

For the sake of comparison we also consider likelihood ratio tests, which have favorable minimax properties under general conditions, see e.g. chapter 8 of Lehmann and Romano (2005). The minimax properties of a test for the null hypothesis of $\beta = b$ depend on the specification of the alternative hypothesis. The likelihood ratio test is a most powerful test when the alternative hypothesis is simple, meaning that it specifies a unique conditional distribution of $(Y_1, ..., Y_n)$ given $\mathcal{X}_n$. However, when the alternative hypothesis is of the form $\beta \neq b$ or $\beta = \tilde{b}$, many possible distributions of $(Y_1, ..., Y_n)$ given $\mathcal{X}_n$ are permitted under the conditional median restriction. This alternative hypothesis is therefore composite. Indeed, we show that the sets of possible distributions of $(Y_1, ..., Y_n)$ given $\mathcal{X}_n$ admitted by both the alternative and null hypothesis have nonempty intersection, so that they are not robustly testable in the sense of Kaido and Zhang (2019, pp. 12-13). A completely randomized test that rejects with probability α irrespective of the data is minimax optimal, suggesting that minimax optimality may not be a suitable criterion for comparing tests in this setting. We further establish a connection between the profile log-likelihood that profiles out the unknown distribution of unobservable heterogeneity and the maximum score criterion function.

Among the aforementioned papers from the literature on maximum score, the most closely related is that of Chen and Lee (2019), who also cast the implications of Manski's (1985) model as conditional moment inequalities for the sake of delivering a new insight, albeit one that is different from ours. Chen and Lee (2019) expand on the conditional moment inequalities used by Komarova (2013) and Blevins (2015) to develop a novel conditional moment inequality characterization of the identified set which involves conditioning on two linear indices instead of on the entire exogenous covariate vector. They apply intersection bound inference from Chernozhukov, Lee, and Rosen (2013) to this conditional moment inequality characterization to achieve asymptotically valid inference. This cleverly exploits the model's semiparametric linear index restriction in order to sidestep the curse of dimensionality. Although a good deal of focus is given to Manski's (1985) binary response model, their method can also be applied to other semiparametric models.

This paper appears to be the first to propose a method that ensures valid finite sample inference for β in Manski's (1985) semiparametric binary response model. This is the main contribution of the paper. Furthermore, although it is well known that identification analysis relies on properties of the support of exogenous variables, this paper is also the first to formally introduce the concept of a set of conditionally observationally equivalent parameters, explicitly defining the set of model parameters logically consistent with the modeling restric-

tions and only information that can be gathered from observable implications conditional on realizations of exogenous variables observed in the finite sample. This concept can be applied to other models, and may therefore be of independent interest.

There are a handful of precedents for employing finite sample inference with other partially identifying models. Manski (2007) considers the problem of predicting choice probabilities for the choices individuals would make if subjected to counterfactual variation in their choice sets. In the absence of the structure afforded by commonly used random utility models, he shows that counterfactual choice probabilities are partially identified, and proposes a procedure for inference using results from Clopper and Pearson (1934). Chernozhukov, Hansen, and Jansson (2009) propose a finite sample inference method for quantile regression models in which the outcome is continuously distributed. Their approach exploits a "conditionally pivotal property" to bound the finite sample distribution of a GMM criterion incorporating moment equalities, but which does not require point identification for its validity.[6] Syrgkanis, Tamer, and Ziani (2018) conduct inference on partially identified parameters in auction models imposing weak assumptions on bidders' information. They propose a method to conduct finite sample inference on moments of functions of the underlying valuation distribution using concentration inequalities. Armstrong and Kolesár (2021) provide methods for optimal inference on average treatment effects that are finite sample valid in the special case in which regression errors are normal, and which are asymptotically valid more generally. Their conditions cover cases where identification may fail due to a lack of overlap of the support of conditioning variables. The recent working paper Li and Henry (2022) proposes a method for finite sample inference in parametric incomplete models based on an optimal transport characterization of the identified set, employing Monte Carlo tests. The approach taken in this paper for finite sample inference in the context of Manski's (1985) binary response model is different from all of these.

The rest of this paper is organized as follows. Section 2 formally sets out the testing problem and the moment inequality representation of the set of conditionally observationally equivalent parameters. Section 3 lays out the main results of the paper, namely the construction of the test statistic and corresponding critical value, and the establishment of our test's finite sample validity as well as the aforementioned finite sample (lower) power bound. Section 4 considers likelihood ratio tests. We establish a power envelope for tests of the hypothesis $\beta = b$ by directing power against simple alternatives and additionally show that the maximum score estimator admits an interpretation as the maximizer of the profile

[6]Chernozhukov, Hansen, and Jansson (2009) study instrumental variable quantile regression with endogenous covariates. In Appendix D, we outline an extension of our test in Section 3 to the case when X is endogenous.

likelihood in which the distribution of unobservable heterogeneity is treated as a nuisance parameter. Section 5 demonstrates the performance of our approach relative to several others by reporting results from Monte Carlo simulations. Section 6 applies our proposed test to the dataset from Horowitz (1993). Section 7 concludes and discusses avenues for future research. All proofs are in the Appendix. Unless otherwise stated, our analysis throughout this paper should be read as conditional on observable covariate vectors $\mathcal{X}_n \equiv (X_1, ..., X_n)$.

2 Model and Moment Restrictions

This section is divided into three subsections, the first of which formally presents the modeling restrictions imposed. The second subsection characterizes the observable implications of the binary response model *conditional* on a size n sequence of covariate vectors, $\mathcal{X}_n$, in contrast to those observable implications obtainable from knowledge of the population distribution of observable variables. Based on these observable implications, this second subsection introduces our definition of the set of conditionally observationally equivalent parameters, $\mathcal{B}_n^*$. The set clarifies which conjectured parameter values a test can feasibly detect, as we show that for any test of $\beta = b$ against $\beta \neq b$ that achieves finite sample size control rejects any null value $b \in \mathcal{B}_n^*$ with probability no greater than the significance level of the test. The third subsection provides a moment inequality representation of $\mathcal{B}_n^*$ that is subsequently used in the construction of our test statistic. It further describes how recently developed incremental enumeration algorithms for hyperplane arrangements can be used to enumerate these inequalities.

2.1 Model

The following assumptions formalize the restrictions of the semiparametric binary response model under study and the requirements on the sampling process. We maintain these Assumptions 1(i)–1(v) throughout this paper.

Assumption 1. *(i) Random vectors* $\{(Y_i, X_i, U_i) : i = 1, ..., n\}$ *reside on a probability space* $(\Omega, \mathfrak{F}, \mathbb{P})$. *(ii) Variables* $\{(Y_i, X_i) : i = 1, ..., n\}$ *are observed. (iii) There is a column vector* $\beta \in \mathbb{R}^K$ *such that* $\mathbb{P}(Y_i = 1\{X_i\beta + U_i \geq 0\} \mid \mathcal{X}_n) = 1$ *and* $\mathbb{P}(U_i \geq 0 \mid \mathcal{X}_n) = 1/2$ *for every* $i = 1, \ldots, n$, *where* $\mathcal{X}_n \equiv (X_1, ..., X_n)$. *(iv) There is a known set* $\mathcal{B} \subseteq \mathbb{R}^K$ *to which* β *belongs. (v) The unobservable variables* $(1\{U_1 \geq 0\}, ..., 1\{U_n \geq 0\})$ *are mutually independent conditional on* $\mathcal{X}_n$.

The requirements of Assumption 1 are slightly weaker than the assumptions used in the existing literature (e.g. Manski, 1975, 1985). Parts (i), (ii), and (iv) are standard.

Although it is not necessary in this paper because we employ partial identification analysis, the parameter space $\mathcal{B}$ can be restricted by imposing one of the usual scale normalizations from the literature, such as $|b_1| = 1$ for all $b \in \mathcal{B}$. Part (iii) imposes the binary response structure and the requirement that $\mathbb{P}(U_i \geq 0 \mid \mathcal{X}_n) = 1/2$ for each i, which follows from the inequality of display (1) in Manski (1985, page 315). Binary response models typically require that U_i is continuously distributed in a neighborhood of zero, in which case this is implied by a conditional median restriction. Strictly speaking, we do not need to impose that each U_i is continuously distributed at zero to ensure our proposed test achieves finite sample size control, and hence we replace the median restriction with this weaker requirement.[7] Part (v) holds if (Y_i, X_i, U_i) are independent and identically distributed, but is much more general. Throughout the text, $\mathbb{E}$ is used to denote population expectation with respect to $\mathbb{P}$, and $\mathbb{E}_n \equiv n^{-1} \sum_{i=1}^{n}$.

In the model specified by Assumption 1, knowledge of β is insufficient to uniquely determine the conditional distribution of $(Y_1, ..., Y_n)$ given $\mathcal{X}_n$, or even the conditional probability of $Y_i = 1$ for any i. One additionally requires knowledge of the joint distribution of $(U_1, ..., U_n)$ given $\mathcal{X}_n$, here denoted G_0. In other words, G_0 is a nuisance parameter that is restricted by Assumption 1, but is otherwise left unspecified. The conditional distribution of $(Y_1, ..., Y_n)$ given $\mathcal{X}_n$ is uniquely determined by (β, G_0):

$$\mathbb{P}(Y_i = y_i \text{ for all } i \mid \mathcal{X}_n) = P_{(\beta, G_0)}(Y_i = y_i \text{ for all } i \mid \mathcal{X}_n), \tag{2.1}$$

where for any $(b, G) \in \mathcal{B} \times \mathcal{G}$, $P_{(b,G)}(Y_i = y_i \text{ for all } i \mid \mathcal{X}_n) = G(1\{X_i b + U_i \geq 0\} = y_i \text{ for all } i \mid \mathcal{X}_n)$.[8] Note that $\mathcal{G}$ denotes the set of possible distributions of $(U_1, ..., U_n)$ given $\mathcal{X}_n$ satisfying Assumption 1.[9]

In this paper, we consider the hypothesis test

$$H_0 : \beta = b \quad \text{versus} \quad H_1 : \beta \neq b, \tag{2.2}$$

where b is an arbitrary element of $\mathcal{B}$. Using the nuisance parameter G_0, the null and alter-

[7] Appendix G discusses implications of additionally require each U_i to be continuously distributed with strictly increasing CDF conditional on $\mathcal{X}_n$.

[8] It is to be understood here that writing that an event holds "for all i" means that it holds for all $i = 1, ..., n$. The expression $G(1\{X_i b + U_i \geq 0\} = y_i \text{ for all } i \mid \mathcal{X}_n)$ denotes the conditional probability of $(1\{X_1 b + U_1 \geq 0\}, ..., 1\{X_n b + U_n \geq 0\}) = (y_1, ..., y_n)$ given $\mathcal{X}_n$ when G is the conditional distribution of $(U_1, ..., U_n)$ given $\mathcal{X}_n$.

[9] In other words, $\mathcal{G}$ is the set of conditional distributions G of $(U_1, ..., U_n)$ given $\mathcal{X}_n$ such that $G(U_i \geq 0 \mid \mathcal{X}_n) = 1/2$ for every $i = 1, \ldots, n$ and that $(1\{U_1 \geq 0\}, ..., 1\{U_n \geq 0\})$ are mutually independent conditional on $\mathcal{X}_n$ under G.

native hypotheses in (2.2) can be equivalently expressed as

$$H_0 : (\beta, G_0) = (b, G) \text{ for some } G \in \mathcal{G},$$

and

$$H_1 : (\beta, G_0) = (\tilde{b}, \tilde{G}) \text{ for some } \tilde{b} \in \mathcal{B} \text{ with } \tilde{b} \neq b \text{ and some } \tilde{G} \in \mathcal{G}.$$

The objective of this paper is to propose a test of (2.2) with binary rejection rule $\phi(b, Y_1, ..., Y_n; \mathcal{X}_n)$ that achieves finite sample size control for $H_0 : \beta = b$ when the null hypothesis is true. The rejection probability of such a test when $\beta = b$ and $G_0 = G$ (conditional on $\mathcal{X}_n$) is

$$P_{(b,G)}(\phi(b, Y_1, ..., Y_n; \mathcal{X}_n) = 1 \mid \mathcal{X}_n) = \sum_{(y_1,...,y_n)\in\{0,1\}^n} \phi(b, y_1, ..., y_n; \mathcal{X}_n) P_{(b,G)} \left(Y_i = y_i \text{ for all } i \mid \mathcal{X}_n\right).$$

A test $\phi(b, Y_1, ..., Y_n; \mathcal{X}_n)$ achieves finite sample size control if

$$\sup_{G\in\mathcal{G}} P_{(b,G)}(\phi(b, Y_1, ..., Y_n; \mathcal{X}_n) = 1 \mid \mathcal{X}_n) \leq \alpha. \tag{2.3}$$

We focus on deterministic tests that achieve finite sample size control conditional on $\mathcal{X}_n$. The above inequality could easily be made an equality by employing a randomized test.

The power result presented in Theorem 5 in Section 3 and the results of Sections 4.1–4.3 additionally invoke the following assumption.

Assumption 2. *Unobservable variables $(U_1, ..., U_n)$ are mutually independent conditional on $\mathcal{X}_n$.*

Assumption 2 is common in the prior literature on maximum score estimation, but is not required for many of the results in this paper. In particular, it is not necessary to establish finite sample size control for our test. The assumption is satisfied in models that restrict (X_i, U_i) to be i.i.d., implied for example by Assumption 3 of Manski (1985). Note however that Assumption 2 does not require $(U_1, ..., U_n)$ to be i.i.d. given $\mathcal{X}_n$, but only mutually independent.

2.2 Observable Implications Conditional on $\mathcal{X}_n$

To conduct finite sample inference, we focus solely on the implications obtainable from the conditional distribution of $(Y_1, ..., Y_n)$ given $\mathcal{X}_n$. The approach does not rely on features of the population conditional distribution of Y given values of X that may be on the support of X but that are not realized in the sample. The set of parameters satisfying all implications of the conditional distribution of $(Y_1, ..., Y_n)$ given $\mathcal{X}_n$ is by definition the set of parameter

vectors $b \in \mathcal{B}$ such that the conditional distribution of $(Y_1, ..., Y_n)$ given $\mathcal{X}_n$ matches that of $\tilde{Y}_i \equiv 1\{X_i b + \tilde{U}_i \geq 0\}$ for a sequence of random variables $(\tilde{U}_1, ..., \tilde{U}_n)$ that satisfy the restrictions placed on the conditional distribution of $U_1, ..., U_n$ given $\mathcal{X}_n$ in Assumption 1. This is the set we refer to as the set of conditionally observationally equivalent parameters, which we denote $\mathcal{B}_n^*$.

Definition 1. *The **set of conditionally observationally equivalent parameters** for β, denoted $\mathcal{B}_n^*$, is the set of $b \in \mathcal{B}$ for which there exist random variables $\{\tilde{Y}_i : i = 1, ..., n\}$ and $\{\tilde{U}_i : i = 1, ..., n\}$ on $(\Omega, \mathfrak{F}, \mathbb{P})$ such that:*

(i): $\mathbb{P}\left(\tilde{Y}_i = 1\{X_i b + \tilde{U}_i \geq 0\} \text{ for all } i \mid \mathcal{X}_n\right) = 1,$

(ii): $(\tilde{Y}_1, ..., \tilde{Y}_n)$ *and* $(Y_1, ..., Y_n)$ *have the same conditional distribution given* $\mathcal{X}_n$*,*

(iii): $\mathbb{P}\left(\tilde{U}_i \geq 0 \mid \mathcal{X}_n\right) = 1/2$ *for every* $i = 1, \ldots, n,$

(iv): $\{1\{\tilde{U}_i \geq 0\} : i = 1, \ldots, n\}$ *are mutually independent given* $\mathcal{X}_n$*.*

The set $\mathcal{B}_n^*$ is thus determined by $\mathcal{X}_n$ and the conditional distribution of $(Y_1, ..., Y_n)$ given $\mathcal{X}_n$. Each parameter vector $b \in \mathcal{B}_n^*$ cannot be distinguished from the population parameter β on the basis of this conditional distribution. The set $\mathcal{B}_n^*$ involves the unknown population distribution of $(Y_1, ..., Y_n)$ given $\mathcal{X}_n$, so it is an unknown object, even after $\mathcal{X}_n$ is realized. Notably, it is not $\text{argmax}_{b \in \mathcal{B}} S_n(b)$, the set of Manski's (1985) maximum score estimators, where the sample score function is defined by $S_n(b) \equiv \frac{1}{n} \sum_{i=1}^n (2Y_i - 1)\text{sgn}(X_i b)$ with $\text{sgn}(X_i b) \equiv 1\{X_i b \geq 0\} - 1\{X_i b < 0\}$. The set $\text{argmax}_{b \in \mathcal{B}} S_n(b)$ is a sample object that can be computed directly from sample data.

In this paper, the set $\mathcal{B}_n^*$ plays a key role in testing (2.2). Let b denote any element of $\mathcal{B}_n^*$ and consider a level α test of (2.2) with null hypothesis $H_0 : \beta = b$. Theorem 1 below establishes that *any* such test that controls size has rejection probability less than or equal to α, conditional on $\mathcal{X}_n$.[10]

Theorem 1. *Let Assumption 1 hold, let b be any element of $\mathcal{B}_n^*$, and let $\phi(b, Y_1, ..., Y_n; \mathcal{X}_n)$ be any rejection rule that achieves finite sample size control for $H_0 : \beta = b$, thus satisfying* (2.3)*. Then*

$$P_{(\beta, G_0)}(\phi(b, Y_1, ..., Y_n; \mathcal{X}_n) = 1 \mid \mathcal{X}_n) \leq \alpha.$$

[10]Theorem 9 in the appendix further shows that this statement remains true if we additionally require that the cumulative distribution function of U_i given $\mathcal{X}_n$ is strictly increasing, in which case the conditional mean of U_i given $\mathcal{X}_n$ is unique.

Here $P_{(\beta,G_0)}(\phi(b, Y_1, ..., Y_n; \mathcal{X}_n) = 1 \mid \mathcal{X}_n)$ is the rejection probability of test $\phi(b, Y_1, ..., Y_n; \mathcal{X}_n)$ for the hypothesis $\beta = b$ under the population data generation process produced by (β, G_0). Theorem 1 allows $b \neq \beta$ as long as b belongs to $\mathcal{B}_n^*$. In this case, the conclusion of the theorem considers the rejection probability of a test ϕ under the population distribution of $Y_1, ..., Y_n$ given $\mathcal{X}_n$ when the null hypothesis is false (i.e., $\beta \neq b$). It establishes that any finite sample valid test for $H_0 : \beta = b$ rejects with probability no higher than α when $b \in \mathcal{B}_n^*$. Theorem 1 holds because when the null and alternative hypotheses restrict the class of possible G_0 no further than the restrictions maintained under Assumption 1, then, for any $b \in \mathcal{B}_n^*$, it is always possible to find a $\tilde{G} \in \mathcal{G}$ that produces $\mathbb{P}(Y_i = y_i \text{ for all } i \mid \mathcal{X}_n) = P_{(b,\tilde{G})}(Y_i = y_i \text{ for all } i \mid \mathcal{X}_n)$, the same distribution induced by β and the unknown population distribution of G_0.

While Theorem 1 shows no test of (2.2) that achieves finite sample size control can reject a null of $\beta = b$ with probability greater than α when $b \in \mathcal{B}_n^*$ conditional on $\mathcal{X}_n$, it is possible to do so when $b \notin \mathcal{B}_n^*$, as we will later show. In models in which conditions for point identification of β are satisfied it is well known that it is possible to consistently estimate β at the $n^{-1/3}$ rate and hence achieve nontrivial *asymptotic* power against $b \neq \beta$. Under such conditions the set $\mathcal{X}_n$ approaches $\text{Supp}(X)$ as n increases, and in the limit the set $\mathcal{B}_n^*$ will then converge to the singleton set $\{\beta\}$. Theorem 1 aligns with the ability to achieve non-trivial *asymptotic* power against $b \neq \beta$ because for sufficiently large n, $b \neq \beta$ implies that $b \notin \mathcal{B}_n^*$.

Our next task in developing our test is to express $\mathcal{B}_n^*$ with a moment inequality representation useful for inference. The following lemma sets out two observable implications for this purpose.

Lemma 1. *Under Assumption 1,*

$$X_i\beta \geq 0 \implies \mathbb{E}[2Y_i - 1 \mid \mathcal{X}_n] \geq 0, \tag{2.4}$$

$$X_i\beta \leq 0 \implies \mathbb{E}[2Y_i - 1 \mid \mathcal{X}_n] \leq 0. \tag{2.5}$$

From the inequalities of the lemma, it further follows that if $X_i\beta = 0$ then $\mathbb{E}[2Y_i - 1 \mid \mathcal{X}_n] = 0$. Moreover (2.4) and (2.5) and their implications described above hold with β replaced by any b that is an element of the set of conditionally observationally equivalent parameters $\mathcal{B}_n^*$. This can be proven by following precisely the same steps as in the proof of the lemma with $\tilde{U}_i$ from Definition 1 replacing U_i.

With Lemma 1 in hand, the following theorem provides a moment inequality characterization of $\mathcal{B}_n^*$.

Theorem 2. *Under Assumption 1, the set of conditionally observationally equivalent parameters is*

$$\mathcal{B}_n^* = \{b \in \mathcal{B} : \mathbb{E}\left[(2Y_i - 1)\, 1\{X_i b \geq 0\} \mid \mathcal{X}_n\right] \geq 0 \geq \mathbb{E}\left[(2Y_i - 1)\, 1\{X_i b \leq 0\} \mid \mathcal{X}_n\right] \text{ for all } i\}\,.$$

The conditional moment inequalities characterizing $\mathcal{B}_n^*$ in Theorem 2 are equivalent to (2.4) and (2.5) for all $i = 1, ..., n$. However, using this conditional moment inequality representation to conduct inference on β is complicated by the fact that in a sample of n observations the distribution of Y_i given $\mathcal{X}_n$ can vary across i, even if $(Y_i, X_i) : i = 1, ..., n$ are identically distributed, and there is only one observation of (Y_i, X_i) for each i. Thus, to make these inequalities operational for inference, some level of aggregation across i is required. The following section considers the use of unconditional moment inequalities for precisely this purpose.

2.3 Observable Implications as Unconditional Moment Inequalities

In order to construct unconditional moment inequalities that hold for each $b \in \mathcal{B}_n^*$, let

$$\{g_u(\cdot, v) : v \in \mathcal{V}_u\}, \qquad \{g_l(\cdot, v) : v \in \mathcal{V}_l\} \tag{2.6}$$

be collections of nonnegative-valued instrument functions indexed by $v \in \mathcal{V}_u$ and $v \in \mathcal{V}_l$, respectively. That is, for any $v_u \in \mathcal{V}_u$ and $v_l \in \mathcal{V}_l$, $g_u(\cdot, v_u) : \mathcal{S}_X \to \mathbb{R}_+$ and $g_l(\cdot, v_l) : \mathcal{S}_X \to \mathbb{R}_+$ are functions that map from $\mathcal{S}_X$, the support of X, to the nonnegative real numbers. Since these functions are nonnegative-valued, Theorem 2 implies that for any $b \in \mathcal{B}_n^*$, and all $v_u \in \mathcal{V}_u$ and $v_l \in \mathcal{V}_l$:

$$\mathbb{E}\left[\mathbb{E}_n\left[(2Y - 1)1\{Xb \geq 0\} g_u(X, v_u)\right] \mid \mathcal{X}_n\right] \geq 0, \tag{2.7}$$

$$\mathbb{E}\left[\mathbb{E}_n\left[(1 - 2Y)1\{Xb \leq 0\} g_l(X, v_l)\right] \mid \mathcal{X}_n\right] \geq 0. \tag{2.8}$$

The test statistic developed in Section 3 incorporates empirical analogs of moment inequalities of this form. Here we focus on instrument functions (2.6) of a particular form, as we now describe, while noting that the steps taken to prove finite sample size control in Theorem 4 can in fact be applied to an analogous test statistic employing *any* choice of nonnegative-valued instrument functions (2.6) that do not depend on $(Y_1, ..., Y_n)$. The choice of instrument functions will in general affect the power of the test. Here we consider a different criterion from an identification perspective – namely, how to choose a sufficiently rich collection of in-

strument functions such that the unconditional inequalities (2.7) and (2.8) fully characterize $\mathcal{B}_n^*$.

For this purpose, consider a comparison of the hypothesized parameter b to a parameter v. For values of X_i such that the sign of $X_i v$ is the same as $X_i b$, both $\beta = v$ and $\beta = b$ deliver the same implication for whether $\mathbb{E}[2Y_i - 1|X_i]$ is nonnegative or nonpositive. When instead $X_i v$ has the opposite sign of $X_i b$, then the two parameter vectors make different predictions for the sign of $\mathbb{E}[2Y_i - 1|X_i]$. Heuristically, parameter vector b does a better job at predicting the sign of $\mathbb{E}[2Y_i - 1|X_i]$ if the sign of $X_i b$ matches that of $\mathbb{E}[2Y_i - 1|X_i]$ more often than the sign of $X_i v$ does. This is the intuition that underlies Manski's (1985) maximum score estimator.

This intuition motivates the use of instrument functions of the form $g_u(X, v) = 1\{Xv < 0\}$ and $g_l(X, v) = 1\{Xv > 0\}$. These functions are nonnegative, so (2.7) and (2.8) are implied for all $v \in \mathcal{B}$ by the conditional inequalities that characterize $\mathcal{B}_n^*$ provided by Theorem 2. As we now establish, ensuring that (2.7) and (2.8) hold for all $v \in \mathcal{B}$ provides a full characterization of $\mathcal{B}_n^*$. However, we do not need to consider these inequalities for all the possible values of v. For any v and $\tilde{v}$ such that $X_i v$ and $X_i \tilde{v}$ have the same sign for all i, the resulting inequalities in (2.7) and (2.8) are the same, so it is redundant to use inequalities that feature both v and $\tilde{v}$. To make use of this observation, we consider the hyperplanes $\{v \in \mathbb{R}^K : X_i v = 0\}$ for $i = 1, ..., n$ and the complement of their union, i.e., $\{v \in \mathbb{R}^K : X_i v \neq 0 \text{ for all } i\}$.[11] Further define $\mathsf{V}(\mathcal{X}_n)$ to be the partition of $\{v \in \mathbb{R}^K : X_i v \neq 0 \text{ for all } i\}$ according to the sequence of inequalities $X_i v < 0$ and $X_i v > 0$, $i = 1, ..., n$, with boundaries given by the hyperplanes $\{v \in \mathbb{R}^K : X_i v = 0\}$.[12] Such a collection of hyperplanes is referred to as a linear *hyperplane arrangement* in the computational geometry literature, and $\mathsf{V}(\mathcal{X}_n)$ is the partition induced by this hyperplane arrangement on $\{v \in \mathbb{R}^K : X_i v \neq 0 \text{ for all } i\}$.[13] The second part of Theorem 3 below implies that the moment inequality characterization using such instrument functions and such a collection of points provides a full characterization of $\mathcal{B}_n^*$ with no loss of identifying information.

[11] Points v at which $X_i v = 0$ for some i need not be considered. For such v the contribution of the term involving X_i to both inequalities (2.7) and (2.8) is zero for all b, and there must exist a perturbation of v, say $\tilde{v}$, such that $X_i \tilde{v} \neq 0$ for all i, which produces a smaller value of the left hand side of both (2.7) and (2.8) for all b. Thus if inequalities employing such values of $\tilde{v}$ are used, values of v for which $X_i v = 0$ for some i are redundant for characterizing $\mathcal{B}_n^*$.

[12] In an early paper comparing computational methods for the maximum score estimator Pinkse (1993, Section 3.3) used these hyperplanes to provide an exact but practically infeasible characterization of the maximum score estimator. More recently Florios and Skouras (2008) developed a mixed integer linear program for efficient computation of the estimator, see also Florios (2018) and Florios, Louka, and Bilias (2022).

[13] In Appendix C we provide a detailed illustration of the hyperplane arrangement and resulting cells for the simplest nontrivial case in which $K = 2$.

Theorem 3. *Let Assumption 1 hold. Then*

(i) If $b \in \mathcal{B}_n^*$ then for any $\mathcal{V}_u \subseteq \mathcal{B}$ and $\mathcal{V}_l \subseteq \mathcal{B}$:

$$\forall v \in \mathcal{V}_u : \mathbb{E}\left[\ \mathbb{E}_n\left[(2Y-1)1\{Xb \geq 0\}1\{Xv < 0\}\right] \mid \mathcal{X}_n\right] \geq 0, \tag{2.9}$$

and

$$\forall v \in \mathcal{V}_l : \mathbb{E}\left[\ \mathbb{E}_n\left[(1-2Y)1\{Xb \leq 0\}1\{Xv > 0\}\right] \mid \mathcal{X}_n\right] \geq 0. \tag{2.10}$$

(ii) If $\mathcal{V}_u = \mathcal{V}_l = \mathcal{V}$ and $\mathcal{V}$ contains at least one element from each cell of the partition $\mathsf{V}(\mathcal{X}_n)$ of $\left\{v \in \mathbb{R}^K : X_i v \neq 0 \text{ for all } i\right\}$ induced by the hyperplane arrangement $\left\{v \in \mathbb{R}^K : X_i v = 0\right\}$, then (2.9) and (2.10) imply that $b \in \mathcal{B}_n^*$, so that $\mathcal{B}_n^* = \{b \in \mathcal{B} :$ (2.9) and (2.10) hold$\}$.

The test statistic developed for inference in Section 3 employs sample analogs of moment inequalities of the form (2.9) and (2.10). The first part of Theorem 3 justifies their use for arbitrary collections of v. The second part establishes that the use of $\mathcal{V}_u = \mathcal{V}_l = \mathcal{V}$, where $\mathcal{V}$ comprises a collection of representatives that includes at least one element from each member of the partition $\mathsf{V}(\mathcal{X}_n)$, exhausts all the identifying power of these moment inequalities, conditional on $\mathcal{X}_n$. The use of values of v that belong to the same element of the partition is redundant; thus it is sufficient to use a set $\mathcal{V}$ that has precisely one representative point from each cell.

The first implication of Theorem 3 will imply asymptotic validity of the test we propose in Section 3 below. That is, asymptotic size control does not rely on the choice of $\mathcal{V}_u$ and $\mathcal{V}_l$: we can use any sets of values $\mathcal{V}_u$ and $\mathcal{V}_l$ to achieve finite sample size control. If however one wishes to use a sufficiently rich collection of moment inequalities to fully characterize $\mathcal{B}_n^*$, then it will be necessary to compute an exhaustive collection of representatives. In this case the number of required representative points grows quickly with both the dimension K of β and the sample size n. From Theorem 1 of Cover (1965), the maximal number of cells from such a dichotomy employing n hyperplanes in $\mathbb{R}^K$ is bounded from above by

$$2\sum_{j=0}^{K-1}\binom{n-1}{j},$$

which can be a very large number. This is equivalently the upper bound on the number of elements of $\mathsf{V}(\mathcal{X}_n)$, attained when every subset of K points from $X_1, ..., X_n$ are linearly independent.

Fortunately algorithms for enumeration of cells induced by the hyperplane arrangement determined by any sequence of points, $X_1, ..., X_n$, have been developed. Such algorithms have

been put to good use recently in econometrics by e.g. Gu and Koenker (2022) for the purpose of computing nonparametric maximum likelihood estimators for binary response models, Gu and Russell (2023) for computing bounds on certain counterfactuals in nonseparable models of binary response, and Gu, Russell, and Stringham (2022) for latent space enumeration for the sake of computing bounds on counterfactuals in a more general class of partially identifying models. Notable contributions in the development of such algorithms in the computational geometry literature include Avis and Fukuda (1996), Sleumer (1998), and Rada and Černý (2018).

While it is not necessary to compute an exhaustive collection of representatives for asymptotic size control, if one wishes to do so we have found the class of algorithms known as incremental enumeration methods useful for this purpose. These algorithms start with an initial hyperplane, for example in our context this could be the hyperplane $\left\{v \in \mathbb{R}^K : X_1 v = 0\right\}$, and take one representative point each from the region above and below the hyperplane. This is the partition of $\mathbb{R}^K$ induced by the singleton hyperplane arrangement $\left\{v \in \mathbb{R}^K : X_1 v = 0\right\}$. Incremental enumeration works by iteratively adding one of the hyperplanes $\left\{v \in \mathbb{R}^K : X_i v = 0\right\}$ for $i = 2, ..., n$, each time splitting $\mathbb{R}^K$ into the two regions above and below the newly added hyperplane, and adding representative points for any newly created cells in which there is not already a representative from prior iterations. Enumerating all hyperplanes in this way produces a representative point for each cell of the partition induced by the hyperplane arrangement.

This type of approach was proposed by Rada and Černý (2018), whose algorithm solves a sequence of linear programs in each iteration to find new representatives from the newly added hyperplane. Because linear programs can be solved very quickly, the algorithm performs well compared to earlier approaches, and we use it here in our simulation studies and our empirical application where $K > 2$. Nonetheless, the number of cells of the hyperplane arrangement can be very large. For our problem the sheer number of points required makes computation of our test statistic based on an exhaustive enumeration of cells costly. In order to speed up computation, one can impose an upper bound on the number of representative points v collected, such that the algorithm stops when this number is attained. While an exhaustive set of representatives is necessary for a sharp characterization of $\mathcal{B}_n^*$, any set of points achieves valid finite sample inference. In our Monte Carlo analysis in Section 5 we used the first 500 points collected by the Rada and Černý (2018) algorithm.[14] With this

[14]In Appendix E we include additional Monte Carlo results comparing the use of 500 and 1000 values of v when $K = 5$, and we find that the difference is quite small. Moreover the relationship between the power curves for these two cases is not monotone. While using an exhaustive collection of v from the cells induced by hyperplane arrangement provides a complete moment inequality characterization of $\mathcal{B}_n^*$, it is not clear that using more values of v is optimal from a power perspective, due to sampling variation.

choice our inference method performed well in comparison to the other inference approaches considered. Nonetheless, the development of incremental enumeration algorithms remains an active area of study, and future innovation in this dimension may speed up the computation of our test statistic when employing more cells.[15]

3 Inference Based on Moment Inequalities

For a given value $b \in \mathcal{B}$, we consider a hypothesis test of

$$H_0 : \beta = b \quad \text{versus} \quad H_1 : \beta \neq b, \tag{3.1}$$

on the basis of n observations $\{(Y_i, X_i) : i = 1, ..., n\}$ following the restrictions of the semi-parametric binary response model given by Assumption 1. In this section we first provide theoretical guarantees for our test, followed by a step-by-step guide for implementation. If one wishes to construct a confidence set for β, the set of b for which H_0 is not rejected by a size α test will provide a confidence set guaranteed to contain β with probability at least $1 - \alpha$. We come back to this point in our empirical application, in which we also consider the problem of conducting marginal inference on individual components of β from a computational standpoint.

3.1 Hypothesis Test

To perform inference based on moment inequalities in Theorem 3, we incorporate sample analogs of the moments appearing in (2.9) and (2.10), which are

$$\begin{aligned} \hat{m}_u(b, v) &\equiv \mathbb{E}_n \left[(2Y - 1) 1\{Xb \geq 0 > Xv\} \right], \quad v \in \mathcal{V}_u, \\ \hat{m}_l(b, v) &\equiv \mathbb{E}_n \left[(1 - 2Y) 1\{Xb \leq 0 < Xv\} \right], \quad v \in \mathcal{V}_l, \end{aligned}$$

into our test statistic

$$T_n(b) \equiv \max \left\{ 0, \sup_{v \in \mathcal{V}_u} \sqrt{n} \left(-\frac{\hat{m}_u(b, v)}{\widehat{\sigma}_u(b, v)} \right), \sup_{v \in \mathcal{V}_l} \sqrt{n} \left(-\frac{\hat{m}_l(b, v)}{\widehat{\sigma}_l(b, v)} \right) \right\}, \tag{3.2}$$

[15] For example, the contemporaneous working paper Gu, Russell, and Stringham (2022) proposes a novel incremental enumeration algorithm that replaces the use of linear programs to find new representative points in each iteration. Instead, for this step they invoke recursion in the dimension of the space under study, which avoids the need to repeatedly solve linear programs.

where:[16]

$$\widehat{\sigma}_u(b,v) \equiv \sqrt{\mathbb{E}_n\left[1\{Xb \geq 0 > Xv\}\right] - \hat{m}_u(b,v)^2}, \tag{3.3}$$

$$\widehat{\sigma}_l(b,v) \equiv \sqrt{\mathbb{E}_n\left[1\{Xb \leq 0 < Xv\}\right] - \hat{m}_l(b,v)^2}. \tag{3.4}$$

The finite sample distribution of $T_n(b)$ under H_0 is unknown. We construct a random variable $T_n^*(b)$ which has a known finite sample distribution given $\mathcal{X}_n$ and which satisfies

$$T_n(b) \leq T_n^*(b) \text{ under } H_0 : \beta = b. \tag{3.5}$$

To this purpose define $(Y_1^*, ..., Y_n^*)$ by

$$Y_i^* = 1\{U_i \geq 0\}, \quad i = 1, ..., n,$$

and define

$$T_n^*(b) \equiv \max\left\{0, \sup_{v \in \mathcal{V}_u} \sqrt{n}\left(-\frac{\hat{m}_u^*(b,v)}{\widehat{\sigma}_u^*(b,v)}\right), \sup_{v \in \mathcal{V}_l} \sqrt{n}\left(-\frac{\hat{m}_l^*(b,v)}{\widehat{\sigma}_l^*(b,v)}\right)\right\}$$

with

$$\begin{aligned}
\hat{m}_u^*(b,v) &\equiv \mathbb{E}_n\left[(2Y^* - 1)1\{Xb \geq 0 > Xv\}\right] \\
\hat{m}_l^*(b,v) &\equiv \mathbb{E}_n\left[(1 - 2Y^*)1\{Xb \leq 0 < Xv\}\right] \\
\widehat{\sigma}_u^*(b,v) &\equiv \sqrt{\mathbb{E}_n\left[1\{Xb \geq 0 > Xv\}\right] - \hat{m}_u^*(b,v)^2}, \\
\widehat{\sigma}_l^*(b,v) &\equiv \sqrt{\mathbb{E}_n\left[1\{Xb \leq 0 < Xv\}\right] - \hat{m}_l^*(b,v)^2}.
\end{aligned}$$

The random variable $T_n^*(b)$ replaces Y with Y^* in $T_n(b)$ defined by (3.2) – (3.4). We do not observe $T_n^*(b)$ because $(Y_1^*, ..., Y_n^*)$ are not observed, but the finite sample distribution of $T_n^*(b)$ given $\mathcal{X}_n$ is known since $(Y_1^*, ..., Y_n^*)$ are independent Bernoulli(1/2) random variables conditional on $\mathcal{X}_n$.

Thus, for a given level $\alpha \in (0,1)$, the critical value used for our test is the conditional $1-\alpha$ quantile of $T_n^*(b)$ given $\mathcal{X}_n$, namely

$$q_{1-\alpha}(b) \equiv \inf\{c \in \mathbb{R} : \mathbb{P}\left(T_n^*(b) \leq c \mid \mathcal{X}_n\right) \geq 1 - \alpha\}. \tag{3.6}$$

This critical value can be computed up to arbitrary accuracy by drawing a large number of

[16] In (3.2) it is to be understood that when $\widehat{\sigma}_c(b,v) = 0$ then $-\hat{m}_c(b,v)/\widehat{\sigma}_c(b,v) = 0$ if $\hat{m}_c(b,v) = 0$, while otherwise $-\hat{m}_c(b,v)/\widehat{\sigma}_c(b,v) = \pm\infty$ with the sign that of $-\hat{m}_c(b,v)$.

simulations, each of which comprises a sequence of n independent Bernoulli random variables. Our proposed test uses the rejection rule $1\{T_n(b) > q_{1-\alpha}(b)\}$, that is, it rejects H_0 in favor of H_1 iff $T_n(b) > q_{1-\alpha}(b)$.

The relationship between $T_n(b)$ and $T_n^*(b)$ in (3.5) implies Theorem 4, establishing finite sample size control of the proposed test. As is the case with all formal mathematical results stated in the paper, the proofs of inequality (3.5) and Theorem 4 are in the Appendix.

Theorem 4. *If Assumption 1 holds, then*

$$\mathbb{P}\left(T_n(b) \leq q_{1-\alpha}(b) \mid \mathcal{X}_n\right) \geq \inf_{G \in \mathcal{G}} P_{(\beta,G)}(T_n(b) \leq q_{1-\alpha}(b) \mid \mathcal{X}_n) \geq 1-\alpha \text{ under } H_0 : \beta = b,$$

and for any value $\tilde{c} < q_{1-\alpha}(b)$ that is fixed conditional on $\mathcal{X}_n$,

$$\inf_{G \in \mathcal{G}} P_{(\beta,G)}\left(T_n(b) \leq \tilde{c} \mid \mathcal{X}_n\right) < 1-\alpha \text{ under } H_0 : \beta = b.$$

Theorem 4 establishes finite sample size control of the rejection rule $1\{T_n(b) > q_{1-\alpha}(b)\}$ for hypothesis test (3.1). While it is possible that $\mathbb{P}\left(T_n(b) \leq q_{1-\alpha}(b) \mid \mathcal{X}_n\right)$ strictly exceeds $1-\alpha$ under H_0, any test using the same test statistic $T_n(b)$ with a smaller critical value $\tilde{c}$ does not establish finite sample size control.[17]

Theorem 5 next establishes a power result for our test as a function of a measure of the violation of moment inequalities that define $\mathcal{B}_n^*$. Specifically, Hoeffding's inequality is used to establish a lower bound on finite sample power for certain violations of the inequalities (2.9) and (2.10).

Theorem 5. *Let Assumptions 1 and 2 hold, and let ρ be any number in $(0,1)$. If there is $v \in \mathcal{V}_u$ such that*

$$\begin{aligned}&\mathbb{E}\left[\, \mathbb{E}_n\left[(2Y-1)1\{Xb \geq 0 > Xv\}\right] \mid \mathcal{X}_n\right] \\ &\quad \leq -\frac{1}{\sqrt{n}}\left(q_{1-\alpha}(b)\sqrt{\frac{\mathbb{E}_n[1\{Xb \geq 0 > Xv\}]}{1+q_{1-\alpha}(b)^2/n}} + \sqrt{2\log(1/\rho)\mathbb{E}_n[1\{Xb \geq 0 > Xv\}]}\right), \quad (3.7)\end{aligned}$$

[17] Our proposed critical value $q_{1-\alpha}(b)$ and the value $\tilde{c}$ in Theorem 4 depend only on $\mathcal{X}_n$. In this paper, we focus on *fixed* critical values, where here because inference is conditional on $\mathcal{X}_n$ we mean fixed conditional on $\mathcal{X}_n$. Also, note that Theorem 4 is silent about different choices of the *test statistic*.

or there is $v \in \mathcal{V}_l$ *such that*

$$\begin{aligned}&\mathbb{E}\left[\ \mathbb{E}_n\left[(1-2Y)1\{Xb \leq 0 < Xv\}\right] \mid \mathcal{X}_n\right] \\ &\quad \leq -\frac{1}{\sqrt{n}}\left(q_{1-\alpha}(b)\sqrt{\frac{\mathbb{E}_n[1\{Xb \leq 0 < Xv\}]}{1+q_{1-\alpha}(b)^2/n}} + \sqrt{2\log(1/\rho)\mathbb{E}_n[1\{Xb \leq 0 < Xv\}]}\right), \quad (3.8)\end{aligned}$$

then the rejection probability is at least $1-\rho$*, i.e.,* $\mathbb{P}(T_n(b) > q_{1-\alpha}(b) \mid \mathcal{X}_n) \geq 1-\rho$*.*

Furthermore, inversion of the conditions in Theorem 5 provides a power guarantee for any parameter value $b \notin \mathcal{B}_n^*$.

Corollary 1. *Let Assumptions 1 and 2 hold. For any* $b \notin \mathcal{B}_n^*$*, the rejection probability* $\mathbb{P}(T_n(b) > q_{1-\alpha}(b) \mid \mathcal{X}_n)$ *is at least the maximum of the following two expressions:*

$$\max_{v \in \mathcal{V}_u}\left(1-\exp\left(-\frac{1}{2}\left(\max\left\{0, \sqrt{n}\zeta_u(b,v) - q_{1-\alpha}(b)(1+q_{1-\alpha}(b)^2/n)^{-1/2}\right\}\right)^2\right)\right), \quad (3.9)$$

$$\max_{v \in \mathcal{V}_l}\left(1-\exp\left(-\frac{1}{2}\left(\max\left\{0, \sqrt{n}\zeta_l(b,v) - q_{1-\alpha}(b)(1+q_{1-\alpha}(b)^2/n)^{-1/2}\right\}\right)^2\right)\right), \quad (3.10)$$

where the quantities in the above expressions are defined as

$$\zeta_u(b,v) \equiv \frac{-\mathbb{E}\left[\ \mathbb{E}_n\left[(2Y-1)1\{Xb \geq 0 > Xv\}\right] \mid \mathcal{X}_n\right]}{\sqrt{\mathbb{E}_n[1\{Xb \geq 0 > Xv\}]}}$$

$$\zeta_u(b,v) \equiv \frac{-\mathbb{E}\left[\ \mathbb{E}_n\left[(1-2Y)1\{Xb \leq 0 < Xv\}\right] \mid \mathcal{X}_n\right]}{\sqrt{\mathbb{E}_n[1\{Xb \leq 0 < Xv\}]}}.$$

The bound provided by Theorem 5 depends on the degree to which the inequalities (2.9) and (2.10) that characterize $\mathcal{B}_n^*$ are violated relative to $\sqrt{\mathbb{E}_n[1\{Xb \geq 0 > Xv\}]}$ and $\sqrt{\mathbb{E}_n[1\{Xb \leq 0 < Xv\}]}$. However, Theorem 5 further implies an explicit mapping between (i) the extent to which a given parameter vector b violates the inequalities that define the set of conditionally observationally equivalent parameters and (ii) a lower bound on the finite sample power of our test for $\beta = b$ that does not depend on sample quantities. To see this, define

$$\begin{aligned}Q_u(b) &\equiv -\min\left\{0, \min_{v \in \mathcal{V}_u}\mathbb{E}\left[\ \mathbb{E}_n\left[(2Y-1)1\{Xb \geq 0 > Xv\}\right] \mid \mathcal{X}_n\right]\right\}, \\ Q_l(b) &\equiv -\min\left\{0, \min_{v \in \mathcal{V}_l}\mathbb{E}\left[\ \mathbb{E}_n\left[(1-2Y)1\{Xb \leq 0 < Xv\}\right] \mid \mathcal{X}_n\right]\right\}.\end{aligned}$$

The values of $Q_u(b)$ and $Q_l(b)$ denote the maximal violation exhibited by b of the inequalities (2.9) and (2.10) that characterize $\mathcal{B}_n^*$ in Theorem 3. Theorem 5 implies that our test is

guaranteed to reject false hypotheses $\beta = b$ with probability at least $1 - \rho$ whenever the measure of violation, $\max\{Q_u(b), Q_l(b)\}$, is at least $C_\alpha(b, 1-\rho)$, defined by

$$C_\alpha(b, 1-\rho) \equiv \frac{1}{\sqrt{n}} \left(q_{1-\alpha}(b)(1 + q_{1-\alpha}(b)^2/n)^{-1/2} + \sqrt{-2\log(\rho)} \right). \tag{3.11}$$

Inversion of this relation also provides an explicit power guarantee as a function of $\max\{Q_u(b), Q_l(b)\}$. The following corollary to Theorem 5 gives the formal results.

Corollary 2. *Let Assumptions 1 and 2 hold and let $\rho \in (0, 1)$.*

1. *If $\max\{Q_u(b), Q_l(b)\} \geq C_\alpha(b, 1-\rho)$, then $\mathbb{P}(T_n(b) > q_{1-\alpha}(b) \mid \mathcal{X}_n) \geq 1 - \rho$.*
2. *For any $b \notin \mathcal{B}_n^*$, the rejection probability $\mathbb{P}(T_n(b) > q_{1-\alpha}(b) \mid \mathcal{X}_n)$ is at least*

$$1 - \exp\left(-\frac{1}{2} \left(\max\left\{0, \sqrt{n}\max\{Q_u(b), Q_l(b)\} - q_{1-\alpha}(b)(1 + q_{1-\alpha}(b)^2/n)^{-1/2} \right\} \right)^2 \right).$$

Corollary 2 can be used to indicate how big $\max\{Q_u(b), Q_l(b)\}$ must be in order for Theorem 5 to guarantee our test has power at least α against a parameter value $\tilde{b}$ irrespective of sample quantities. It should be noted however that the power bounds delivered by Theorem 5 provide power guarantees which may not be sharp. That is, the test may achieve higher power than this bound guarantees.

3.2 Implementation

The following steps describe how to perform inference using test statistic $T_n(b)$ with the critical value $q_{1-\alpha}(b)$ described above.

1. Compute $\mathcal{V}_u$ and $\mathcal{V}_l$ to use in $T_n(b)$.
2. Compute $T_n(b)$ as defined in (3.2) - (3.4).
3. Draw r samples of n i.i.d. Bernoulli(1/2) variables $Y_1^*, ..., Y_n^*$.
4. Compute $T_n^*(b)$ in each sample from step 3 and set $q_{1-\alpha}(b)$ to the $1-\alpha$ quantile of $T_n^*(b)$ in these r samples.
5. Reject the null hypothesis $\beta = b$ of (3.1) if $T_n(b) > q_{1-\alpha}(b)$, otherwise do not reject.

Step one can be implemented in several different ways. For instance, it can be done by exhaustively computing representatives from $\mathsf{V}(\mathcal{X}_n)$ using incremental cell enumeration algorithms for hyperplane arrangements such as those of Rada and Černý (2018) and Gu, Russell,

and Stringham (2022) as described in Section 2.3. Alternatively, one can randomly select points in $\mathcal{B}$. Exhaustively computing all representatives can be computationally demanding depending on n and K, so imposing a maximal number or randomly selecting points can help to maintain computational feasibility even when the number of elements of $\mathsf{V}(\mathcal{X}_n)$ is prohibitive.

Steps three and four are used to approximate the critical value $q_{1-\alpha}(b)$, the conditional $1-\alpha$ quantile of $T_n^*(b)$ given $\mathcal{X}_n$, defined in (3.6). It can be computed up to arbitrary accuracy by choosing a sufficiently high value of r. If the test is carried out for multiple null values of b, then the same r samples of $(Y_1^*, ..., Y_n^*)$ can be used for each one.

4 Likelihood Ratios, Minimax Tests, and Maximum Score

In this section, we consider the use of likelihood ratio tests. In subsections 4.1–4.3 we consider their performance among least favorable configurations for the sake of determining minimax optimal tests under a sequence of null and alternative hypotheses, the key distinction between subsections being whether each hypothesis is simple or composite. When considering the minimax optimality of the likelihood ratio test, the relevant consideration is the set of conditional distributions of $(Y_1, ..., Y_n)$ given $\mathcal{X}_n$ allowed under each hypothesis. A point null hypothesis for β is a composite hypothesis because it does not specify the conditional distribution of $(U_1, ..., U_n)$ given $\mathcal{X}_n$, and therefore admits many different possible conditional distributions of $(Y_1, ..., Y_n)$ given $\mathcal{X}_n$ even under the maintained assumption that for all i, $\mathbb{P}(U_i \geq 0 \mid \mathcal{X}_n) = 1/2$. We therefore begin by first considering cases in which the null and alternative hypotheses each uniquely specify both β and G_0, and hence uniquely determine the distribution $P_{(\beta,G_0)}$, before then moving on to consideration of composite hypotheses. In Section 4.4 we then move beyond analysis under least favorable configurations and instead consider likelihood ratio tests that treat the distribution of unobservable heterogeneity as a nuisance parameter over which to optimize subject to the restriction of the hypothesized parameter value. Under standard conditions the profile log-likelihood is shown to be a monotone transformation of the maximum score objective function.

4.1 Simple Null Hypotheses and Simple Alternative Hypotheses

When both hypotheses are simple, they have the representation

$$H_0 : (\beta, G_0) = (b, G) \quad \text{versus} \quad H_1 : (\beta, G_0) = \left(\tilde{b}, \tilde{G}\right). \tag{4.1}$$

The likelihood under the null hypothesis is

$$P_{(b,G)}(Y_i = y_i \text{ for all } i \mid \mathcal{X}_n) = G(1\{X_i b + U_i \geq 0\} = y_i \text{ for all } i \mid \mathcal{X}_n), \tag{4.2}$$

and the likelihood under the alternative hypothesis is

$$P_{(\tilde{b},\tilde{G})}(Y_i = y_i \text{ for all } i \mid \mathcal{X}_n) = \tilde{G}(1\{X_i \tilde{b} + U_i \geq 0\} = y_i \text{ for all } i \mid \mathcal{X}_n).$$

The likelihood ratio test ϕ_{LR} is the test that rejects H_0 in favor of H_1 with probability $\phi_{LR}(Y_1, ..., Y_n; \mathcal{X}_n)$ defined by

$$\phi_{LR}(y_1, ..., y_n; \mathcal{X}_n) = \begin{cases} 1 & \text{if } P_{(\tilde{b},\tilde{G})}(Y_i = y_i \text{ for all } i \mid \mathcal{X}_n) > kP_{(b,G)}(Y_i = y_i \text{ for all } i \mid \mathcal{X}_n) \\ \xi & \text{if } P_{(\tilde{b},\tilde{G})}(Y_i = y_i \text{ for all } i \mid \mathcal{X}_n) = kP_{(b,G)}(Y_i = y_i \text{ for all } i \mid \mathcal{X}_n) \\ 0 & \text{otherwise,} \end{cases}$$

where k and $\xi \in [0, 1]$ are chosen such that $\sum_{(y_1,...,y_n)\in\{0,1\}^n} \phi_{LR}(y_1, ..., y_n; \mathcal{X}_n) P_{(b,G)}(Y_i = y_i \text{ for all } i \mid \mathcal{X}_n) = \alpha$. [18]

When the null and alternative distributions of a hypothesis test are simple, then by the Neyman-Pearson Lemma (e.g., Chapter 3.2 of Lehmann and Romano (2005)), the likelihood ratio test is the uniformly most powerful test. However, the hypothesis test of interest, (3.1), differs because neither the null or alternative hypotheses uniquely specify G_0.

If Assumption 2 additionally holds, the likelihood simplifies to the product of individual likelihoods. Define

$$\bar{p}_i \equiv G\left(U_i \geq -X_i b \mid \mathcal{X}_n\right), \tag{4.3}$$

and

$$\tilde{p}_i \equiv \tilde{G}\left(U_i \geq -X_i \tilde{b} \mid \mathcal{X}_n\right), \tag{4.4}$$

[18] This test uses randomization in the event that $P_{(\tilde{b},\tilde{G})}(Y_i = y_i \text{ for all } i \mid \mathcal{X}_n) = kP_{(b,G)}(Y_i = y_i \text{ for all } i \mid \mathcal{X}_n)$. If a non-randomized implementation of the likelihood ratio test is preferred, that can be done in the usual way by making the modification

$$\phi_{LR}(y_1, ..., y_n; \mathcal{X}_n) = \begin{cases} 1 & \text{if } P_{(\tilde{b},\tilde{G})}(Y_i = y_i \text{ for all } i \mid \mathcal{X}_n) > kP_{(b,G)}(Y_i = y_i \text{ for all } i \mid \mathcal{X}_n) \\ 0 & \text{if } P_{(\tilde{b},\tilde{G})}(Y_i = y_i \text{ for all } i \mid \mathcal{X}_n) = kP_{(b,G)}(Y_i = y_i \text{ for all } i \mid \mathcal{X}_n) \\ 0 & \text{otherwise,} \end{cases}$$

with k chosen as small as possible subject to the constraint that $\sum_{(y_1,...,y_n)\in\{0,1\}^n} \phi_{LR}(y_1, ..., y_n; \mathcal{X}_n) P_{(b,G)}(Y_i = y_i \text{ for all } i \mid \mathcal{X}_n) \leq \alpha$. The conclusion of Theorem 6 would then be that the likelihood ratio test is a most powerful non-randomized test of (4.5) subject to achieving finite sample size control. Subsequent results could be similarly modified without substantive change of conclusions.

The test $\phi_{LR}(Y_1, ..., Y_n; \mathcal{X}_n)$ can then be written as

$$\phi_{LR}(Y_1, ..., Y_n; \mathcal{X}_n) = \begin{cases} 1 & \text{if } \prod_{i=1}^n \tilde{p}_i^{Y_i}(1-\tilde{p}_i)^{1-Y_i} > k \prod_{i=1}^n \bar{p}_i^{Y_i}(1-\bar{p}_i)^{1-Y_i}, \\ \xi & \text{if } \prod_{i=1}^n \tilde{p}_i^{Y_i}(1-\tilde{p}_i)^{1-Y_i} = k \prod_{i=1}^n \bar{p}_i^{Y_i}(1-\bar{p}_i)^{1-Y_i}, \\ 0 & \text{otherwise,} \end{cases}$$

where k and $\xi \in [0, 1]$ are chosen such that $\sum_{(y_1,...,y_n)\in\{0,1\}^n} \phi_{LR}(y_1, ..., y_n; \mathcal{X}_n) \prod_{i=1}^n \bar{p}_i^{y_i}(1 - \bar{p}_i)^{1-y_i} = \alpha$.

4.2 Composite Null Hypothesis and Simple Alternative Hypothesis

In a step closer to the two-sided hypothesis test (3.1), consider the hypothesis test

$$H_0 : \beta = b \quad \text{versus} \quad H_1 : (\beta, G_0) = \left(\tilde{b}, \tilde{G}\right), \tag{4.5}$$

which comprises the same null hypothesis of (3.1) and the alternative hypothesis of (4.1). This test features a composite null hypothesis and a simple alternative hypothesis for (β, G_0).

Using Theorem 3.8.1 of Lehmann and Romano (2005), we can reduce H_0 to a simple hypothesis by finding the least favorable distribution of $(Y_1, ..., Y_n)$ given $\mathcal{X}_n$. This entails pairing the null value $\beta = b$ with the distribution G_0 among those satisfying the restrictions maintained in Assumptions 1 and 2 for which the likelihood ratio test has minimal power against H_1. Using the independence restriction of Assumption 2, the distributions of $(Y_1, ..., Y_n)$ given $\mathcal{X}_n$ are given by the marginal Bernoulli probabilities for each i, so that the hypotheses can be characterized by the implied collections of probabilities $(\bar{p}_1, ..., \bar{p}_n)$ and $(\tilde{p}_1, ..., \tilde{p}_n)$ coincident with (4.3) and (4.4) under the null and alternative, respectively. Thus the null hypothesis in (4.5) can be written as

$$H_0 : (\beta, G_0) = \left(b, \prod_{i=1}^n G_i\right) \text{ for some } (G_1, ..., G_n).$$

For each observation i consider the probabilities $\bar{p}_i$ and $\tilde{p}_i$. Note that under Assumptions 1 and 2, $\bar{p}_i - 1/2$ is nonpositive (nonnegative) if $X_i b$ is nonpositive (nonnegative), and likewise $\tilde{p}_i - 1/2$ is nonpositive (nonnegative) if $X_i\tilde{b}$ is nonpositive (nonnegative). When $X_i b$ and $\tilde{p}_i - 1/2$ have the same sign, then there exists a conditional distribution of unobservable U_i given $\mathcal{X}_n$ such that $\bar{p}_i$ can be made equal to $\tilde{p}_i$. For such i, the least favorable distribution will thus have $\bar{p}_i = \tilde{p}_i$. When instead $X_i b$ and $\tilde{p}_i - 1/2$ have the opposite sign, then in general

$\bar{p}_i$ cannot be equal to $\tilde{p}_i$ under the null due to the requirement in Assumption 1(iii) that $\mathbb{P}(U_i \geq 0 \mid \mathcal{X}_n) = 1/2$. The least favorable probability $\bar{p}_i$ will however be made as close as possible to $\tilde{p}_i$ while adhering to this assumption if the conditional distribution of U_i given $\mathcal{X}_n$ allocates all mass to regions in which $|U_i| > |X_i b|$, so that $\bar{p}_i = 1/2$.

We now formally let $\bar{p}_i$ denote the null probability of $Y_i = 1$ given $\mathcal{X}_n$ in keeping with the above intuition for constructing the least favorable distribution under the composite null $\beta = b$ as follows:

$$\bar{p}_i = \begin{cases} 1/2 & \text{if } X_i b(\tilde{p}_i - 1/2) < 0 \text{ or } X_i b = 0, \\ \tilde{p}_i & \text{otherwise.} \end{cases} \tag{4.6}$$

These probabilities lie within the null set in (4.5) because $\bar{p}_i = 1/2$ is possible for any value of b, and $\bar{p}_i = \tilde{p}_i$ is possible when $\tilde{p}_i - 1/2$ and $X_i b$ are either both positive or negative. By the Neyman-Pearson Lemma, the likelihood ratio test for the simple null of $\bar{p}_1, ..., \bar{p}_n$ against $\tilde{p}_1, ..., \tilde{p}_n$ is most powerful. Theorem 6 verifies that the likelihood ratio test for this simple hypothesis achieves finite sample size control under the composite null $\beta = b$, thus enabling application of Theorem 3.8.1 of Lehmann and Romano (2005) and establishing that these null probabilities are least favorable.

To proceed consider the corresponding test $\bar{\phi}_{LR}$ for (4.5) that makes use of this configuration of implied probabilities under the null. In other words, we consider the likelihood ratio test with

$$\bar{\phi}_{LR}(Y_1, ..., Y_n; \mathcal{X}_n) = \begin{cases} 1 & \text{if } \prod_{i=1}^n \tilde{p}_i^{Y_i}(1-\tilde{p}_i)^{1-Y_i} > k \prod_{i=1}^n \bar{p}_i^{Y_i}(1-\bar{p}_i)^{1-Y_i}, \\ \xi & \text{if } \prod_{i=1}^n \tilde{p}_i^{Y_i}(1-\tilde{p}_i)^{1-Y_i} = k \prod_{i=1}^n \bar{p}_i^{Y_i}(1-\bar{p}_i)^{1-Y_i}, \\ 0 & \text{otherwise,} \end{cases} \tag{4.7}$$

where k and $\xi \in [0, 1]$ are chosen such that

$$\sum_{(y_1,...,y_n) \in \{0,1\}^n} \bar{\phi}_{LR}(y_1, ..., y_n; \mathcal{X}_n) \prod_{i=1}^n \bar{p}_i^{y_i}(1-\bar{p}_i)^{1-y_i} = \alpha. \tag{4.8}$$

There is the following result, leveraging Theorem 3.8.1 of Lehmann and Romano (2005).

Theorem 6. *Let Assumptions 1 and 2 hold. Consider the null and alternative hypotheses stated in* (4.5) *and the test* $\bar{\phi}_{LR}$ *defined in* (4.7) *and* (4.8)*. If the null hypothesis is true then the rejection probability* $\mathbb{E}\left[\bar{\phi}_{LR}(Y_1, ..., Y_n) \mid \mathcal{X}_n\right]$ *is no greater than* α*. Moreover,*

$$(y_1, ..., y_n) \mapsto \prod_{i=1}^n \bar{p}_i^{y_i}(1-\bar{p}_i)^{1-y_i} \tag{4.9}$$

is the least favorable distribution of $(Y_1, ..., Y_n)$ *given* $\mathcal{X}_n$ *under* H_0 *against* H_1*, and the test* $\bar{\phi}_{LR}$ *is a most powerful test of* H_0 *against* H_1*.*

Moreover, because the likelihood ratio test (4.7) is a most powerful test for hypothesis (4.7) in which the alternative hypothesis specifies a particular distribution G in addition to $\tilde{b}$, it provides a (possibly unattainable) power envelope for the test (3.1) which does not specify G_0 under the alternative hypothesis.

4.3 Composite Null Hypothesis and Composite Alternative Hypothesis

When the researcher's goal is to test $\beta = b$ against the simple alternative hypothesis in (4.5) that completely specifies the conditional distribution of $(Y_1, ..., Y_n)$ given $\mathcal{X}_n$, then the likelihood ratio test implemented by adopting the rejection probability specified in (4.6) – (4.8) is most powerful. When instead the researcher wishes to control power against a composite alternative hypothesis, such as that of

$$H_0 : \beta = b \quad \text{versus} \quad H_1 : \beta = \tilde{b}, \tag{4.10}$$

Theorem 6 is silent because each conditional distribution of $(Y_1, ..., Y_n)$ given $\mathcal{X}_n$ allowed under the alternative hypothesis will result in a different least favorable configuration and thus a different likelihood ratio test. However, following arguments in Chapter 8.1 of Lehmann and Romano (2005), it is straightforward to construct the least favorable pair of distributions for this test. Note that for *any* $\tilde{b}$ hypothesized under the alternative, distributions $\tilde{G}_i$ can be specified such that for all i:

$$\tilde{p}_i \equiv \tilde{G}\left(U_i \geq -X_i\tilde{b} \mid \mathcal{X}_n\right) = 1/2. \tag{4.11}$$

In particular, this is achieved by allocating all mass to regions on which $|U_i| \geq |X_i\tilde{b}|$, while obeying the constraint $\mathbb{P}(U_i \geq 0 \mid \mathcal{X}_n) = 1/2$.[19] Such a combination of $\left(\tilde{b}, \tilde{G}\right)$ under the alternative hypothesis yields $\tilde{p}_i = 1/2$ for all i. This conclusion holds irrespective of the hypothesized value of $\tilde{b}$ in (4.5). Indeed it also holds for the value of $\beta = b$ hypothesized under the null. Correspondingly the least favorable $\{\bar{p}_i : i = 1, ..., n\}$ under the null given by (4.6) when all $\tilde{p}_i = 1/2$ is given by $\bar{p}_i = 1/2$ for all i. Thus from Theorem 6 we have the following implications for the two-sided test of $\beta = b$ against $\beta = \tilde{b}$.

[19] If the distribution of U_i were restricted to have positive density on $\mathbb{R}$, then G_i could be specified to allocate probability $1 - \epsilon$ to $\{u_i : |u_i| \geq |X_i b'|\}$ for any small $\epsilon > 0$.

Corollary 3. *Let Assumptions 1 and 2 hold. The least favorable pair of* $\{\bar{p}_i : i = 1, ..., n\}$ *and* $\{\tilde{p}_i : i = 1, ..., n\}$ *for the test* (4.5) *is given by* $\bar{p}_i = 1/2$ *for all* i *and* $\tilde{p}_i = 1/2$ *for all* i*. Moreover, since the conclusion holds for any* b *and* $\tilde{b}$*, this is also the least favorable pair for the two-sided hypothesis test of* $\beta = b$ *in* (3.1)*.*

This corollary is a direct implication of Theorem 6. While it may be conceptually appealing to construct a minimax optimal likelihood ratio test for the two-sided test (4.10) based on the least favorable pair, we see from (4.7) and (4.8) that this results in a test which rejects the null hypothesis with probability α irrespective of the data. This is because the hypothesis $\beta = b$ always includes the distribution under which $\bar{p}_i = 1/2$ for all i, for any hypothesized value of b. Thus the sets of feasible conditional distributions of $(Y_1, ..., Y_n)$ given $\mathcal{X}_n$ compatible with each of the two hypotheses, $\beta = b$ and $\beta = \tilde{b}$ overlap. Therefore, parameter values b and $\tilde{b}$ are potentially observationally equivalent, as defined in Chesher and Rosen (2017), even if $\tilde{b}$ lies outside the set of conditionally observationally equivalent parameters, and hypothesis test (4.10) is not robustly testable in the sense of Kaido and Zhang (2019). In words, under both hypotheses there exists a specification for the nuisance distribution that produces $\bar{p}_i = 1/2$ for all i.[20] Under this conditional distribution of Y_i given $\mathcal{X}_n$ the hypotheses are indistinguishable, and a test that rejects with probability α independent of the data is then most powerful.

4.4 Profile Log-Likelihood and Maximum Score

In this section we shift attention from characterizing minimax optimal tests to the study of the likelihood ratio test statistic, which comprises the ratio of profile log-likelihoods restricted by the null and alternative hypotheses. When considering the null hypothesis $\beta = b$, the profile log-likelihood imposes $\beta = b$ and achieves the maximal log-likelihood over all possible values of the nuisance parameter G_0.[21] Namely, the profile log-likelihood is given by $\ell_{(Y_1,...,Y_n)}(b)$, where

$$\ell_{(y_1,...,y_n)}(b) \equiv \sup_{G \in \mathcal{G}} \log G(1\{X_i b + U_i \geq 0\} = y_i \text{ for all } i \mid \mathcal{X}_n). \tag{4.12}$$

The function $\ell_{(Y_1,...,Y_n)}(b)$ has a closed-form expression as given in the next theorem.

[20] In fact, there exists distributions of $(U_1, ..., U_n)$ given $\mathcal{X}_n$ that produce $\bar{p}_i = 1/2$ for all i irrespective of whether $\beta = b$ or $\beta = \tilde{b}$, namely those distributions that allocate all mass to regions on which $|U_i| \geq \max\left\{|X_i b|, |X_i \tilde{b}|\right\}$ while also satisfying $\mathbb{P}(U_i \geq 0 \mid \mathcal{X}_n) = 1/2$.

[21] Recall that G_0 denotes the *joint* distribution of $U_1, ..., U_n$ conditional on $\mathcal{X}_n$. Our analysis thus differs from results of Cosslett (1983) and Manski and Thompson (1989) providing comparisons of maximum score estimation of β to maximum likelihood estimators constructed under the additional assumptions that $U_1, ..., U_n$ are identically and independently distributed with U_i and X_i independent for all i.

Theorem 7. *Let Assumption 1 hold. Then:*

$$\ell_{(Y_1,...,Y_n)}(b) = \frac{n\log(2)}{2}(\tilde{S}_n(b) - 1), \tag{4.13}$$

where

$$\tilde{S}_n(b) \equiv \frac{1}{n}\sum_{i=1}^{n}(1\{(2Y_i - 1)X_ib > 0\} - 1\{(2Y_i - 1)X_ib \leq 0\}).$$

The function $\tilde{S}_n(b)$ closely resembles Manski's (1985) maximum score objective function:

$$S_n(b) \equiv \frac{1}{n}\sum_{i=1}^{n}(2Y_i - 1)\text{sgn}(X_ib) \text{ with } \text{sgn}(X_ib) \equiv 1\{X_ib \geq 0\} - 1\{X_ib < 0\}.$$

Since

$$S_n(b) = \tilde{S}_n(b) + \frac{2}{n}\sum_{i=1}^{n}1\{X_ib = 0, Y_i = 1\},$$

the difference between $\tilde{S}_n(b)$and $S_n(b)$ lies *only* in how each function treats observations with $X_ib = 0$.[22] Note however that under Manski's conditions for point identification X_ib is continuously distributed for any b, so that $X_ib = 0$ occurs with zero probability. Thus we have the following Corollary.

Corollary 4. *Suppose that* $\mathbb{P}(X_ib = 0) = 0$ *for all* i. *Then* $\underset{b\in\mathcal{B}}{\text{argmax}}\, S_n(b) = \underset{b\in\mathcal{B}}{\text{argmax}}\, \ell_{(Y_1,...,Y_n)}(b)$ *almost surely.*

Evidently, under the usual conditions imposed for point identification in the literature, an estimator that maximizes the profile log-likelihood $\ell_{(Y_1,...,Y_n)}(b)$ is a maximum score estimator, and *vice versa*. Moreover the likelihood ratio statistic is

$$LR(b) = 2\left(\ell_{(Y_1,...,Y_n)}(\widehat{\beta}_{MS}) - \ell_{(Y_1,...,Y_n)}(b)\right) = 2n\log(2)\left(\tilde{S}_n(\widehat{\beta}_{MS}) - \tilde{S}_n(b)\right), \tag{4.14}$$

where $\widehat{\beta}_{MS}$ is a maximum score estimator, i.e. any element of the set maximizers of $\tilde{S}_n(b)$. Note that for any $b \neq \beta$ there exists a distribution G of $(U_1, ..., U_n)$ given $\mathcal{X}_n$ such that $U_1, ..., U_n$ are independent given $\mathcal{X}_n$ and that

$$G\left(U_i < \max_i \max\left(|X_i\beta|, |X_ib|\right) \mid \mathcal{X}_n\right) = G\left(U_i > \max_i \max\left(|X_i\beta|, |X_ib|\right) \mid \mathcal{X}_n\right) = \frac{1}{2}.$$

[22]Indeed Manski (1985) combines the two implications of $Med(U|X) = 0$: (1) $x\beta > 0 \iff E[2Y - 1|X = x] > 0$ and (2) $x\beta = 0 \iff E[2Y - 1|X = x] = 0$ to $x\beta \geq 0 \iff E[2Y - 1|X = x] \geq 0$. He notes that the former implication is stronger, but the latter is more convenient, and he further notes that the convention that $\text{sgn}(0) = 1$ is "convenient and inconsequential", see Manski (1985), pages 2–3. In his analysis, by virtue of the conditions shown to ensure point identification, Xb is continuously distributed for all $b \in \mathcal{B}$ so that $Xb = 0$ is a probability zero event.

If any such distribution is the population conditional distribution of $(U_1, ..., U_n)$ given $\mathcal{X}_n$, then $LR(b) = 0$ in accordance with our finding that there exist least favorable distributions of $G_{U|\mathcal{X}_n}$ for which any test has trivial power.

It is possible to construct a critical value for $LR(b)$ similarly to (3.6) as follows. Let $(Y_1^*, ..., Y_n^*)$ be independent (conditional on $\mathcal{X}_n$) random variables with $\mathbb{P}(Y^* = 0 \mid \mathcal{X}_n) = \mathbb{P}(Y^* = 1 \mid \mathcal{X}_n) = 1/2$. Using $(Y_1^*, ..., Y_n^*)$ in place of $(Y_1, ..., Y_n)$, we can construct the likelihood ratio test statistic $LR^*(b)$. Then a valid critical value is obtained by doing this for a large number of repetitions, and using the conditional $1 - \alpha$ quantile of $LR^*(b)$ given $\mathcal{X}_n$ across repeated samples as the critical value for $LR(b)$. However, this is computationally demanding because it computes maximum score estimators repeatedly, one for each sample of $(Y_1^*, ..., Y_n^*)$. Furthermore, Figure 9 in Appendix E shows that in the simulation designs of Section 5 with sample size of 100, the test proposed in Section 3 is more powerful than the test based on $LR(b)$ for all designs considered, while both approaches achieve finite sample size control.

5 Monte Carlo Experiments

In this section we present Monte Carlo results that illustrate the relative performance of our test compared to three other available inference methods. Specifically, in addition to our approach, we consider inference using the smoothed maximum score estimator introduced by Horowitz (1992), moment inequality inference proposed by Chen and Lee (2018), and inference using the bootstrap for cube root asymptotic developed by Cattaneo, Jansson, and Nagasawa (2020).[23]

We employ four different data generation processes in our simulations each for covariates of dimensions $K = 2$ and $K = 5$. Our simulation designs are the same as Horowitz (1992) when $K = 2$, see also Horowitz (2002) and Cattaneo, Jansson, and Nagasawa (2020). In this case the true parameter value is specified as $\beta = (1, 1)'$ with X bivariate normal such that $X_1 \sim N(0, 1)$ and $X_2 \sim N(1, 1)$, independent of one another. When $K = 5$, we extend the $K = 2$ design by specifying that $\beta = (1, 1, 0, 0, 0)'$ and that $X = (X_1, ..., X_K)$ is a vector of uncorrelated normal random variables with $\mathbb{E}[X_k] = 1\{k \neq 1\}$ and $Var(X_k) = 1$ for $k = 1, ..., K$. In all cases the first component of β is normalized to one, and in the simulations inference is conducted for different values of the second component.

The distribution of unobservable U in each design is as follows, with U independent of X in designs 1-3 and V independent of X in design 4.

[23] As mentioned at the end of the previous section, in Appendix E we also provide a brief comparison in a subset of the cases considered here to finite sample inference based on the likelihood ratio statistic.

- Design 1: U is distributed logistic with mean zero and variance normalized to one.
- Design 2: U is uniformly distributed on $[-\sqrt{3}, \sqrt{3}]$.
- Design 3: U is distributed Student-t with three degrees of freedom, normalized to have variance one.
- Design 4: $U = 0.25(1 + 2Z^2 + Z^4)V$ where $Z = X_1 + \cdots + X_K$ and V is distributed logistic with mean zero, normalized to have variance one.

We report the results of conducting tests of the null hypothesis $\beta = b$ against the alternative hypothesis $\beta \neq b$, where $b = (1, b_2)'$ for $K = 2$ and $b = (1, b_2, 0, 0, 0)'$ for $K = 5$. Figures 1-6 present non-rejection frequencies in which the value of b_2 ranges from -1 to 3 in increments of 0.1 with intermediate values interpolated for illustration. The nominal significance level of the test in all cases is $\alpha = 0.1$. For each of the four designs for U, and both specifications for covariates of dimension $K = 2$ and $K = 5$, we report results for samples of size $n \in \{100, 250, 1000\}$. All results for all cases are based on 500 simulation draws.

The results in the figures are labeled RU, CL, HSA, HSB, and CJN for (1) Rosen and Ura, (2) Chen and Lee, (3) Horowitz smoothed asymptotic, (4) Horowitz smoothed bootstrap, and (5) Cattaneo, Jansson, and Nagasawa, respectively. The details of each procedure used are outlined below, before then discussing the results shown in Figures 1-6.

To implement our procedure (RU) we followed the implementation steps described in Section 3.2. For computation of $\mathcal{V}_u$ and $\mathcal{V}_l$ we use one set $\mathcal{V}$ for both. When $K = 2$, we analytically enumerate all cells induced by the hyperplane arrangement $\{v \in \mathbb{R}^K : X_i v = 0\}$, as described in Appendix C. When $K = 5$, we use the first 500 representatives produced by the Rada and Černý (2018) algorithm for computational simplicity, as described in Section 2.3.[24] For computation of the critical value we used $r = 500$ samples of Bernoulli(1/2) variables $(Y_1^*, ..., Y_n^*)$.

For the implementation of CL, we follow the Index approach used by Chen and Lee (2019) in their simulation studies, using the inference procedure described in their Section 4. To follow Chen and Lee (2019) as closely as possible, we translated the Gauss code for their simulations to R. The approach applies intersection bound inference from Chernozhukov, Lee, and Rosen (2013) to a double index representation of the conditional moment inequalities implied by the binary response model, using kernel functions. To compute their statistic, we took the minimum value of the conditional moments over 500 randomly drawn values

[24] As noted in footnote 14, as a sensitivity check to this choice we also provide a comparison to using the first 1000 representatives produced by the Rada and Černý (2018) algorithm in Appendix E, and find only minor difference; see Figures 10 – 12.

for the conditioning index, matching the number of representative points $\mathcal{V}$ used with our approach, and to compute critical values we used 500 draws of n standard normal random variables, matching our use of 500 draws of $Y_1^*, ..., Y_1^*$. for our critical values.

The smoothed maximum score estimator introduced by Horowitz (1992) estimates β by maximizing $\sum_{i=1}^{n} (2Y_i - 1) K(X_i b/\sigma_n)$ over b, where $K(\cdot)$ is a kernel function and σ_n is the bandwidth.[25] Inference using smoothed maximum score was conducted using both asymptotic critical values (HSA) and bootstrap critical values (HSB). For inference using asymptotic critical values we used the asymptotic normal approximation afforded by Horowitz (1992) Theorem 2, while bootstrap critical values were implemented following Horowitz (2002). The results reported in this paper use an undersmoothing factor of 1/2. We used a fourth-order kernel function as also used in the Monte Carlo simulations in both of those earlier papers.

Bootstrap-based inference for cube root asymptotics (CJN) was implemented following Section 4.1 of Cattaneo, Jansson, and Nagasawa (2020), in which they applied their method to inference on the maximum score estimated. We used their plug-in estimator for the Hessian of the expected score criterion, which they denoted $\tilde{\mathbf{H}}_n^{MS}$. Their code applies to the case in which $K = 2$, and we thus only include illustrations for the performance of CJN in these cases.

Figures 1-6 present non-rejection frequencies for b_2 ranging from -1 to 3 for $\alpha = 0.10$. Figures 1-3 show the results with $K = 2$, and Figures 4-6 show the results with $K = 5$. Our construction of the critical value is guaranteed to achieve finite sample size control (for any given n), while the other methods have been previously shown to achieve asymptotic size control. The simulation results in this section confirm these theoretical results. We see that when the null hypothesis is correct ($b_2 = 1$) RU achieves finite sample size control in every design and for all sample sizes. HSB performs very well for our simulation designs, nearly achieving finite sample size control for all cases. Its improved performance relative to HSA aligns with the earlier observations from Horowitz (1992, 2002).

Looking first at Figures 1-3, we see that the methods based on asymptotic approximation generally perform better for larger sample sizes, as we might expect. HSA and CJN both over-reject the null at these sample sizes, but to a progressively lesser degree as we move from $n = 100$ in Figure 1 to $n = 1000$ in Figure 3.[26] CL achieves size control for all sample

[25] The smoothed maximum score objective function is not convex, and to the best of our knowledge there is no guaranteed algorithm for computing its exact global maximizer. For the results in this section, we use the NLOPT_LD_TNEWTON algorithm from Dembo and Steihaug (1983) implemented in NLopt, Johnson (2007–2019), executed through R's nloptr package, Ypma (2018), using the true parameter value as a starting value and the parameter space $\{1\} \times [-1, 3]^{K-1}$.

[26] In unreported results with $n = 2000$ we found continued improvement in the performance of CJN consistent with asymptotic theory, with non-rejection probability achieving the 90% target for $b_2 = 1$ in

sizes for designs 1–3, but not for design 4, which features heteroskedasticity. The overall pattern is similar when $K = 5$, as shown in Figures 4–6, although the performance of HSA in general and CL in design 4 is not as good as with $K = 2$. We conjecture here that this could be because with more covariates a larger sample size is required for the asymptotic approximations to work well.

Overall we find that the performance of our proposed inference approach held up well in comparison to the different methods considered. We conclude that no single approach dominates. CL performs well in designs 1–3. In some of these cases, depending on K and n it is outperformed by RU, and in some cases it performs better than RU, while in design 4 CL did not perform well at these sample sizes. HSB generally performed very well in all designs, but from the figures we see that it does not dominate either RU or CL. It is also important to note that asymptotic theory for smoothed maximum score invokes stronger assumptions than Assumptions 1 and 2 in this paper, and these additional assumptions hold in the data generating processes used in these illustrations.[27]

Finally, we note an asymmetry in power shared by our method and Chen and Lee (2019). Among the alternatives plotted in our figures, our method and CL are more powerful for $b_2 < 1$ than for $b_2 > 1$, where 1 is the true parameter value for the second component of β. This asymmetry in power reflects the reliance of both methods on moment inequalities that measure the disagreement between the sign of $X\beta$ and Xb. In these designs there is an inherent asymmetry in $\mathbb{P}(X\beta < 0 < Xb \text{ or } Xb < 0 < X\beta)$ between $b_2 < 1$ and $b_2 > 1$, and this is precisely what the asymmetry in the figures reflects. To illustrate this point, Figure 7, plots $\mathbb{P}(X\beta < 0 < Xb \text{ or } Xb < 0 < X\beta)$ for different values of b_2.[28] The figure illustrates the stark difference in $\mathbb{P}(X\beta < 0 < Xb \text{ or } Xb < 0 < X\beta)$ on either side of $b_2 = 1$. This asymmetry is a consequence of the particular hypothesized values for b that are tested and for which non-rejection frequencies are plotted in the figures.

In terms of computation time, testing a single parameter value as implemented here took roughly 0.6 seconds (23 seconds for 41 points) for a single parameter value in the most demanding case considered, in which $n = 1000$ and $K = 5$.[29] By way of comparison, testing a

design 4.

[27]Assumption 9 of Horowitz (1992) imposes that the function $z \mapsto \mathbb{P}(U \leq -z \mid X\beta = z, \tilde{X})$ is differentiable where $\tilde{X}$ is defined such that $(X\beta, \tilde{X})$ has the same information as X. In unreported Monte Carlo simulations of designs 1–4 modified such that the unobserved variable U was right-censored at 0 (i.e., changing the value of U to zero when $U \geq 0$) we found that HSB did not control size in some cases. For brevity we do not include these results here, as it is not surprising that tests employing HSB may not always perform well under a violation of the stated assumptions from Horowitz (1992). Nonetheless we note that our inference method is robust to violations of those additional assumptions. We thank an anonymous reviewer for suggesting such a comparison.

[28]Note that $\mathbb{P}(X\beta < 0 < Xb \text{ or } Xb < 0 < X\beta)$ is the same across designs 1-4 and $K = 2, 5$.

[29]Specifically, this computation time was obtained on a MacBook Air (M1, 2020) with 16 GB Memory.

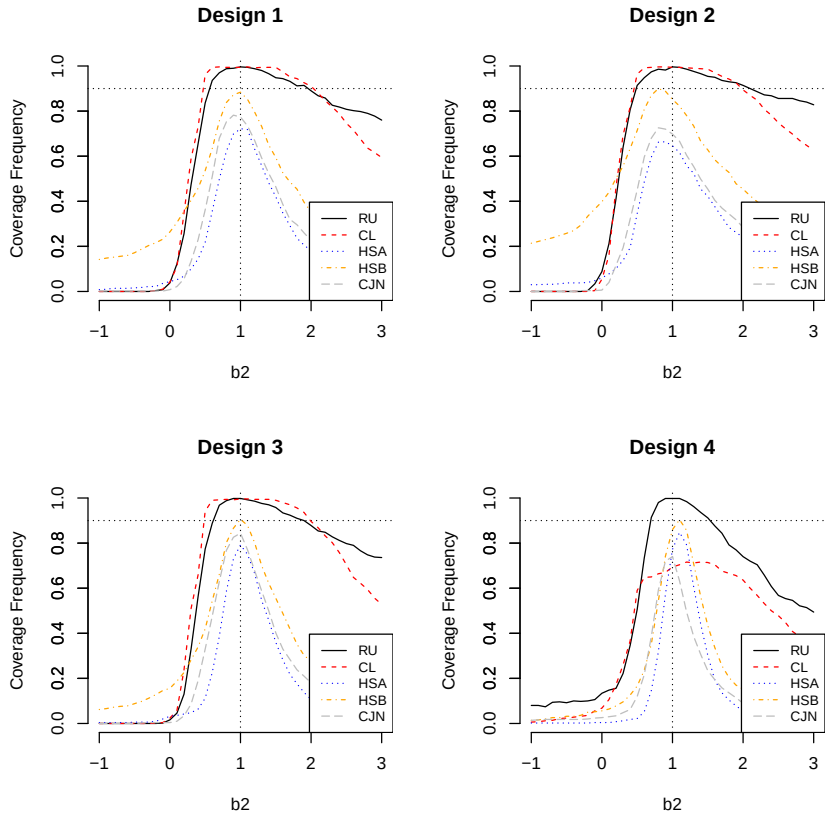


Figure 1: Non-rejection frequencies with $1 - \alpha = 90\%$ with $n = 100$ and $K = 2$. RU, CL, HSA, HSB, and CJN stand for (1) Rosen and Ura, (2) Chen and Lee, (3) Horowitz smoothed asymptotic, (4) Horowitz smoothed bootstrap, and (5) Cattaneo, Jansson, and Nagasawa, respectively.

single parameter value using our implementation of CL took roughly 1.8 seconds (72 seconds for 41 points).

6 Empirical Application

In this section, we apply the proposed inference method to the empirical application of Horowitz (1993). The dataset comprises a cross section of observations of 842 randomly sampled individuals from the Washington DC area transportation study. The outcome variable is an indicator of transportation choice between automobile ($Y = 1$) and public transit ($Y = 0$) as a method of commuting. The set of covariates is the number of cars in a household ($CARS$), the difference in the commuting cost ($DCOST$) in dollars, the difference in out-of-vehicle travel time in minutes ($DOVTT$), and the difference in in-vehicle

Using computers with different specifications or computing clusters will of course impact computation time.

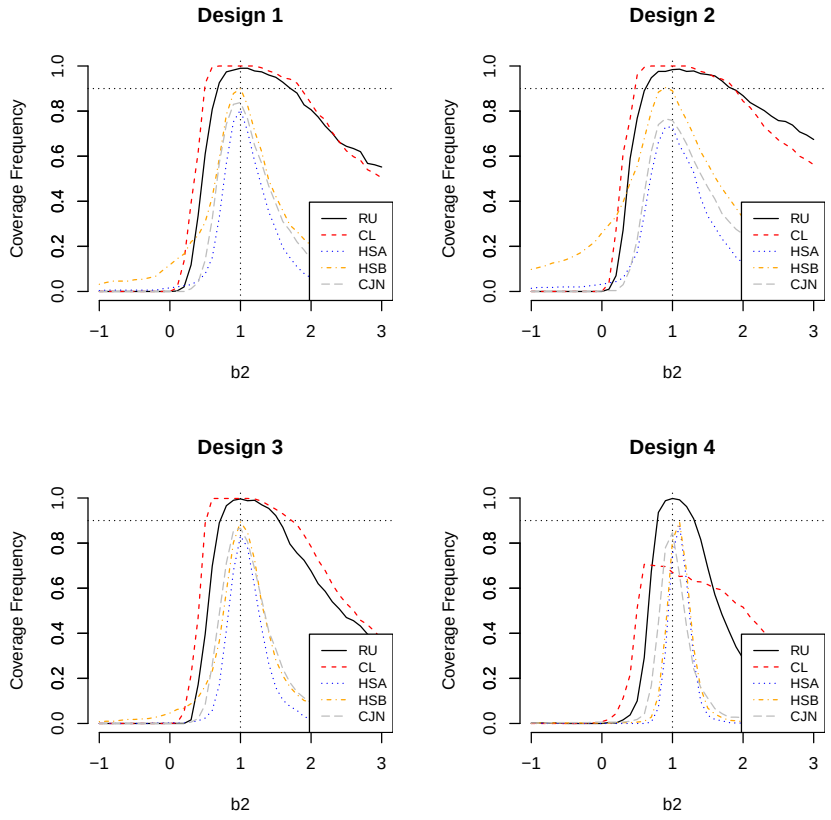


Figure 2: Non-rejection frequencies with $1-\alpha = 90\%$ with $n = 250$ and $K = 2$. RU, CL, HSA, HSB, and CJN stand for (1) Rosen and Ura, (2) Chen and Lee, (3) Horowitz smoothed asymptotic, (4) Horowitz smoothed bootstrap, and (5) Cattaneo, Jansson, and Nagasawa, respectively.

travel time in minutes ($DIVTT$). All differences are given by the amount for public transit minus that for automobile. Table 1 reports summary statistics for the variables. All the covariates are normalized to mean zero in the following analysis.

In the application of our inference method, we consider the analysis separately for households with 0, 1, and 2 cars, for which there are 79, 302, and 316 observations, respectively. This allows each of the other variables to affect these households differently. When computing the critical value for our test, we draw $r = 500$ sequences of n independent Bernoulli variables. We use $\alpha = 0.1$ as the nominal significance level of our test throughout.

We first conduct a hypothesis test for the joint significance of all covariates by testing $H_0 : \beta = 0$ against $H_0 : \beta \neq 0$. This is done separately for each value of $CARS$, by computing the test statistic in (3.2) and the critical value $q_{1-\alpha}(b)$ in Section 3. The results are reported in Table 2. For all three groups, the null hypothesis of $H_0 : \beta = 0$ is rejected. The usual conditions from the literature that ensure point identification as well as the standard

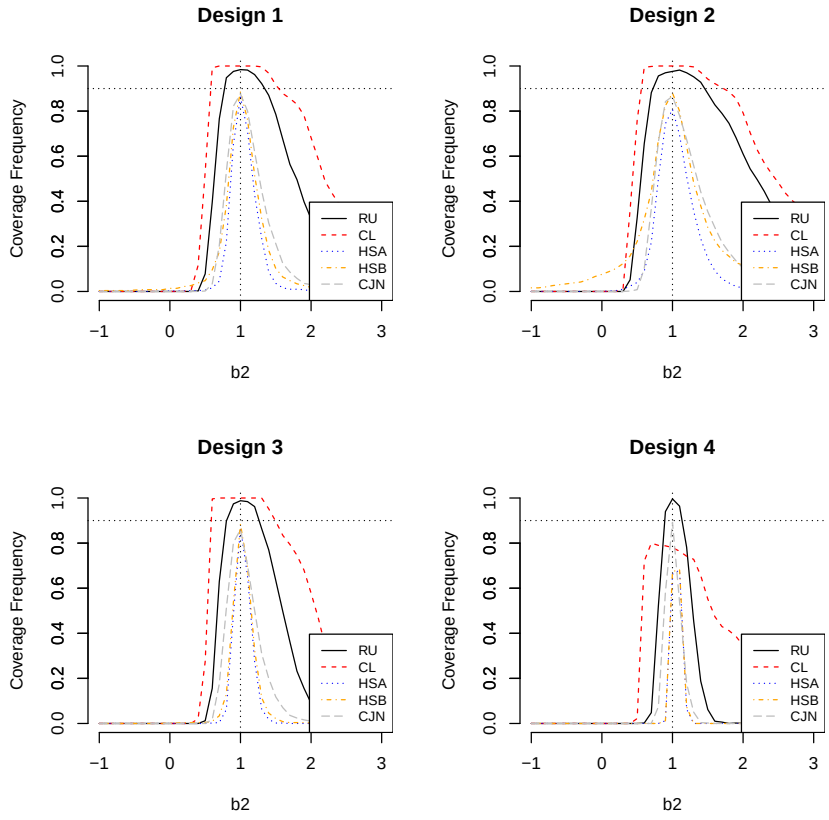


Figure 3: Non-rejection frequencies with $1-\alpha = 90\%$ with $n = 1000$ and $K = 2$. RU, CL, HSA, HSB, and CJN stand for (1) Rosen and Ura, (2) Chen and Lee, (3) Horowitz smoothed asymptotic, (4) Horowitz smoothed bootstrap, and (5) Cattaneo, Jansson, and Nagasawa, respectively.

normalizations imposed require that β has at least one non-zero component. Inference methods that rely on point identification as a condition for asymptotic inference assume that $H_0 : \beta = 0$ is false. Our inference method allows us to test this hypothesis directly, and it in fact provides a finite sample exact test when this hypothesis is correct.[30]

For a scale normalization one can restrict either the norm of the parameter vector or the magnitude of an individual parameter component to a constant. We normalize the magnitude of the intercept to one.[31] The conditional probability of $Y = 1$ given $CARS = 0$ is 0.21,

[30]Note that our test is calibrated to a least favorable configuration, which for any β is a DGP in which U has mass sufficiently far in the tails such that the sign of Y agrees with the sign of U almost surely. The sign of Y is also completely determined by that of U when $\beta = 0$.

[31]Horowitz (1993) normalizes the coefficient of $DCOST$ to one under the assumption that it is positive. The sign of this coefficient is difficult to determine from this dataset. When we use all observations, the score function takes a maximum of 0.906 if we normalize the coefficient of $DCOST$ to 1, whereas it takes a maximum of 0.899 if we normalize the coefficient of $DCOST$ to -1. We have thus chosen an alternative normalization, although it is reasonable to expect preference for automobiles to be increasing in $DCOST$.

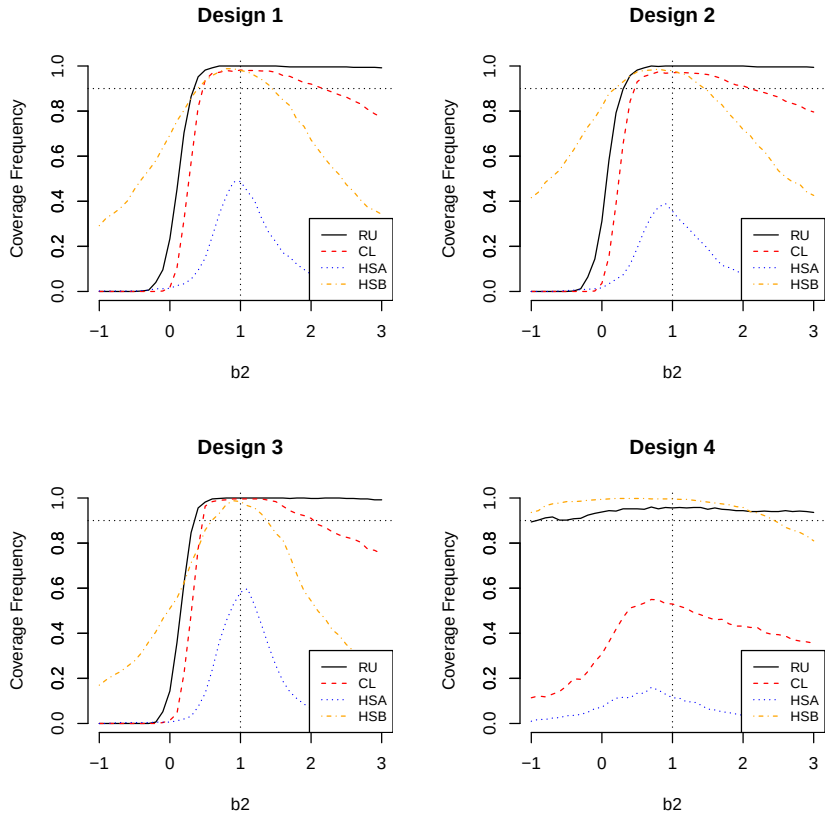


Figure 4: Non-rejection frequencies with $1-\alpha = 90\%$ with $n = 100$ and $K = 5$. RU, CL, HSA, and HSB stand for (1) Rosen and Ura, (2) Chen and Lee, (3) Horowitz smoothed asymptotic, and (4) Horowitz smoothed bootstrap, respectively.

which is less than $1/2$, so we normalize the intercept in this case to -1. For $CARS = 1$ and $CARS = 2$ the conditional probabilities are 0.85 and 0.95, respectively, so we normalize the intercept in these cases to $+1$. For all other parameter components, we set the parameter space for β to be the rectangle spanning from -10 to 10 in each dimension. We also compute the set of maximum score estimators, $\text{argmax}_{b\in\mathcal{B}} S_n(b)$. In Table 3, we report projections of this set onto each parameter component. For computation we used Florios and Skouras' (2008) exact characterization of weighted maximum score estimators using mixed-integer linear programming.[32]

We construct 90% confidence intervals for each parameter component β_k by inverting our test from Section 3 and reporting the endpoints of the resulting confidence region

$$\{b_k : b \in \mathcal{B} \text{ and } T_n(b) \leq q_{1-\alpha}(b)\}. \tag{6.1}$$

[32]For solving mixed integer programs we used the Gurobi optimizer (Gurobi Optimization, LLC (2023)).

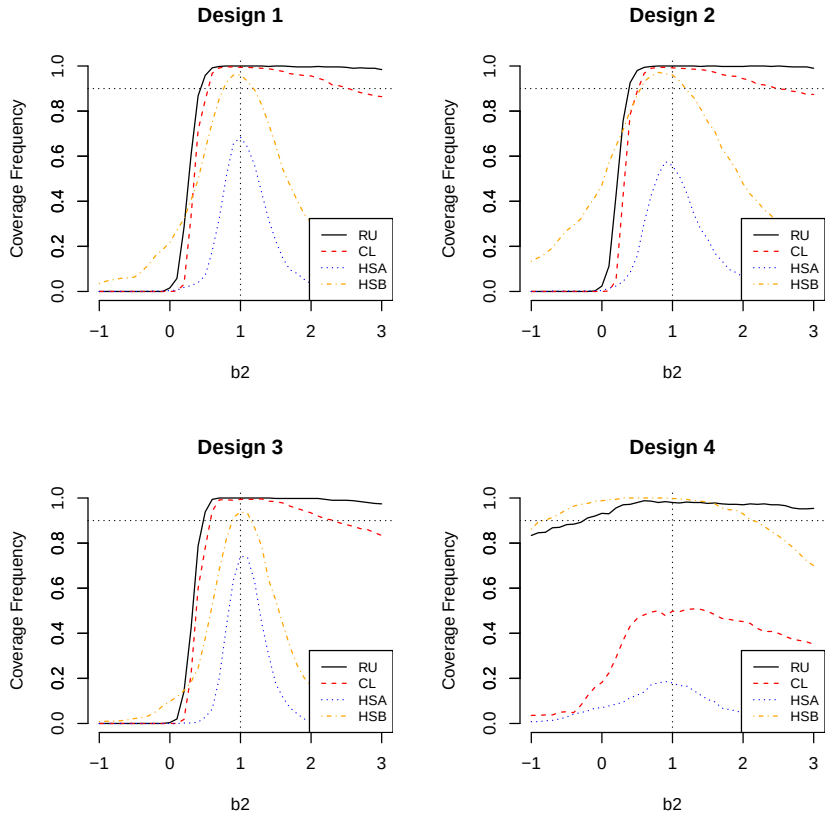


Figure 5: Non-rejection frequencies with $1-\alpha = 90\%$ with $n = 250$ and $K = 5$. RU, CL, HSA, and HSB stand for (1) Rosen and Ura, (2) Chen and Lee, (3) Horowitz smoothed asymptotic, and (4) Horowitz smoothed bootstrap, respectively.

In order to scan the parameter space appropriately when performing test inversion, we first used a quadratic mixed-integer program that characterizes a superset of our confidence region, and thus itself provides a conservative $1-\alpha$ confidence interval for each component. Specifically, we solve the quadratic mixed-integer program:

$$\text{maximize/minimize } b_k \tag{6.2}$$

over $(b, \{(Z_{ui}, Z_{li}) : i = 1, ..., n\}) \in \mathcal{B} \times \{0,1\}^{2n}$ subject to

$$X_i b \leq C Z_{ui}, -X_i b \leq C Z_{li}, Z_{ui} + Z_{li} = 1 \text{ for every } i = 1, ..., n; \tag{6.3}$$

$$-\sqrt{n}\frac{\mathbb{E}_n\left[(2Y-1)1\{Xv<0\}Z_u\right]}{\sqrt{\mathbb{E}_n\left[1\{0\ >Xv\}Z_u\right]-\left(\mathbb{E}_n\left[(2Y-1)1\{Xv<0\}Z_u\right]\right)^2}} \leq \bar{c}v \text{ for every } v \in \mathcal{V}_u; \text{ and} \tag{6.4}$$

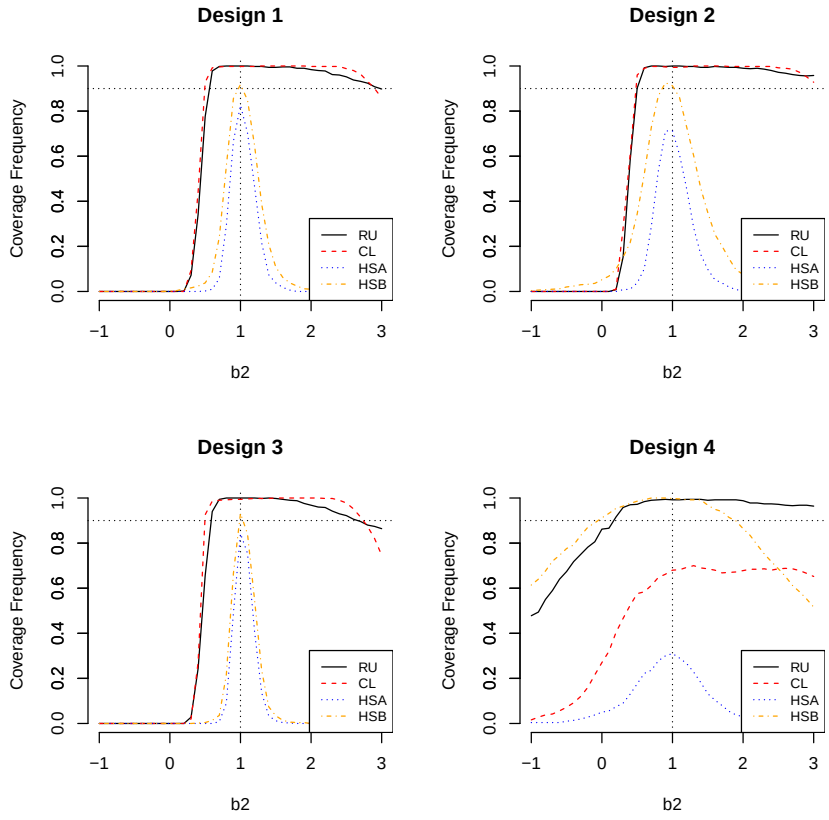


Figure 6: Non-rejection frequencies with $1-\alpha = 90\%$ with $n = 1000$ and $K = 5$. RU, CL, HSA, and HSB stand for (1) Rosen and Ura, (2) Chen and Lee, (3) Horowitz smoothed asymptotic, and (4) Horowitz smoothed bootstrap, respectively.

$$-\sqrt{n}\frac{\mathbb{E}_n\left[(1-2Y)1\{Xv>0\}Z_l\right]}{\sqrt{\mathbb{E}_n\left[1\{0\ <Xv\}Z_l\right]-\left(\mathbb{E}_n\left[(1-2Y)1\{Xv>0\}Z_l\right]\right)^2}} \leq \bar{cv} \text{ for every } v\in\mathcal{V}_l, \quad (6.5)$$

where C is a sufficiently large positive number. Lemma 2 in the Appendix formally shows that this mixed-integer program yields a superset of $\{b_k : b\in\mathcal{B} \text{ and } T_n(b)\leq \bar{cv}\}$, where $\bar{cv}$ is the maximum value of $q_{1-\alpha}(b)$ over $b\in\mathcal{B}$. This mixed-integer program differs from, but is inspired by that developed by Florios and Skouras (2008) for estimation.[33]

Table 4 reports the 90% region based on the above mixed-integer program. We constructed the regions separately according to the value of $CARS$. Regarding computation time, it took 17.9 seconds to make Table 4 for $CARS = 0$, 137.9 minutes for $CARS = 1$, and 55.3 minutes for $CARS = 2$.

Table 5 reports the resulting endpoints of the projection of a 90% confidence region

[33]For the sake of computation we use the first 500 representatives produced by the Rada and Černý (2018) algorithm, compute the critical value for each of the first 500 values produced, and set $\bar{cv}$ to the maximum. As in the previous section, we use these same 500 values as $\mathcal{V}_l$ and $\mathcal{V}_u$. When we conduct test inversion as described by (6.1), we use 10000 uniform random draws from the rectangular region reported in Table 4.

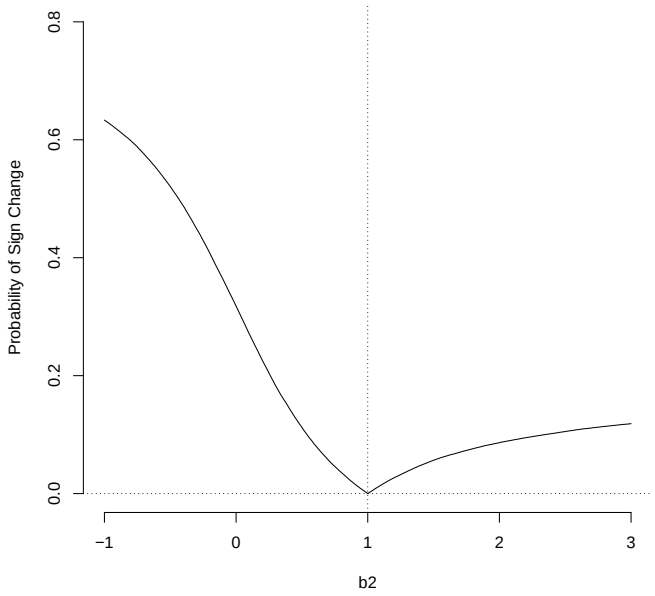


Figure 7: The proportion of observations for which $X\beta \neq Xb$.

	$CARS = 0$		$CARS = 1$		$CARS = 2$	
	mean	std. dev.	mean	std. dev.	mean	std. dev.
Y	0.21	0.41	0.85	0.36	0.95	0.22
$DCOST$	-20.97	39.46	-16.45	36.99	-10.55	39.00
$DOVTT$	7.51	7.88	12.91	10.27	13.61	10.24
$DIVTT$	12.63	14.38	16.52	16.19	17.38	19.16

Table 1: Summary statistics for the covariates.

($\alpha = 0.1$) onto each parameter component β_k using test inversion as described by (6.1), again separately for each value of $CARS$. Regarding computation time, it took 3.9 minutes to make Table 5 for $CARS = 0$, 19.2 minutes for $CARS = 1$, and 16.7 minutes for $CARS = 2$.

Overall, Tables 3 and 5 deliver mixed results. The interval estimates reported in Table 3 suggest that the coefficients on $DCOST$ and $DOVTT$ are both positive, while that for $DIVTT$ is negative for $CARS = 0$ and positive for $CARS = 1$. The interval estimates for each non-normalized coefficient cross zero for $CARS = 2$, so the implied sign in this case is ambiguous. This is not surprising given that 90% of individuals with $CARS = 2$ report commuting by automobile. Comparing the confidence intervals in Table 5 to the interval estimates in Table 3 for $CARS = 0$ and $CARS = 1$, we see that we cannot sign any parameter components at the 90% level. The confidence intervals for $CARS = 0$ are very wide, which makes sense since there were only 79 households in the sample with no cars.

	$CARS = 0$	$CARS = 1$	$CARS = 2$
Test statistic	8.15	20.20	37.19
Critical value	3.24	3.02	3.14

Table 2: Hypothesis tests for $\beta = 0$ with $\alpha = 0.1$.

Covariate	$CARS = 0$		$CARS = 1$		$CARS = 2$	
	lower	upper	lower	upper	lower	upper
Intercept	-1	-1	1	1	1	1
$DCOST$	0.0235	0.0277	0.0075	0.0108	-0.0110	0.0114
$DOVTT$	0.0312	0.2403	0.0284	0.0475	-0.0343	0.0600
$DIVTT$	-0.0684	-0.0106	0.0009	0.0133	-0.0129	0.0199

Table 3: Lower and upper bounds on each parameter component in the maximum score set estimate. The intercept is normalized to -1 for $CARS = 0$ and 1 for $CARS = 1, 2$.

Comparing the sets reported in Tables 3 and 5 is also useful for getting a sense for the relative effect of the size of set estimates versus sampling uncertainty. We see that in some cases the confidence intervals are substantially wider than the set estimates, while in other cases they are fairly close. For example, for $CARS = 1$ the interval estimate for the coefficient on $DCOST$ has length 0.0033 while the corresponding confidence interval has length 0.0380. For $CARS = 2$ the interval estimate and confidence interval are much closer with lengths 0.0224 and 0.0309, respectively. Interestingly, the confidence intervals produced by the mixed-integer program (6.2) – (6.5) are not much wider than those obtained by test inversion.

7 Conclusion

In this paper we proposed an approach to conduct finite sample inference on the parameters of Manski's (1985) semiparametric binary response model, for which the maximum score estimator has been shown to be cube-root consistent with a non-normal asymptotic distribution when there is point identification. Our approach circumvents the need to accommodate the complicated asymptotic behavior of this point estimator. Since our goal was finite sample inference, we considered the problem of making inference conditional on the n covariate vectors observable in a finite sample. With covariates taking only a finite number of observed values, the parameter vector β is not point identified. We therefore employed moment inequality implications for β for the sake of constructing our test statistic for inference, as the moment inequalities are valid no matter whether β is point identified. In order to exposit what observable implications can be distilled on only the basis of exogenous variables ob-

Covariate	$CARS = 0$		$CARS = 1$		$CARS = 2$	
Intercept	-1	-1	1	1	1	1
$DCOST$	-3.6578	10.0000	-0.0200	0.0233	-0.0178	0.0158
$DOVTT$	-10.0000	10.0000	-0.0573	0.1220	-0.0511	0.0928
$DIVTT$	-6.2916	10.0000	-0.0445	0.0599	-0.0300	0.0515

Table 4: Confidence intervals based on (6.2)-(6.5) for the coefficients with $1-\alpha = 90\%$. The intercept is normalized to -1 for $CARS = 0$ and 1 for $CARS = 1, 2$.

Covariate	$CARS = 0$		$CARS = 1$		$CARS = 2$	
	lower	upper	lower	upper	lower	upper
Intercept	-1	-1	1	1	1	1
$DCOST$	-3.1418	9.9983	-0.0161	0.0219	-0.0164	0.0145
$DOVTT$	-9.8789	9.9982	-0.0512	0.1106	-0.0477	0.0783
$DIVTT$	-5.3020	9.9981	-0.0416	0.0548	-0.0239	0.0476

Table 5: Confidence intervals for coefficients based on test inversion with $1-\alpha = 90\%$. The intercept is normalized to -1 for $CARS = 0$ and 1 for $CARS = 1, 2$.

served in the finite sample, we defined the notion of the set of conditionally observationally equivalent parameters $\mathcal{B}_n^*$. We showed how to make use of the full set of observable implications conditional on the size n sequence of exogenous variables in our construction of a test statistic $T_n(b)$. Finite sample valid critical values were established, and were shown to be easily computed by making use of many simulations of size n sequences of independent Bernoulli variables. A finite sample power (lower) bound was also presented and the results of some Monte Carlo experiments were reported, illustrating the performance of the test.

Several interesting directions for future research are possible. First, further study of the relative costs and benefits of using more values of v, and possibly using the full set of representative points afforded by cell enumeration, remains an open line of investigation. While a sharp moment inequality representation of $\mathcal{B}_n^*$ requires use of an exhaustive set of representatives, this may not be optimal for conducting inference. This parallels an earlier observation in the literature on partial identification, namely that it is possible that incorporating redundant or imprecisely-estimated moments into test statistics that use moment inequalities can result in less precise inference even when those moments are implied by identification analysis. Moreover, using more values of v increases computational requirements.

Second, the maximum score estimator is one of several estimators in the econometrics literature that consistently estimate a model parameter that may only be identified under support conditions that can never be satisfied by an empirical distribution based on a finite sample. Some such estimators, like the maximum score estimator, exhibit slower than $n^{-1/2}$

convergence rates. Other such estimators, such as the maximum rank correlation estimator of Han (1987) achieve the parametric rate. In this paper we have exploited the particular structure of the semi-parametric binary response model, but there may nonetheless be potential to extend ideas in this paper to such settings to alleviate dependence on conditions not satisfied by empirical distributions that result from finite data.

A third possible avenue pertains to optimal testing. One direction could be to exploit the likelihood ratio test analysis in this paper to consider minimax testing rates as $n \to \infty$ under additional assumptions on the distribution of U_i.[34] Minimax optimal estimation has recently been investigated in a setting with high-dimensional covariates by Mukherjee, Banerjee, and Ritov (2021) in an asymptotic framework under sufficient conditions for point identification. Investigation of minimax optimal estimators and tests could be interesting to consider when conditions for point identification are not guaranteed to hold. More generally, in future work we aim to continue to explore the interplay between partial identification and testability, and in particular the implications of not having point identification based on an underlying discrete data distribution, as one always has when using the empirical distribution obtained in a finite data set.

References

Abrevaya, J., and J. Huang (2005): "On the Bootstrap of the Maximum Score Estimator," *Econometrica*, 73(4), 1175–1204.

Andrews, D. W. K., and X. Shi (2013): "Inference Based on Conditional Moment Inequalities," *Econometrica*, 81(2), 609–666.

Armstrong, T., and M. Kolesár (2021): "Finite-Sample Optimal Estimation and Inference on Average Treatment Effects Under Unconfoundedness," *Econometrica*, 89(3), 1141–1177, arXiv:1712.04594v2.

Avis, D., and K. Fukuda (1996): "Reverse Search for Enumeration," *Discrete Applied Mathematics*, 65(1-3), 21–46.

Blevins, J. R. (2015): "Non-standard Rates of Convergence of Criterion-Function-Based Set Estimators for Binary Response Models," *The Econometrics Journal*, 18(2), 172–199.

Cattaneo, M. D., M. Jansson, and K. Nagasawa (2020): "Bootstrap-Based Inference for Cube Root Asymptotics," *Econometrica*, 88(5), 2203–2219.

Chen, L.-Y., and S. Lee (2018): "Best Subset Binary Prediction," *Journal of Econometrics*, 206(1), 39 – 56.

[34] This line of research was suggested by Tim Armstrong.

——— (2019): "Breaking the Curse of Dimensionality in Conditional Moment Inequalities for Discrete Choice Models," *Journal of Econometrics*, 210(2), 482–497.

Chernozhukov, V., C. Hansen, and M. Jansson (2009): "Finite Sample Inference for Quantile Regression Models," *Journal of Econometrics*, 152(2), 93 – 103.

Chernozhukov, V., S. Lee, and A. M. Rosen (2013): "Intersection Bounds: Estimation and Inference," *Econometrica*, 81(2), 667–737.

Chesher, A., and A. M. Rosen (2017): "Generalized Instrumental Variable Models," *Econometrica*, 85(3), 959–989.

Clopper, C., and E. S. Pearson (1934): "The Use of Confidence or Fiducial Limits Illustrated in the Case of the Binomial," *Biometrika*, 26(4), 404–413.

Cosslett, S. R. (1983): "Distribution-Free Maximum Likelihood Estimator of the Binary Choice Model," *Econometrica*, 51(3), 404–413.

Cover, T. M. (1965): "Geometrical and Statistical Properties of Systems of Linear Inequalities with Applications in Pattern Recognition," *IEEE Transactions on Electrical Computers*, EC-14(3), 326–334.

Delgado, M. A., J. M. Rodríguez-Poo, and M. Wolf (2001): "Subsampling Inference in Cube Root Asymptotics with an Application to Manski's Maximum Score Estimator," *Economics Letters*, 73(2), 241–250.

Dembo, R. S., and T. Steihaug (1983): "Truncated-Newton algorithms for large-scale unconstrained optimization," *Mathematical Programming*, 26, 190–212.

Florios, K. (2018): "A hyperplanes intersection simulated annealing algorithm for maximum score estimation," *Econometrics and Statistics*, 8, 37–55.

Florios, K., A. Louka, and Y. Bilias (2022): "Tabu Search for Maximum Score Estimator Computation," Available at `https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4091006`.

Florios, K., and S. Skouras (2008): "Exact computation of max weighted score estimators," *Journal of Econometrics*, 146(1), 86–91.

Gu, J., and R. Koenker (2022): "Nonparametric Maximum Likelihood Methods for Binary Response Models With Random Coefficients," *Journal of the American Statistical Association*, 117(538), 732–751.

Gu, J., and T. Russell (2023): "Partial identification in nonseparable binary response models with endogenous regressors," *Journal of Econometrics*, 235, 528 – 562.

Gu, J., T. Russell, and T. Stringham (2022): "Counterfactual Identification and Latent Space Enumeration in Discrete Outcome Models," working paper, Carleton University and University of Toronto.

Gurobi Optimization, LLC (2023): "Gurobi Optimizer Reference Manual," .

Han, A. K. (1987): "Non-parametric Analysis of a Generalized Regression Model: The Maximum Rank Correlation Estimator," *Journal of Econometrics*, 35, 303–316.

Hoeffding, W. (1963): "Probability Inequalities for Sums of Bounded Random Variables," *Journal of the American Statistical Association*, 58(301), 13–30.

Horowitz, J. L. (1992): "A Smoothed Maximum Score Estimator for the Binary Response Model," *Econometrica*, 60(3), 505–531.

——— (1993): "Semiparametric estimation of a work-trip mode choice model," *Journal of Econometrics*, 58(1-2), 49–70.

——— (2002): "Bootstrap Critical Values for Tests Based on the Smoothed Maximum Score Estimator," *Journal of Econometrics*, 141-167(2), 505–531.

Johnson, S. G. (2007–2019): "The NLopt nonlinear-optimization package," `https://github.com/stevengj/nlopt`.

Jun, S. J., J. Pinkse, and Y. Wan (2015a): "Classical Laplace estimation for n3-consistent estimators: Improved convergence rates and rate-adaptive inference," *Journal of Econometrics*, 187(1), 201–216.

——— (2015b): "Integrated Score Estimation," *Econometric Theory*, 33(6), 1418–1456.

Kaido, H., and Y. Zhang (2019): "Robust Likelihood Ratio Test for Incomplete Economic Models," Working paper, arXiv:1910.04610.

Khan, S., T. Komarova, and D. Nekipelov (2024): "Sharp and Robust Estimation of Partially Identified Discrete Response Models," Working paper, arXiv:2310.02414v4.

Kim, J., and D. Pollard (1990): "Cube Root Asymptotics," *Annals of Statistics*, 18(1), 191–219.

Komarova, T. (2013): "Binary Choice Models with Discrete Regressors: Identification and Misspecification," *Journal of Econometrics*, 177(1), 14 – 33.

Léger, C., and B. MacGibbon (2006): "On the Bootstrap in Cube Root Asymptotics," *The Canadian Journal of Statistics / La Revue Canadienne de Statistique*, 34(1), 29–44.

Lehmann, E. L., and J. P. Romano (2005): *Testing Statistical Hypotheses, Third Edition*. Springer.

Li, L., and M. Henry (2022): "Finite Sample Inference in Incomplete Models," Working paper, arXiv:2204.00473.

Manski, C. F. (1975): "Maximum Score Estimation of the Stochastic Utility Model of Choice," *Journal of econometrics*, 3(3), 205–228.

——— (1985): "Semiparametric Analysis of Discrete Response: Asymptotic Properties of the Maximum Score Estimator," *Journal of Econometrics*, 27(3), 313–333.

——— (2007): "Partial Identification of Counterfactual Choice Probabilities," *International Economic Review*, 48(4), 1393–1410.

Manski, C. F., and T. S. Thompson (1986): "Operational Characteristics of Maximum Score Estimation," *Journal of Econometrics*, 32(1), 85–108.

——— (1989): "Estimation of Best Predictors of Binary Response," *Journal of Econometrics*, 40(1), 97–123.

Molchanov, I. (2017): *Theory of Random Sets. Second Edition.* Springer-Verlag, London, UK.

Mukherjee, D., M. Banerjee, and Y. Ritov (2021): "Optimal Linear Discriminators For The Discrete Choice Model In Growing Dimensions," *Annals of Statistics*, 49(6), 3324–3357.

Patra, R. K., E. Seijo, and B. Sen (2018): "A Consistent Bootstrap Procedure for the Maximum Score Estimator," *Journal of Econometrics*, 205(2), 488–507.

Pinkse, C. (1993): "On the Computation of Semiparametric Estimates in Limited Dependent Variable Models," *Journal of Econometrics*, 58(1), 185 – 205.

Powell, J. L. (1994): "Estimation of Semiparametric Models," in *The Handbook of Econometrics*, ed. by R. F. Engle, and D. L. McFadden, vol. 4. North-Holland.

Rada, M., and M. Černý (2018): "A New Algorithm for Enumeration of Cells of Hyperplane Arrangements and a Comparison with Avis and Fukuda's Reverse Search," *Siam Journal of Discrete Mathematics*, 32(1), 455 – 473.

Seo, M. H., and T. Otsu (2018): "Local M-estimation with discontinuous criterion for dependent and limited observations," *Ann. Statist.*, 46(1), 344–369.

Sleumer, N. (1998): "Output-sensitive Cell Enumeration in Hyperplane Arrangements," in *Algorithm Theory — SWAT'98*, ed. by S. Arnborg, and L. Ivansson, pp. 300–309. Springer-Verlag Berlin Heidelberg.

Syrgkanis, V., E. Tamer, and J. Ziani (2018): "Inference on Auctions with Weak Assumptions on Information," Working paper, arXiv:1710.03830.

Ypma, J. (2018): "Introduction to nloptr: an R interface to NLopt," `https://cran.r-project.org/web/packages/nloptr/vignettes/nloptr.pdf`.

Online Appendix

A Proofs

Proof of Theorem 1. It suffices to show

$$P_{(\beta,G_0)}\left(\phi(b,Y_1,...,Y_n;\mathcal{X}_n)=1 \mid \mathcal{X}_n\right) \leq \sup_{G\in\mathcal{G}} P_{(b,G)}(\phi(b,Y_1,...,Y_n;\mathcal{X}_n)=1 \mid \mathcal{X}_n). \tag{A.1}$$

By the definition of $\mathcal{B}_n^*$, there exists a conditional distribution of $(U_1,\cdots,U_n)$ given $\mathcal{X}_n$, say $\tilde{G}$, that satisfies all the requirements of Assumption 1 and is such that

$$P_{(\beta,G_0)}\left(Y_i=y_i \text{ for all } i \mid \mathcal{X}_n\right) = P_{(b,\tilde{G})}\left(Y_i=y_i \text{ for all } i \mid \mathcal{X}_n\right).$$

Then we have

$$\begin{aligned}
&P_{(\beta,G_0)}\left(\phi(b,Y_1,...,Y_n;\mathcal{X}_n)=1 \mid \mathcal{X}_n\right)\\
&= \sum_{(y_1,...,y_n)\in\{0,1\}^n} \phi(b,y_1,...,y_n;\mathcal{X}_n) P_{(\beta,G_0)}(Y_i=y_i \text{ for all } i \mid \mathcal{X}_n)\\
&= \sum_{(y_1,...,y_n)\in\{0,1\}^n} \phi(b,y_1,...,y_n;\mathcal{X}_n) P_{(b,\tilde{G})}\left(Y_i=y_i \text{ for all } i \mid \mathcal{X}_n\right),
\end{aligned}$$

which implies Eq. (A.1). □

Proof of Lemma 1. If $X_i\beta \geq 0$, then $Y_i = 1\{X_i\beta+U_i\geq 0\} \geq 1\{U_i\geq 0\}$ and therefore $\mathbb{E}[2Y_i-1 \mid \mathcal{X}_n] \geq 2\mathbb{P}(U_i\geq 0 \mid \mathcal{X}_n)-1 = 0$. If $X_i\beta\leq 0$, then $Y_i = 1\{X_i\beta+U_i\geq 0\} \leq 1\{U_i\geq 0\}$ and therefore $\mathbb{E}[2Y_i-1 \mid \mathcal{X}_n] \leq 2\mathbb{P}(U_i\geq 0 \mid \mathcal{X}_n)-1 = 0$. □

Proof of Theorem 2. Directly from Lemma 1, $b\in\mathcal{B}_n^*$ implies

$$\mathbb{E}\left[(2Y_i-1)\,1\{X_ib\geq 0\} \mid \mathcal{X}_n\right] \geq 0 \text{ and } \mathbb{E}\left[(1-2Y_i)\,1\{X_ib\leq 0\} \mid \mathcal{X}_n\right] \geq 0.$$

To demonstrate the other direction, let b be any element of $\mathcal{B}$ such that $\mathbb{E}[(2Y_i-1)1\{X_ib\geq 0\} \mid \mathcal{X}_n] \geq 0$ and $\mathbb{E}[(1-2Y_i)1\{X_ib\leq 0\} \mid \mathcal{X}_n] \geq 0$ for every $i=1,...,n$. Then

$$X_ib\geq 0 \implies \mathbb{P}(Y_i=1 \mid \mathcal{X}_n) \geq 1/2, \tag{A.2}$$

$$X_ib\leq 0 \implies \mathbb{P}(Y_i=1 \mid \mathcal{X}_n) \leq 1/2. \tag{A.3}$$

To show that b is in $\mathcal{B}_n^*$ as defined in Definition 1, we now construct a collection of random variables $\{\tilde{U}_i : i=1,...,n\}$ such that for all $i=1,...,n$: (i) $\mathbb{P}(Y_i = 1\{X_ib+\tilde{U}_i\geq 0\} \mid \mathcal{X}_n) = 1$,

and (ii) $\mathbb{P}(\forall i, 1\{U_i \geq 0\} = 1\{\tilde{U}_i \geq 0\} \mid \mathcal{X}_n) = 1$. To do so, let $\kappa_i : i = 1, ..., n$ be n positive random variables defined on $(\Omega, \mathfrak{F}, \mathbb{P})$ and consider each of the cases $X_i\beta < 0$, $X_i\beta = 0$, and $X_i\beta > 0$ in turn as follows.

Case 1: $X_i\beta < 0$. By Lemma 1,

$$\mathbb{E}[2Y_i - 1 \mid \mathcal{X}_n] \leq 0. \tag{A.4}$$

Let

$$\tilde{U}_i \equiv \begin{cases} \max\{-X_i b, 0\} + \kappa_i & \text{if } U_i \geq -X_i\beta \\ 0 & \text{if } 0 \leq U_i < -X_i\beta \\ \min\{-X_i b, 0\} - \kappa_i & \text{if } U_i < 0. \end{cases}$$

Then $1\{\tilde{U}_i \geq 0\} = 1\{U_i \geq 0\}$, which verifies (ii). To verify (i), note that:

$$\begin{aligned} 1\{X_i b + \tilde{U}_i \geq 0\} &= 1\{X_i\beta + U_i \geq 0\} + 1\{0 \leq U_i < -X_i\beta,\ X_i b \geq 0\} \\ &= Y_i + 1\{0 \leq U_i < -X_i\beta, X_i b \geq 0\}, \end{aligned}$$

because

$$X_i b + \tilde{U}_i = \begin{cases} \max\{X_i b, 0\} + \kappa_i & \text{if } U_i \geq -X_i\beta \\ X_i b & \text{if } 0 \leq U_i < -X_i\beta \\ \min\{X_i b, 0\} - \kappa_i & \text{if } U_i < 0. \end{cases}$$

Therefore,

$$\mathbb{P}(Y_i = 1\{X_i b + \tilde{U}_i \geq 0\} \mid \mathcal{X}_n) = \mathbb{P}(1\{0 \leq U_i < -X_i\beta, X_i b \geq 0\} = 0 \mid \mathcal{X}_n).$$

Thus if $X_i b < 0$, then $\mathbb{P}(Y_i = 1\{X_i b + \tilde{U}_i \geq 0\} \mid \mathcal{X}_n) = 1$. Suppose instead that $X_i b \geq 0$. Then

$$\begin{aligned} \mathbb{P}(Y_i = 1\{X_i b + \tilde{U}_i \geq 0\} \mid \mathcal{X}_n) &= \mathbb{P}(1\{0 \leq U_i < -X_i\beta, X_i b \geq 0\} = 0 \mid \mathcal{X}_n) \\ &= \mathbb{P}(1\{0 \leq U_i < -X_i\beta\} = 0 \mid \mathcal{X}_n) \\ &= \mathbb{P}(U_i < 0 \mid \mathcal{X}_n) + \mathbb{P}(U_i \geq -X_i\beta \mid \mathcal{X}_n) \\ &= \frac{1}{2} + \mathbb{P}(U_i \geq -X_i\beta \mid \mathcal{X}_n). \end{aligned}$$

Since $X_i b \geq 0$ inequality (A.2) holds. Combining this with (A.4) implies that $\mathbb{P}(Y_i = 1 \mid \mathcal{X}_n) = 1/2$, implying in turn that $\mathbb{P}(Y_i = 1\{X_i b + \tilde{U}_i \geq 0\} \mid \mathcal{X}_n) = 1$, which verifies (i).

Case 2: $X_i\beta = 0$. Let

$$\tilde{U}_i \equiv \begin{cases} \max\{-X_i b, 0\} + \kappa_i & \text{if } U_i \geq 0 \\ \min\{-X_i b, 0\} - \kappa_i & \text{if } U_i < 0. \end{cases} \tag{A.5}$$

Then $1\{\tilde{U}_i \geq 0\} = 1\{U_i \geq 0\}$, which verifies (ii). It further follows from (A.5) that

$$X_i b + \tilde{U}_i = \begin{cases} \max\{0, X_i b\} + \kappa_i & \text{if } U_i \geq 0 \\ \min\{0, X_i b\} - \kappa_i & \text{if } U_i < 0. \end{cases}$$

Consequently since $X_i\beta = 0$, $X_i b + \tilde{U}_i \geq 0$ if and only if $X_i\beta + U_i \geq 0$, verifying (i).
Case 3: $X_i\beta > 0$. The proof is similar to Case 1 with

$$\tilde{U}_i \equiv \begin{cases} \max\{-X_i b, 0\} + \kappa_i & \text{if } U_i \geq 0 \\ -X_i b & \text{if } -X_i\beta \leq U_i < 0 \\ \min\{-X_i b, 0\} - \kappa_i & \text{if } U_i < -X_i b. \end{cases}$$

□

Proof of Theorem 3. By Theorem 2, $b \in \mathcal{B}_n^*$ implies (2.9) and (2.10). For the rest of the proof, we are going to demonstrate the other direction. Namely, we are going to show (2.9) and (2.10) imply

$$\forall i = 1, ..., n: \quad \mathbb{E}\left[(2Y_i - 1)\, 1\{X_i b \geq 0\} \mid \mathcal{X}_n\right] \geq 0 \text{ and } \mathbb{E}\left[(1 - 2Y_i)\, 1\{X_i b \leq 0\} \mid \mathcal{X}_n\right] \geq 0, \tag{A.6}$$

which establishes $b \in \mathcal{B}^*$ from the characterization of $\mathcal{B}^*$ provided by Theorem 2. Let $v^* \in \mathcal{V}$ be such that for all i with $X_i\beta \neq 0$, $X_i v^*$ has the same sign as $X_i\beta$. The existence of v^* is guaranteed by setting $v^* = \beta + \varepsilon\lambda$ for an appropriately selected vector $\lambda \in \mathbb{R}^K$ and sufficiently small $\varepsilon > 0$ such that $X_i v^* \cdot X_i\beta = (X_i\beta)^2 + \varepsilon(X_i\lambda \cdot X_i\beta) > 0$ for all i with

$X_i\beta \neq 0$. It follows that

$$\begin{aligned}
-\frac{1}{n}\sum_{i=1}^{n}\left|\mathbb{E}\left[2Y_i - 1 \mid \mathcal{X}_n\right]\right| 1\{X_i b \geq 0, X_i\beta < 0\} &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[2Y_i - 1 \mid \mathcal{X}_n\right] 1\{X_i b \geq 0, X_i\beta < 0\} \\
&= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[2Y_i - 1 \mid \mathcal{X}_n\right] 1\{X_i b \geq 0, X_i v^* < 0\} \\
&= \mathbb{E}\left[\, \mathbb{E}_n\left[(2Y-1)1\{Xb \geq 0, Xv^* < 0\}\right] \mid \mathcal{X}_n\right] \\
&\geq 0, \qquad \text{(A.7)}
\end{aligned}$$

where the first two equalities follow by Lemma 1, noting in particular for the second equality the implication that $X_i\beta = 0$ implies that $E\left[2Y_i - 1 \mid \mathcal{X}_n\right] = 0$. The third equality follows from taking $1\{X_i b \geq 0, X_i v^* < 0\}$ inside the conditional expectation, and then interchanging the sum and the conditional expectation. The inequality in the last line holds because b satisfies the inequality in (2.9) with $v = v^*$. Following similar steps employing Lemma 1 and the inequality in (2.10) with $v = v^*$ additionally gives

$$\begin{aligned}
-\frac{1}{n}\sum_{i=1}^{n}\left|\mathbb{E}\left[2Y_i - 1 \mid \mathcal{X}_n\right]\right| 1\{X_i b \leq 0, X_i\beta > 0\} &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[1 - 2Y_i \mid \mathcal{X}_n\right] 1\{X_i b \leq 0, X_i\beta > 0\} \\
&= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[1 - 2Y_i \mid \mathcal{X}_n\right] 1\{Xb \leq 0, Xv^* > 0\} \\
&= \mathbb{E}\left[\, \mathbb{E}_n\left[(1-2Y)\, 1\{Xb \leq 0, Xv^* > 0\}\right] \mid \mathcal{X}_n\right] \\
&\geq 0. \qquad \text{(A.8)}
\end{aligned}$$

Since $-\left|\mathbb{E}\left[2Y_i - 1 \mid \mathcal{X}_n\right]\right| 1\{X_i b \geq 0, X_i\beta < 0\}$ and $-\left|\mathbb{E}\left[2Y_i - 1 \mid \mathcal{X}_n\right]\right| 1\{X_i b \leq 0, X_i\beta > 0\}$ must both be non-positive for every i, we have that for all $i = 1, ..., n$:

$$\left|\mathbb{E}\left[2Y_i - 1 \mid \mathcal{X}_n\right]\right| 1\{X_i b \geq 0, X_i\beta < 0\} = \left|\mathbb{E}\left[2Y_i - 1 \mid \mathcal{X}_n\right]\right| 1\{X_i b \leq 0, X_i\beta > 0\} = 0. \quad \text{(A.9)}$$

From this (A.6) holds for every $i = 1, ..., n$ by considering the three cases: $\mathbb{E}\left[2Y_i - 1 \mid \mathcal{X}_n\right] = 0$, $\mathbb{E}\left[2Y_i - 1 \mid \mathcal{X}_n\right] > 0$, and $\mathbb{E}\left[2Y_i - 1 \mid \mathcal{X}_n\right] < 0$. For every i with $\mathbb{E}\left[2Y_i - 1 \mid \mathcal{X}_n\right] = 0$, the inequalities in (A.6) hold with equality. For every i with $\mathbb{E}\left[2Y_i - 1 \mid \mathcal{X}_n\right] > 0$, we have $X_i\beta > 0$ from Lemma 1, and therefore (A.9) implies $X_i b > 0$, which in turn implies (A.6). For every i with $\mathbb{E}\left[2Y_i - 1 \mid \mathcal{X}_n\right] < 0$, we have $X_i\beta < 0$ from Lemma 1, and therefore (A.9) implies $X_i b < 0$, which in turn implies (A.6). □

Proof of Theorem 4. Thoughout the proof of this theorem, we assume $H_0 : \beta = b$. First we will establish that $P_{(\beta,G)}(T_n(b) \leq q_{1-\alpha}(b) \mid \mathcal{X}_n) \geq 1 - \alpha$ for every $G \in \mathcal{G}$. Note that if (3.5)

holds, then $P_{(\beta,G)}\left(T_n(b) \leq q_{1-\alpha}(b) \mid \mathcal{X}_n\right) \geq \mathbb{P}\left(T_n^*(\beta) \leq q_{1-\alpha}(b) \mid \mathcal{X}_n\right) \geq 1-\alpha$, so it will suffice to show inequality (3.5). Since

$$\begin{cases} Y_i \geq Y_i^* & \text{if } X_i\beta \geq 0 \\ Y_i \leq Y_i^* & \text{if } X_i\beta \leq 0, \end{cases}$$

for every $i = 1, ..., n$, we have

$$\mathbb{E}_n\left[(2Y-1)1\{X\beta \geq 0 > Xv\}\right] \geq \mathbb{E}_n\left[(2Y^*-1)1\{X\beta \geq 0 > Xv\}\right], \forall v \in \mathcal{V}_u$$

and

$$\mathbb{E}_n\left[(1-2Y)1\{X\beta \leq 0 < Xv\}\right] \geq \mathbb{E}_n\left[(1-2Y^*)1\{X\beta \leq 0 < Xv\}\right], \forall v \in \mathcal{V}_l.$$

By the construction of $T^*(\beta)$ and $T_n(\beta)$, it suffices to show that, for every $v \in \mathbb{R}^K$, the two functions,

$$t \mapsto \max\left\{0, \sqrt{n}\frac{-t}{\sqrt{\mathbb{E}_n[1\{X\beta \geq 0 > Xv\}] - t^2}}\right\} \tag{A.10}$$

and

$$t \mapsto \max\left\{0, \sqrt{n}\frac{-t}{\sqrt{\mathbb{E}_n[1\{X\beta \leq 0 < Xv\}] - t^2}}\right\} \tag{A.11}$$

are weakly decreasing. For the rest of the proof, we focus on the first function

$$f(t) \equiv \max\left\{0, \sqrt{n}\frac{-t}{\sqrt{\mathbb{E}_n[1\{X\beta \geq 0 > Xv\}] - t^2}}\right\}.$$

Consider t_1 and t_2 with $t_1 < t_2$. If $t_2 \geq 0$, we have $f(t_1) \geq 0 = f(t_2)$. For the rest of the proof, therefore, we are going to show $f(t_1) \geq f(t_2)$ when $t_1 < t_2 < 0$. Since $t_2^2 < t_1^2$, we have

$$\mathbb{E}_n[1\{X\beta \geq 0 > Xv\}] - t_1^2 < \mathbb{E}_n[1\{X\beta \geq 0 > Xv\}] - t_2^2,$$

so

$$0 < \sqrt{\mathbb{E}_n[1\{X\beta \geq 0 > Xv\}] - t_1^2} \leq \sqrt{\mathbb{E}_n[1\{X\beta \geq 0 > Xv\}] - t_2^2}.$$

Since $-t_1 > -t_2 > 0$, we have

$$\frac{-t_1}{\sqrt{\mathbb{E}_n[1\{X\beta \geq 0 > Xv\}] - t_1^2}} > \frac{-t_2}{\sqrt{\mathbb{E}_n[1\{X\beta \geq 0 > Xv\}] - t_2^2}}.$$

Therefore, $f(t_1) > f(t_2)$.

Now we establish that for any $\tilde{c} < q_{1-\alpha}(b)$, we have that $\inf_{G\in\mathcal{G}} P_{(b,G)}(T_n(b) \leq \tilde{c} \mid \mathcal{X}_n) < 1-\alpha$. To see why this is so, let G be any distribution of $U_1, ..., U_n$ given $\mathcal{X}_n$ such that $U_1, ..., U_n$ are independent given $\mathcal{X}_n$ and that

$$G\left(U_i < -|X_i b| \mid \mathcal{X}_n\right) = G\left(U_i > |X_i b| \mid \mathcal{X}_n\right) = \frac{1}{2}.$$

Any such G satisfies Assumption 1, and together with $\beta = b$ implies that for each i, $\mathbb{P}\left(Y_i = 1|\mathcal{X}_n\right) = 1/2$. □

Proof of Theorem 5. In this proof, we focus on Eq. (3.7). Define $W = (2Y-1)1\{Xb \geq 0 > Xv\}$. First, we are going to show that

$$\sqrt{n}\mathbb{E}_n[W] < -q_{1-\alpha}(b)\sqrt{\frac{\mathbb{E}_n[W^2]}{1+q_{1-\alpha}(b)^2/n}} \implies T_n(b) > q_{1-\alpha}(b). \tag{A.12}$$

Suppose $\sqrt{n}\mathbb{E}_n[W] < -q_{1-\alpha}(b)\sqrt{\frac{\mathbb{E}_n[W^2]}{1+q_{1-\alpha}(b)^2/n}}$. Note that

$$\mathbb{E}_n[W] < 0 \tag{A.13}$$

and

$$n\mathbb{E}_n[W]^2 > \mathbb{E}_n[W^2]\frac{q_{1-\alpha}(b)^2}{1+q_{1-\alpha}(b)^2/n}.$$

The second inequality implies $n\mathbb{E}_n[W]^2 > \mathbb{E}_n[W^2]q_{1-\alpha}(b)^2 - \mathbb{E}_n[W]^2 q_{1-\alpha}(b)^2$. Using Eq. (A.13), $-\sqrt{n}\mathbb{E}_n[W] > q_{1-\alpha}(b)\sqrt{\mathbb{E}_n[W^2] - \mathbb{E}_n[W]^2}$ and then

$$\sqrt{n}\frac{-\mathbb{E}_n[W]}{\sqrt{\mathbb{E}_n[W^2] - \mathbb{E}_n[W]^2}} > q_{1-\alpha}(b)$$

which implies $T_n(b) > q_{1-\alpha}(b)$.

Then, we are going to show $\mathbb{P}(T_n(b) > q_{1-\alpha}(b) \mid \mathcal{X}_n) \geq 1-\rho$. Using Eq. (A.12), we have

$$\mathbb{P}(T_n(b) > q_{1-\alpha}(b) \mid \mathcal{X}_n) \geq \mathbb{P}\left(\sqrt{n}\mathbb{E}_n[W] < -q_{1-\alpha}(b)\sqrt{\frac{\mathbb{E}_n[W^2]}{1+q_{1-\alpha}(b)^2/n}} \mid \mathcal{X}_n\right).$$

Eq. (3.7) implies

$$\begin{aligned}
&\mathbb{P}(T_n(b) > q_{1-\alpha}(b) \mid \mathcal{X}_n) \\
&\geq \mathbb{P}\left(\mathbb{E}_n[W] < \mathbb{E}\left[\mathbb{E}_n[W] \mid \mathcal{X}_n\right] + \sqrt{\frac{2\log(1/\rho)\mathbb{E}_n[1\{Xb \geq 0 > Xv\}]}{n}} \mid \mathcal{X}_n \right) \\
&= 1 - \mathbb{P}\left(\mathbb{E}_n[W] \geq \mathbb{E}\left[\mathbb{E}_n[W] \mid \mathcal{X}_n\right] + \sqrt{\frac{2\log(1/\rho)\mathbb{E}_n[1\{Xb \geq 0 > Xv\}]}{n}} \mid \mathcal{X}_n \right).
\end{aligned}$$

Since $-1\{Xb \geq 0 > Xv\} \leq W_i \leq 1\{Xb \geq 0 > Xv\}$ for every $i = 1, ..., n$, Hoeffding (1963)'s inequality implies

$$\mathbb{P}(T_n(b) > q_{1-\alpha}(b) \mid \mathcal{X}_n) \geq 1 - \exp\left(-\frac{2n^2 \left(\sqrt{\frac{2\log(1/\rho)\mathbb{E}_n[1\{Xb \geq 0 > Xv\}]}{n}} \right)^2}{4\sum_{i=1}^n 1\{X_i b \geq 0 > X_i v\}} \right) = 1 - \rho.$$

□

Proof of Corollary 1. Let $1 - \rho$ denote the maximum of the power bounds (3.9) and (3.10) in the statement of the corollary. If the expressions $\max\{0, \cdot\}$ in (3.9) and (3.10) are both zero for all $v \in \mathcal{V}_u$ and $v \in \mathcal{V}_l$, then the implication of the corollary is trivially satisfied. Thus suppose instead that the maximum of (3.9) and (3.10) is greater than zero. It follows that there is either a $v \in \mathcal{V}_u$ such that

$$\rho = \exp\left(-\frac{1}{2}\left(\sqrt{n}\zeta_u(b, v) - q_{1-\alpha}(b)(1 + q_{1-\alpha}(b)^2/n)^{-1/2} \right)^2 \right),$$

or a $v \in \mathcal{V}_l$ such that

$$\rho = \exp\left(-\frac{1}{2}\left(\sqrt{n}\zeta_l(b, v) - q_{1-\alpha}(b)(1 + q_{1-\alpha}(b)^2/n)^{-1/2} \right)^2 \right),$$

implying (3.9) in the former case and (3.10) in the latter. The conclusion of the corollary then follows from Theorem 5. □

Proof of Corollary 2. The first part of the corollary can be shown by first noting that $\max\{Q_u(b), Q_l(b)\} \geq C_\alpha(b, 1 - \rho)$ implies that at least one of the following inequalities

hold:

$$\mathbb{E}\left[\ \mathbb{E}_n\left[(2Y-1)1\{Xb \geq 0 > Xv\}\right] \mid \mathcal{X}_n\right] \leq -C_\alpha(b, 1-\rho),$$
$$\mathbb{E}\left[\ \mathbb{E}_n\left[(1-2Y)1\{Xb \leq 0 < Xv\}\right] \mid \mathcal{X}_n\right] \leq -C_\alpha(b, 1-\rho).$$

Then because $-C_\alpha(b, 1-\rho)$ is less than or equal to each of the expressions on the right hand side of inequalities (3.7) and (3.8), at least one of (3.7) and (3.8) is true and Theorem 5 delivers the result.

For the second claim of the corollary, consider first that if $\sqrt{n}\max\{Q_u(b), Q_l(b)\} \leq q_{1-\alpha}(b)(1+q_{1-\alpha}(b)^2/n)^{-1/2}$ the result holds trivially. Thus, suppose instead that $\sqrt{n}\max\{Q_u(b), Q_l(b)\} > q_{1-\alpha}(b)(1+q_{1-\alpha}(b)^2/n)^{-1/2}$ and consider

$$\rho = \exp\left(-\frac{1}{2}\left(\sqrt{n}\max\{Q_u(b), Q_l(b)\} - q_{1-\alpha}(b)(1+q_{1-\alpha}(b)^2/n)^{-1/2}\right)^2\right).$$

Then $\max\{Q_u(b), Q_l(b)\} = C_\alpha(b, 1-\rho)$ and the desired implication follows from the first part of the corollary. □

Proof of Theorem 6. Assume w.l.o.g. that $\{i : \bar{p}_i \neq \tilde{p}_i\} = \{1, ..., \bar{n}\}$, which can be achieved by rearranging i's. Then $(\bar{p}_1, ..., \bar{p}_{\bar{n}}) = (1/2, ..., 1/2)$ and the test in (4.7) simplifies as

$$\bar{\phi}(Y_1, ..., Y_{\bar{n}}; \mathcal{X}_n) = \begin{cases} 1 & \text{if } \prod_{i=1}^{\bar{n}} \tilde{p}_i^{Y_i}(1-\tilde{p}_i)^{1-Y_i} > k2^{-\bar{n}}, \\ \xi & \text{if } \prod_{i=1}^{\bar{n}} \tilde{p}_i^{Y_i}(1-\tilde{p}_i)^{1-Y_i} = k2^{-\bar{n}}, \\ 0 & \text{otherwise.} \end{cases}$$

Define

$$\mathrm{RP}(p_1, ..., p_{\bar{n}}) \equiv \sum_{(y_1,...,y_{\bar{n}})\in\{0,1\}^{\bar{n}}} \bar{\phi}(y_1, ..., y_{\bar{n}}; \mathcal{X}_n) \prod_{i=1}^{\bar{n}} p_i^{y_i}(1-p_i)^{1-y_i}.$$

For the rest of the proof, we first show that $\mathrm{RP}(p_1, ..., p_{\bar{n}})$ is the probability that $\bar{\phi}(y_1, ..., y_n)$ rejects the null hypothesis when $\mathbb{P}\{Y_i = 1 \mid \mathcal{X}_n\} = p_i$ for all i, that is,

$$\mathrm{RP}(p_1, ..., p_{\bar{n}}) = \sum_{(y_1,...,y_n)\in\{0,1\}^n} \bar{\phi}(y_1, ..., y_n) \prod_{i=1}^{n} p_i^{y_i}(1-p_i)^{1-y_i}, \tag{A.14}$$

where the right-hand side of the above equation is the sum of 2^n terms instead of $2^{\bar{n}}$. From this, it follows from (4.8) that $\mathrm{RP}(\bar{p}_1, ..., \bar{p}_{\bar{n}}) = \alpha$. Subsequently we show that $(\bar{p}_1, ..., \bar{p}_{\bar{n}})$ is the constrained maximizer of $\mathrm{RP}(p_1, ..., p_{\bar{n}})$ with respect to $p_1, ..., p_{\bar{n}}$, subject to $p_1, ..., p_{\bar{n}}$ being compatible with the null hypothesis. Thus (4.7) achieves finite sample size control,

i.e.,

$$\sum_{(y_1,...,y_n)\in\{0,1\}^n} \bar{\phi}(y_1,...,y_n)\prod_{i=1}^{n} p_i^{y_i}(1-p_i)^{1-y_i} \leq \alpha$$

for any sequence $(p_1,...,p_n)$ under the null hypothesis. By Theorem 3.8.1 of Lehmann and Romano (2005) the test (4.7) is a most powerful test of the composite null $\beta = b$ against the simple alternative H_1, and (4.9) is the least favorable distribution of $(Y_1,...,Y_n)$ given $\mathcal{X}_n$.

First, to establish (A.14), we have

$$\begin{aligned}
\mathrm{RP}(p_1,...,p_{\bar{n}}) &= \sum_{(y_1,...,y_{\bar{n}})\in\{0,1\}^{\bar{n}}} \bar{\phi}(y_1,...,y_{\bar{n}};\mathcal{X}_n)\prod_{i=1}^{\bar{n}} p_i^{y_i}(1-p_i)^{1-y_i} \\
&= \sum_{(y_1,...,y_{\bar{n}})\in\{0,1\}^{\bar{n}}} \bar{\phi}(y_1,...,y_{\bar{n}};\mathcal{X}_n)\sum_{(y_{\bar{n}+1},...,y_n)}\prod_{i=1}^{n} p_i^{y_i}(1-p_i)^{1-y_i} \\
&= \sum_{(y_1,...,y_n)\in\{0,1\}^n} \bar{\phi}(y_1,...,y_n)\prod_{i=1}^{n} p_i^{y_i}(1-p_i)^{1-y_i},
\end{aligned}$$

where the second equality holds because $\sum_{(y_{\bar{n}+1},...,y_n)}\prod_{i=\bar{n}+1}^{n} p_i^{y_i}(1-p_i)^{1-y_i} = 1$ and the third equality follows because $\bar{\phi}((Y_1,...,Y_n);\mathcal{X}_n)$ does not depend on $y_{\bar{n}+1},...,y_n$.

For the next step of the proof, notice that (4.6) implies that, under the null hypothesis, $(\mathbb{P}(Y_i = 1 \mid \mathcal{X}_n) - 1/2)(\tilde{p}_i - 1/2) \leq 0$ for every $i = 1,...,\bar{n}$. Thus we next show that $(\bar{p}_1,...,\bar{p}_{\bar{n}})$ is the constrained maximizer of $\mathrm{RP}(p_1,...,p_{\bar{n}})$ subject to $(p_i - 1/2)(\tilde{p}_i - 1/2) \leq 0$ for all $i = 1,...,\bar{n}$. Let $\bar{\phi}(d, y_{-j})$ be shorthand for $\bar{\phi}(y_1,...,y_{j-1},d,y_{j+1},...,y_n)$, where $d \in \{0,1\}$. Note that

$$\begin{cases} \bar{\phi}(1,y_{-j}) - \bar{\phi}(0,y_{-j}) \geq 0 & \text{if } \tilde{p}_j > 1/2, \\ \bar{\phi}(1,y_{-j}) - \bar{\phi}(0,y_{-j}) \leq 0 & \text{if } \tilde{p}_j < 1/2, \end{cases}$$

for every $j = 1,...,\bar{n}$, and that $\tilde{p}_j \neq 1/2$ because $\bar{p}_j \neq \tilde{p}_j$ for all $j = 1,...,\bar{n}$. Since

$$\begin{aligned}
&\frac{\partial}{\partial p_j}\mathrm{RP}(p_1,...,p_{\bar{n}}) \\
&= \sum_{(y_1,...,y_{\bar{n}})\in\{0,1\}^{\bar{n}}} (2y_j - 1)\bar{\phi}(y_1,...,y_{\bar{n}};\mathcal{X}_n)\prod_{i\neq j} p_i^{y_i}(1-p_i)^{1-y_i} \\
&= \sum_{y_{-j}\in\{0,1\}^{\bar{n}-1}} \bar{\phi}(1,y_{-j})\prod_{i\neq j} p_i^{y_i}(1-p_i)^{1-y_i} - \sum_{y_{-j}\in\{0,1\}^{\bar{n}-1}} \bar{\phi}(0,y_{-j})\prod_{i\neq j} p_i^{y_i}(1-p_i)^{1-y_i} \\
&= \sum_{y_{-j}\in\{0,1\}^{\bar{n}-1}} (\bar{\phi}(1,y_{-j}) - \bar{\phi}(0,y_{-j}))\prod_{i\neq j} p_i^{y_i}(1-p_i)^{1-y_i},
\end{aligned}$$

we have

$$\begin{cases} \frac{\partial}{\partial p_j}\mathrm{RP}(p_1,...,p_{\bar{n}}) \geq 0 & \text{if } \tilde{p}_j > 1/2, \\ \frac{\partial}{\partial p_j}\mathrm{RP}(p_1,...,p_{\bar{n}}) \leq 0 & \text{if } \tilde{p}_j < 1/2. \end{cases}$$

Thus $(1/2,...,1/2)$ maximizes $\mathrm{RP}(p_1,...,p_{\bar{n}})$ with respect to $(p_1,...,p_{\bar{n}})$ subject to $(p_i - 1/2)(\tilde{p}_i - 1/2) \leq 0$ for all $i = 1,...,\bar{n}$, completing the proof. □

Proof of Corollary 3. The corollary follows directly from Theorem 6 and the reasoning given in the text. □

Proof of Theorem 7. If we can show

$$\sup_{G\in\mathcal{G}} G(1\{X_ib + U_i \geq 0\} = y_i \text{ for all } i \mid \mathcal{X}_n) = \prod_{i=1}^{n}\left(1 - \frac{1}{2}\cdot 1\left\{(2y_i - 1)X_ib \leq 0\right\}\right)$$

for every $(y_1,...,y_n) \in \{0,1\}^n$, then

$$\begin{aligned} \ell_{(Y_1,...,Y_n)}(b) &= \sum_{i=1}^{n}\log\left(1 - \frac{1}{2}\cdot 1\left\{(2Y_i - 1)X_ib \leq 0\right\}\right) \\ &= \sum_{i=1}^{n}(1\left\{(2Y_i - 1)X_ib \leq 0\right\}\log(1/2) + 1\left\{(2Y_i - 1)X_ib > 0\right\}\log 1) \\ &= -\log(2)\sum_{i=1}^{n}1\left\{(2Y_i - 1)X_ib \leq 0\right\} \\ &= \frac{n\log(2)}{2}(\frac{1}{n}\sum_{i=1}^{n}(1\left\{(2Y_i - 1)X_ib > 0\right\} - 1\left\{(2Y_i - 1)X_ib \leq 0\right\}) - 1) \\ &= \frac{n\log(2)}{2}(\tilde{S}_n(b) - 1). \end{aligned}$$

Let $(y_1,...,y_n)$ be any element of $\{0,1\}^n$. For the rest of the proof, we are going to show

$$\sup_{G\in\mathcal{G}} G(1\{X_ib + U_i \geq 0\} = y_i \text{ for all } i \mid \mathcal{X}_n) = \prod_{i=1}^{n}\left(1 - \frac{1}{2}\cdot 1\left\{(2y_i - 1)X_ib \leq 0\right\}\right).$$

First, we are going to show

$$G(1\{X_ib + U_i \geq 0\} = y_i \text{ for all } i \mid \mathcal{X}_n) \leq \prod_{i=1}^{n}\left(1 - \frac{1}{2}\cdot 1\left\{(2y_i - 1)X_ib \leq 0\right\}\right)$$

for every $G \in \mathcal{G}$. Note that

$$(1\{X_ib + U_i \geq 0\} = y_i \text{ for all } i) \implies (1\{X_ib + U_i \geq 0\} = y_i \text{ for all } i \text{ such that } (2y_i - 1)X_ib \leq 0)$$

and that

$$(1\{X_ib + U_i \geq 0\} = y_i \text{ and } (2y_i - 1)X_ib \leq 0) \implies U_i \in \mathcal{U}(y_i).$$

where $\mathcal{U}(0) \equiv (-\infty, 0)$ and $\mathcal{U}(1) \equiv [0, \infty)$. Therefore,

$$G(1\{X_ib + U_i \geq 0\} = y_i \text{ for all } i \mid \mathcal{X}_n) \leq G(U_i \in \mathcal{U}(y_i) \text{ for all } i \text{ such that } (2y_i - 1)X_ib \leq 0).$$

By Assumption 1, the events $U_i \in \mathcal{U}(y_i)$, $i = 1, ..., n$ are mutually independent and $G(U_i \in \mathcal{U}(y_i)) = 1/2$. Then

$$G(1\{X_ib+U_i \geq 0\} = y_i \text{ for all } i \mid \mathcal{X}_n) \leq \prod_{i:(2y_i-1)X_ib<0} 1/2 = \prod_{i=1}^{n} \left(1 - \frac{1}{2} \cdot 1\left\{(2y_i - 1)X_ib \leq 0\right\}\right).$$

Second, we are going to show

$$G(1\{X_ib + U_i \geq 0\} = y_i \text{ for all } i \mid \mathcal{X}_n) = \prod_{i=1}^{n} \left(1 - \frac{1}{2} \cdot 1\left\{(2y_i - 1)X_ib \leq 0\right\}\right)$$

for some $G \in \mathcal{G}$. Consider the distribution of $(U_1, ..., U_n)$ given $\mathcal{X}_n$ such that $(U_1, ..., U_n)$ are independent given $\mathcal{X}_n$ and that

$$U_i = \begin{cases} \begin{cases} \frac{2y_i+1}{2}|X_ib| & \text{with probability } 1/2 \text{ given } \mathcal{X}_n \\ -\frac{3-2y_i}{2}|X_ib| & \text{with probability } 1/2 \text{ given } \mathcal{X}_n \end{cases} & \text{if } X_ib \neq 0 \\ \begin{cases} 1 & \text{with probability } 1/2 \text{ given } \mathcal{X}_n \\ -1 & \text{with probability } 1/2 \text{ given } \mathcal{X}_n \end{cases} & \text{if } X_ib = 0. \end{cases}$$

Such G belongs to $\mathcal{G}$. By construction,

$$G(1\{X_ib + U_i \geq 0\} = y_i \text{ for all } i \mid \mathcal{X}_n) = \prod_{i=1}^{n} G(1\{X_ib + U_i \geq 0\} = y_i \mid \mathcal{X}_n).$$

If $X_ib = 0$, then

$$\begin{aligned} G(1\{X_ib + U_i \geq 0\} = y_i \mid \mathcal{X}_n) &= \frac{1}{2}1\{1\{1 \geq 0\} = y_i\} + \frac{1}{2}1\{1\{-1 \geq 0\} = y_i\} \\ &= \frac{1}{2} \\ &= 1 - \frac{1}{2} \cdot 1\left\{(2y_i - 1)X_ib \leq 0\right\}. \end{aligned}$$

If $X_i b \neq 0$, then

$$
\begin{aligned}
&G(1\{X_i b + U_i \geq 0\} = y_i \mid \mathcal{X}_n) \\
&= \frac{1}{2} 1\{1\{X_i b + \frac{2y_i + 1}{2}|X_i b| \geq 0\} = y_i\} + \frac{1}{2} 1\{1\{X_i b - \frac{3 - 2y_i}{2}|X_i b| \geq 0\} = y_i\} \\
&= \frac{1}{2} 1\{X_i b + \frac{3}{2}|X_i b| \geq 0, y_i = 1\} + \frac{1}{2} 1\{X_i b + \frac{1}{2}|X_i b| < 0, y_i = 0\} \\
&\quad + \frac{1}{2} 1\{X_i b - \frac{1}{2}|X_i b| \geq 0, y_i = 1\} + \frac{1}{2} 1\{X_i b - \frac{3}{2}|X_i b| < 0, y_i = 0\} \\
&= \frac{1}{2} 1\{y_i = 1\} + \frac{1}{2} 1\{X_i b < 0, y_i = 0\} + \frac{1}{2} 1\{X_i b > 0, y_i = 1\} + \frac{1}{2} 1\{y_i = 0\} \\
&= 1 - \frac{1}{2} \cdot 1\left\{(2y_i - 1) X_i b \leq 0\right\}.
\end{aligned}
$$

□

B Almost Sure Convergence of $\mathcal{X}_n$ and $\mathcal{B}_n^*$

Suppose that Assumption 1 holds, and in addition, as assumed in prior papers in the maximum score literature, that observations (Y_i, X_i) comprise a random sample from population distribution $\mathbb{P}_{YX}$. Let $\mathcal{S}_X$ denote the support of X. Then the identified set for β is

$$\mathcal{B}^* = \{b \in \mathcal{B} : \mathbb{E}\left[(2Y - 1)\, 1\{Xb \geq 0\} \mid X\right] \geq 0 \geq \mathbb{E}\left[(2Y - 1)\, 1\{Xb \leq 0\} \mid X\right] \text{ almost surely}\}.$$

This representation follows from using the same steps as in the proof of Theorem 2, but replacing conditioning on $\mathcal{X}_n$ by conditioning on X, and replacing Y_i and X_i with Y and X throughout. Equivalently, this set can be expressed as

$$\mathcal{B}^* = \{b \in \mathcal{B} : \mathbb{P}\{X \in \mathcal{R}_b\} = 0\},$$

where

$$\mathcal{R}_b \equiv \{x \in \mathcal{S}_X : (xb \geq 0 \text{ and } \mathbb{E}\left[2Y - 1 | X = x\right] < 0) \text{ or } (xb \leq 0 \text{ and } \mathbb{E}\left[2Y - 1 | X = x\right] > 0)\}.$$

The set $\mathcal{R}_b$ is the set of values of x for which knowledge of $\mathbb{P}\{Y = 1 | X = x\}$ would enable a researcher to distinguish b from β under the maintained assumptions. If X takes such values with positive probability, then b is observationally distinct from β, while if X takes such values with probability zero, then b and β are observationally equivalent.[1] We

[1]The set of values of x for which knowledge of $\mathbb{P}\{Y = 1 | X = x\}$ allows one to distinguish b from β is defined differently in Manski (1985, LMDR Identification), specifically in the notation of the present paper

then have the following result.

Theorem 8. *Suppose that Assumption 1 holds and that (Y_i, X_i) are independent and identically distributed with support $\mathcal{S}_X$. Then (i) $\mathcal{X}_n \xrightarrow{a.s.} \mathcal{S}_X$, and (ii) $\mathcal{B}_n^* \xrightarrow{a.s.} \mathcal{B}^*$.*

Proof Theorem 8. To prove both parts of the Theorem we make use of Proposition 1.7.23 on page 137 of Molchanov (2017) which states that a sequence of random sets $\mathcal{A}_n$ converges almost surely to a nonstochastic set $\mathcal{A}$ on $\mathbb{R}^K$ if and only if for the following conditions hold:[2]

1. If $\mathcal{K} \cap \mathcal{A} = \emptyset$ for any compact set $\mathcal{K} \subseteq \mathbb{R}^K$ then $\mathbb{P}\{\mathcal{A}_n \cap \mathcal{K} \neq \emptyset \text{ i.o.}\} = 0$.

2. If $\mathcal{G} \cap \mathcal{A} \neq \emptyset$ for any open set $\mathcal{G} \subseteq \mathbb{R}^K$ then $\mathbb{P}\{\mathcal{A}_n \cap \mathcal{G} = \emptyset \text{ i.o.}\} = 0$.

First we show (i) with the sequence of random sets $\mathcal{A}_n = \mathcal{X}_n$ and nonstochastic set $\mathcal{A} = \mathcal{S}_X$. Condition 1 follows because $\mathcal{K} \cap \mathcal{S}_X = \emptyset$ and $\mathbb{P}\{\mathcal{X}_n \subseteq \mathcal{S}_X\} = 1$ implies that $\mathbb{P}\{\mathcal{X}_n \cap \mathcal{K} \neq \emptyset\} = 0$ for all n. Next we show Condition 2. Let $\mathcal{G} \subseteq \mathbb{R}^K$ be an open set with $\mathcal{G} \cap \mathcal{S}_X \neq \emptyset$. Then $p_{\mathcal{G}} \equiv \mathbb{P}\{X_i \in \mathcal{G}\} > 0$, because $\mathcal{S}_X$ is by definition the complement of the largest set on which $\mathbb{P}_X$ vanishes, and $p_{\mathcal{G}} = 0$ would imply that $\mathbb{P}_X$ vanishes on $\mathcal{S}_X^c \cup \mathcal{G}$, where $\mathcal{S}_X^c$ denotes the complement of $\mathcal{S}_X$. That $p_{\mathcal{G}} > 0$ implies that $\mathcal{X}_n \cap \mathcal{G} \neq \emptyset$ infinitely often with probability one, completing the proof of (i).

Second, we similarly show (ii) by verifying Conditions 1 and 2 with $\mathcal{A}_n = \mathcal{B}_n^*$ and $\mathcal{A} = \mathcal{B}^*$. Let $\mathcal{K}$ be a compact set on $\mathbb{R}^K$ such that $\mathcal{K} \cap \mathcal{B}^* = \emptyset$. Consider the event $\mathcal{E}_n \equiv \{\mathcal{K} \cap \mathcal{B}_n^* \neq \emptyset\}$. Since $\mathcal{K} \cap \mathcal{B}^* = \emptyset$, $\mathcal{E}_n$ implies that $\exists b \notin \mathcal{B}^*$ with $b \in \mathcal{B}_n$, equivalently that $\exists b \notin \mathcal{B}^*$ s.t. $\mathcal{X}_n \cap \mathcal{R}_b \neq \emptyset$. However, $\mathcal{X}_n \cap \mathcal{R}_b \neq \emptyset$ holds if and only if there is some $i \in \{1, ..., n\}$ such that either (a) $X_i b \geq 0$ and $\mathbb{E}[2Y - 1 | X = X_i] < 0$, or (b) $X_i b \leq 0$ and $\mathbb{E}[2Y - 1 | X = X_i] > 0$, and from the definition of $\mathcal{B}^*$ this occurs with probability zero *for any* $b \notin \mathcal{B}^*$. Hence $\mathbb{P}\{\mathcal{E}_n\} = 0$ for any n, and consequently $\mathbb{P}\{\mathcal{E}_n \text{ i.o.}\} = 0$, so Condition 1 is verified. Now let $\mathcal{G}$ be an open set such that $\mathcal{G} \cap \mathcal{B}^* \neq \emptyset$. It is easy to see that $\mathcal{B}_n^* \supseteq \mathcal{B}^*$ and hence $\mathcal{G} \cap \mathcal{B}_n^* \neq \emptyset$. Thus $\mathbb{P}\{\mathcal{G} \cap \mathcal{B}_n^* = \emptyset \text{ i.o.}\} = 0$, which verifies Condition 2 and completes the proof. □

---

as

$$\mathcal{R}_b' \equiv \{x \in \mathcal{S}_X : (xb \geq 0 \text{ and } \mathbb{E}[2Y - 1 | X = x] < 0) \text{ or } (xb < 0 \text{ and } \mathbb{E}[2Y - 1 | X = x] \geq 0)\}.$$

This makes use of the implications in display (1') on page 315 of Manski (1985), which it is remarked are "more convenient" for the analysis in that paper than the implications in display (1). Indeed, under Manski (1985) Assumption 2 it follows that $Xb \neq 0$ almost surely for any b, and consequently $\mathbb{P}\{X \in \mathcal{R}_b\} = \mathbb{P}\{X \in \mathcal{R}_b'\}$. If the distribution of X does not satisfy these additional restrictions, for example if X has a discrete distribution, or if alternative restrictions are imposed on the conditional distribution of U given X, different observable implications leading to different identified sets may be obtained. Appendix G compares, for example, observable implications obtained if the conditional CDFs of U given X are assumed strictly increasing in addition to $\mathbb{P}\{U \geq 0 | X\} = 1/2$.

[2]Proposition 1.7.23 in Molchanov (2017) covers the more general case in which $\mathcal{A}$ resides on a locally compact Hausdorff second countable space, of which K-dimensional Euclidean space is a special case.

C Cell Enumeration with Two Covariates

When the covariates X_i have only two components, characterization of the hyperplane arrangement is greatly simplified. This makes it an ideal case in which to illustrate determination and computation of the cells of the resulting partition. Because the distribution of the unobservables is nonparametrically specified, a scale normalization may be imposed, for example by restricting the first component of β to have absolute value of one. Due to the scale normalization, the circumstance in which there are only two components of X_i is the simplest non-trivial case in which to study the semiparametric binary response model.

The panels of Figure 8 illustrate the construction of the partition V comprised of cells induced by the hyperplane arrangement with bivariate X_i in a simple example in which $n = 3$. Panel (a) depicts X_1, X_2, and X_3, into which panel (b) additionally incorporates the values of v_i for which $X_i v_i = 0$. Panel (c) drops the covariate vectors X_i and panel (d) uses different colors to depict the interiors of the elements of V. Theorem 3 indicates that in order to fully characterize $\mathcal{B}_n^*$, it suffices to employ moment inequalities of the form (2.9) and (2.10) with a set $\mathcal{V}$ of values for v containing one element from the interior of each of these six different colored regions.

To understand the further simplification afforded by the scale normalization, first note that for any v with $v_1 \neq 0$ we can normalize the first component of v such that $|v_1| = 1$ in the instrument functions that appear in (2.9) and (2.10) for the same reason that one can normalize the first component of the parameter vector β.[3] Specifically for any $v \in \mathbb{R}^K$ with $v_1 \neq 0$,

$$X_i v \gtreqless 0 \iff X_{i1} \frac{v_1}{|v_1|} + X_{i2} \frac{v_2}{|v_1|} \gtreqless 0. \tag{C.1}$$

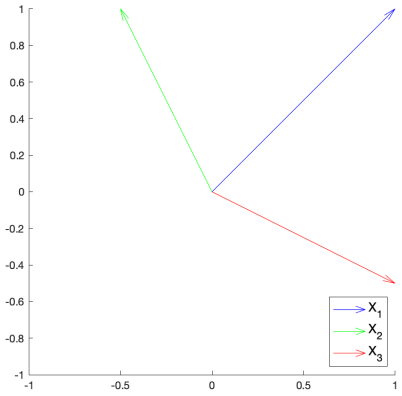
Consequently it suffices in the inequalities of (2.9) and (2.10) to use only values of v with $|v_1| = 1$, and we impose the common scale normalization on the parameter space that $\mathcal{B}$ comprises a compact subset of $\mathbb{R}^2$ such that for all $b \in \mathcal{B}$, $|b_1| = 1$. Geometrically, this means that we can select the required six points for the collection $\mathcal{V}$ from the regions illustrated in Figure 8 panel (d) by focusing solely on values in which the first component is -1 or 1.

For the sake of computing such sets of representative values of v, possibly with much larger n, we may proceed making use of the normalization $|v_1| = 1$ imposed as follows. Recall that only values of v in the interior of the cells need to be considered when constructing $\mathcal{V}$. Values of v satisfying $X_i v = 0$ for some i need not be considered, but they determine the boundaries

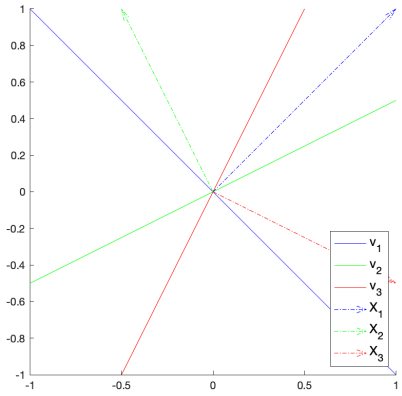
[3] The alternative normalization that $\|v\| = 1$ could also be used.

Figure 8: A schematic illustration of the cell partition V for $n = 3$.

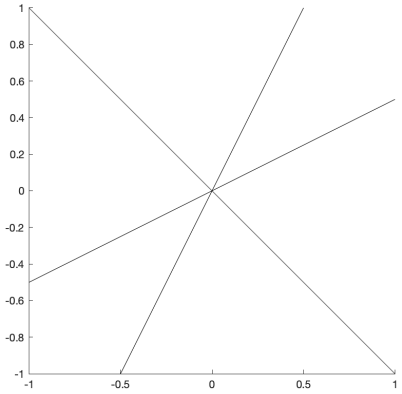
(a) Vectors X_1, X_2, and X_3 in $\mathbb{R}^K$.



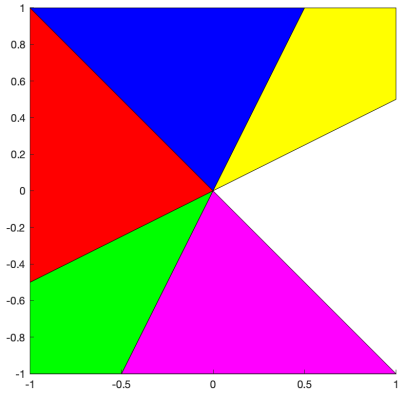
(b) X_1, X_2, X_3 with v_1, v_2, v_3 such that each $X_i v_i = 0$.



(c) Lines defined by those v_1, v_2, v_3 for which $X_i v_i = 0$.



(d) Partitions of $\mathbb{R}^K$ defined by inequalities $X_i v_i \lesseqgtr 0$.



of the cells. These boundary points can be separated into two cases:

$$\text{if } v_1 = +1: \quad X_i v = 0 \iff -\frac{X_{i1}}{X_{i2}} = v_2$$

$$\text{if } v_1 = -1: \quad Xv = 0 \iff \frac{X_{i1}}{X_{i2}} = v_2.$$

Thus membership of a given v in any cell depends only on where v_2 lies relative to the ordered sequence of values of $-Z_i$ if $v_1 = +1$ and those of Z_i if $v_1 = -1$, where $Z_i \equiv \frac{X_{i1}}{X_{i2}}$.[4]

Consequently, a set $\mathcal{V}$ that contains one element from each member of V can be obtained

[4]Here it is to be understood that when $X_{i2} = 0$, Z_i is defined to be $\pm\infty$ according to the sign of X_{i1} if $X_{i1} \neq 0$ and $Z_i = 0$ if $X_{i1} = 0$.

by dividing the real line into intervals according to ordered sequences of values of $-Z_i$ and Z_i, $i = i, ..., n$, and then collecting all pairs $v = (1, v_2)$ such that v_2 lies in the interior of the first sequence of intervals, and all pairs $v = (-1, v_2)$ such that v_2 lies in the interior of the second sequence of intervals. Specifically, let $\vartheta_1 \leq \cdots \leq \vartheta_n$ denote the order statistics of Z_i, so that $\vartheta_1 \equiv \min_i Z_i$ and $\vartheta_n \equiv \max_i Z_i$ and consider the following ordered sequences of intervals:

$$\mathsf{I}_u \equiv \{(-\infty, -\vartheta_n), (-\vartheta_n, -\vartheta_{n-1}), \cdots, (-\vartheta_2, -\vartheta_1), (-\vartheta_1, \infty)\}, \tag{C.2}$$

and

$$\mathsf{I}_l \equiv \{(-\infty, \vartheta_1), (\vartheta_1, \vartheta_2), \cdots, (\vartheta_{n-1}, \vartheta_n), (\vartheta_n, \infty)\}, \tag{C.3}$$

such that I_u and I_l each comprise $n + 1$ open non-overlapping intervals on $\mathbb{R}$. Then $\mathcal{V}$ is constructed by a collection of points $(1, v_2)$ with one value of v_2 taken from each of the intervals of I_u, and points $(-1, v_2)$ with one value of v_2 taken from each of the intervals of I_l.

D Extension to the Case with Endogenous Covariates

In this section, we outline an extension of the test in Section 3 to the case when U satisfies the conditional median restriction given an instrumental variable Z rather than the covariate X. The model in this section is based on Chernozhukov, Hansen, and Jansson (2009), which considers a finite-sample inference on the instrumental variable quantile regression for a continuous outcome variable.

Suppose there are n observed random variables $\{(Y_i, X_i, Z_i) : i = 1, ..., n\}$. We assume there are an unknown parameter β and n unobserved random variables $(U_1, ..., U_n)$ such that $Y_i = 1\{X_i\beta + U_i \geq 0\}$ and that $(Y_1^*, ..., Y_n^*)$ are independent given $(Z_1, ..., Z_n)$ with

$$\mathbb{P}(Y_i^* = 1 \mid Z_1, ..., Z_n) = 1/2 \text{ where } Y_i^* = 1\{U_i \geq 0\}.$$

For this model, we can show the following observable implications:[5]

$$\mathbb{E}[2Y_i \cdot 1\{X_i\beta < 0\} - 1 \mid Z_1, ..., Z_n] \leq 0$$
$$\mathbb{E}[2(1 - Y_i) \cdot 1\{X_i\beta > 0\} - 1 \mid Z_1, ..., Z_n] \leq 0.$$

[5]These implications follow from the inequalities $\mathbb{E}[2Y_i \cdot 1\{X_i\beta < 0\} - 1 \mid Z_1, ..., Z_n] \leq \mathbb{E}[2Y_i^* - 1 \mid Z_1, ..., Z_n]$ and $\mathbb{E}[2(1 - Y_i) \cdot 1\{X_i\beta > 0\} - 1 \mid Z_1, ..., Z_n] \leq \mathbb{E}[2(1 - Y_i^*) - 1 \mid Z_1, ..., Z_n]$. These inequalities are shown as follows. Note that $Y_i \leq Y_i^*$ if $X_i\beta < 0$ and that $Y_i \geq Y_i^*$ if $X_i\beta > 0$. Then $Y_i \cdot 1\{X_i\beta < 0\} \leq Y_i^* \cdot 1\{X_i\beta < 0\} \leq Y_i^*$ and $(1 - Y_i) \cdot 1\{X_i\beta > 0\} \leq (1 - Y_i^*) \cdot 1\{X_i\beta > 0\} \leq 1 - Y_i^*$.

For a given parameter value b, we consider the following test statistic

$$T_n(b) = \max_{g \in \mathcal{G}} \max \left\{ \mathbb{E}_n \left[(2Y \cdot 1\{Xb < 0\} - 1)\, g(Z) \right], \mathbb{E}_n \left[(2(1-Y) \cdot 1\{Xb > 0\} - 1)\, g(Z) \right] \right\},$$

where $\mathcal{G}$ is a set of nonnegative-valued functions of Z. As in Section 3, we do not derive the finite sample distribution of $T_n(b)$ under $\beta = b$. We instead consider the random variable

$$T_n^*(b) = \max_{g \in \mathcal{G}} \max \left\{ \mathbb{E}_n \left[(2Y^* - 1)\, g(Z) \right], \mathbb{E}_n \left[(2(1-Y^*) - 1)\, g(Z) \right] \right\},$$

Note that the finite sample distribution of $T_n^*(b)$ given $(Z_1, ..., Z_n)$ is known since $(2Y_1^* - 1, ..., 2Y_n^* - 1)$ are independent Rademacher random variables conditional on $(Z_1, ..., Z_n)$. Therefore, we can construct $q_{1-\alpha}$ as the conditional $1-\alpha$ quantile of $T_n^*(b)$ given $(Z_1, ..., Z_n)$. Finite sample size control for the test $1\{T_n(b) > q_{1-\alpha}\}$ can be shown in a similar way to Section 3, since it can be shown that $T_n(b) \leq T_n^*(b)$ when $\beta = b$.[6] Note that exogenous variables Z do not enter the model through a linear index, so hyperplane enumeration does not have the same benefit for the construction of instrument functions as in the case where X is exogenous. Alternative instrument functions $g(\cdot)$ such as those of Andrews and Shi (2013) for converting conditional moment inequalities to unconditional ones could be used.

E Additional Results for the Monte Carlo Experiments

In Figure 9, we use the simulation designs of Section 5 and implement the test based on $LR(b)$ in Section 4 as well as the test proposed in Section 3. In Figures 10 – 12, we use the simulation designs of Section 5 and implement our test proposed in Section 3 using the first 500 and 1000 representatives produced by the Rada and Černý (2018) algorithm.

F Mixed Integer Program in Section 6

The following lemma shows that the mixed-integer program (6.2) – (6.5) in Section 6 yields a superset of $\{b_k : b \in \mathcal{B} \text{ and } T_n(b) \leq \bar{cv}\}$, which in turn is a superset of $\{b_k : b \in \mathcal{B} \text{ and } T_n(b) \leq q_{1-\alpha}(b)\}$ in (6.1).

Lemma 2. *Let C be any constant with $\max_{i=1,...,n} |X_i b| < C$. 1. If $T_n(b) \leq \bar{cv}$, then there is a sequence $\{(Z_{ui}, Z_{li}) : i = 1, ..., n\} \in \{0,1\}^{2n}$ such that (6.3)-(6.5) hold. 2. If*

[6] The statement of $T_n(\beta) \leq T_n^*(\beta)$ follows from $Y_i \cdot 1\{X_i\beta < 0\} \leq Y_i^*$ and $(1-Y_i) \cdot 1\{X_i\beta > 0\} \leq 1 - Y_i^*$ as shown in the previous footnote.

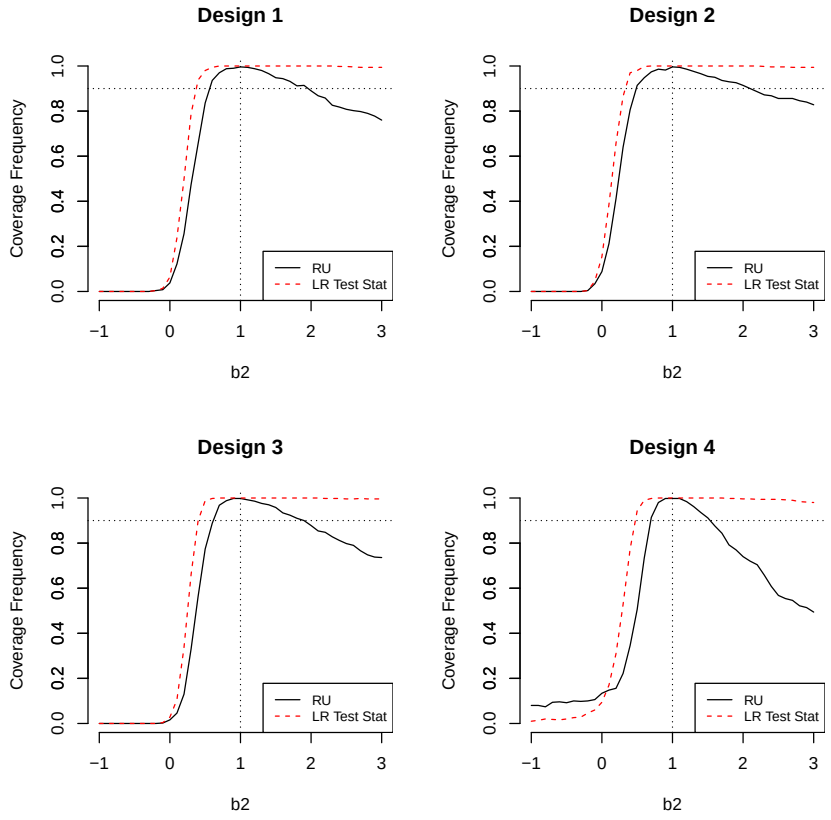


Figure 9: Non-rejection frequencies with $1-\alpha = 90\%$ with $n = 100$ and $K = 2$. RU and LR stands for (1) Rosen and Ura and (2) Likelihood Ratio, respectively.

$X_i b \neq 0$ for every $i = 1, ..., n$, then the converse is also true (i.e., if there is a sequence $\{(Z_{ui}, Z_{li}) : i = 1, ..., n\} \in \{0,1\}^{2n}$ such that (6.3)-(6.5) hold, then $T_n(b) \leq \bar{cv}$).

Proof of Lemma 2. First, we are going to show the first statement. For every $i = 1, ..., n$, define $(Z_{ui}, Z_{li}) = (1\{X_i b > 0\} + 1\{X_i b = 0, Y_i = 1\}, 1\{X_i b < 0\} + 1\{X_i b = 0, Y_i = 0\})$. This choice of (Z_{ui}, Z_{li}) implies (6.3). Moreover, we have

$$\hat{m}_u(b, v) = \mathbb{E}_n[(2Y - 1)1\{Xb \geq 0 > Xv\}] \leq \mathbb{E}_n[(2Y - 1)1\{Xv < 0\}Z_u], \forall v \in \mathcal{V}_u$$

and

$$\hat{m}_l(b, v) = \mathbb{E}_n[(1 - 2Y)1\{Xb \leq 0 < Xv\}] \leq \mathbb{E}_n[(1 - 2Y)1\{Xv > 0\}Z_l], \forall v \in \mathcal{V}_l.$$

The proof of Theorem 4 has shown that the functions in (A.10) and (A.11) are weakly

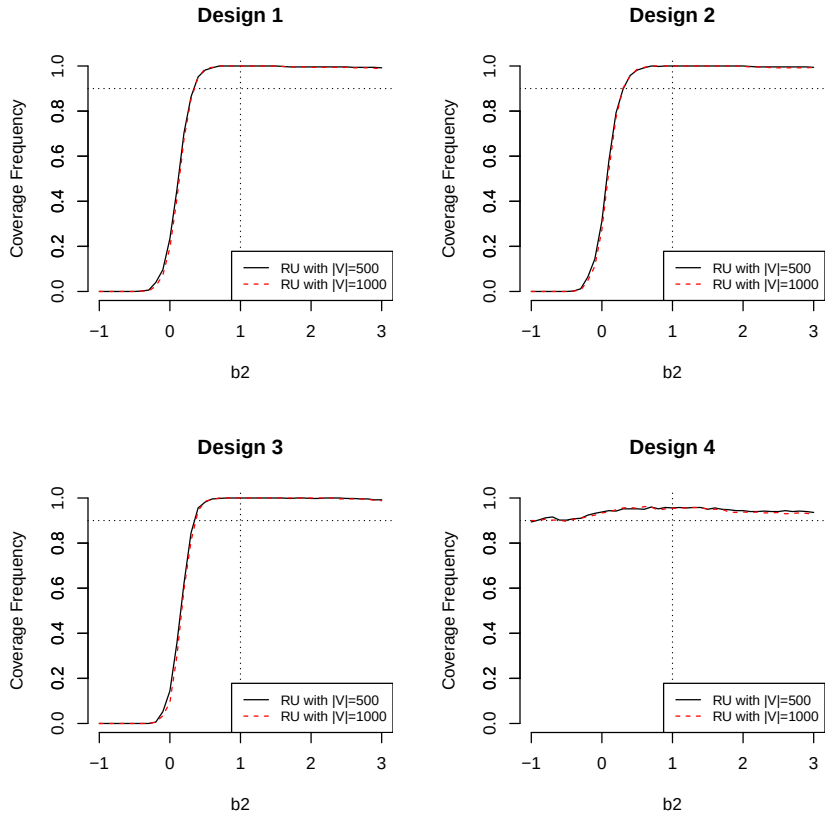


Figure 10: A comparison using 500 and 1000 values of v with $K = 5$ and $n = 100$. RU stands for Rosen and Ura.

decreasing, which implies

$$\sqrt{n}\left(-\frac{\hat{m}_u(b,v)}{\widehat{\sigma}_u(b,v)}\right) \geq \sqrt{n}\left(-\frac{\mathbb{E}_n[(2Y-1)1\{Xv<0\}Z_u]}{\sqrt{\mathbb{E}_n\left[1\{Xb\geq 0>Xv\}\right]-(\mathbb{E}_n[(2Y-1)1\{Xv<0\}Z_u])^2}}\right) \text{ and}$$

$$\sqrt{n}\left(-\frac{\hat{m}_l(b,v)}{\widehat{\sigma}_l(b,v)}\right) \geq \sqrt{n}\left(-\frac{\mathbb{E}_n[(1-2Y)1\{Xv>0\}Z_l]}{\sqrt{\mathbb{E}_n\left[1\{Xb\leq 0<Xv\}\right]-(\mathbb{E}_n[(1-2Y)1\{Xv>0\}Z_l])^2}}\right).$$

Since $T_n(b) = \max\left\{0, \sup_{v\in\mathcal{V}_u}\sqrt{n}\left(-\frac{\hat{m}_u(b,v)}{\widehat{\sigma}_u(b,v)}\right), \sup_{v\in\mathcal{V}_l}\sqrt{n}\left(-\frac{\hat{m}_l(b,v)}{\widehat{\sigma}_l(b,v)}\right)\right\}$, the condition $T_n(b) \leq \bar{cv}$ implies (6.4)-(6.5).

Next, we are going to show the second statement. Assume $X_ib \neq 0$ for every $i = 1, ..., n$. By (6.3), we know $(X_ib > 0 \implies Z_{ui} = 1)$ and $(X_ib < 0 \implies Z_{li} = 1)$. Since $X_ib \neq 0$ and $Z_{ui} + Z_{li} = 1$, it follows that $(Z_{ui}, Z_{li}) = (1\{X_ib \geq 0\}, 1\{X_ib \leq 0\})$. Therefore,

$$\sqrt{n}\left(-\frac{\hat{m}_u(b,v)}{\widehat{\sigma}_u(b,v)}\right) = \sqrt{n}\left(-\frac{\mathbb{E}_n[(2Y-1)1\{Xv<0\}Z_u]}{\sqrt{\mathbb{E}_n\left[1\{Xb\geq 0>Xv\}\right]-(\mathbb{E}_n[(2Y-1)1\{Xv<0\}Z_u])^2}}\right) \text{ and}$$

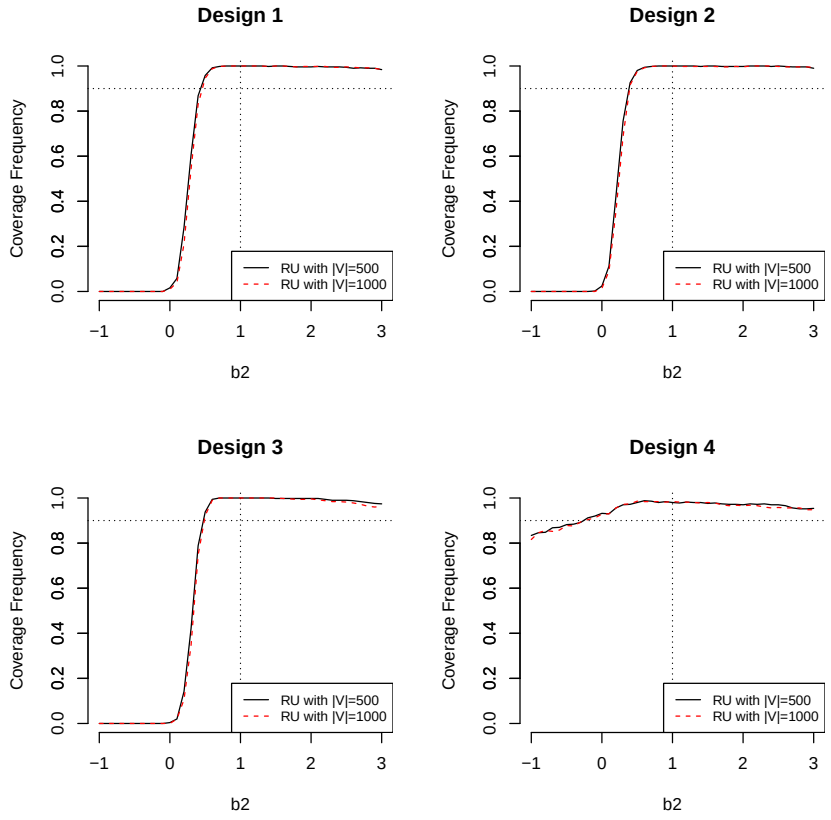


Figure 11: A comparison using 500 and 1000 values of v with $K = 5$ and $n = 250$. RU stands for Rosen and Ura.

$$\sqrt{n}\left(-\frac{\hat{m}_l(b,v)}{\widehat{\sigma}_l(b,v)}\right) = \sqrt{n}\left(-\frac{\mathbb{E}_n[(1-2Y)1\{Xv>0\}Z_l]}{\sqrt{\mathbb{E}_n\left[1\{Xb\leq 0<Xv\}\right]-(\mathbb{E}_n[(1-2Y)1\{Xv>0\}Z_l])^2}}\right).$$

By (6.4)-(6.5), $T_n(b) = \max\left\{0, \sup_{v\in\mathcal{V}_u}\sqrt{n}\left(-\frac{\hat{m}_u(b,v)}{\widehat{\sigma}_u(b,v)}\right), \sup_{v\in\mathcal{V}_l}\sqrt{n}\left(-\frac{\hat{m}_l(b,v)}{\widehat{\sigma}_l(b,v)}\right)\right\} \leq \bar{cv}$. □

G Implications of the conditional median restriction

Assumption 1(v) specifies that $\mathbb{P}(U_i \geq 0|\mathcal{X}_n) = 1/2$ for all i. This delivers the observable implications of Lemma 1 and the representation of the set of conditionally observationally equivalent parameters provided in Theorem 2:

$$\mathcal{B}_n^* = \{b\in\mathcal{B} : \mathbb{E}\left[(2Y_i-1)\,1\{X_ib\geq 0\}\mid\mathcal{X}_n\right]\geq 0\geq\mathbb{E}\left[(2Y_i-1)\,1\{X_ib\leq 0\}\mid\mathcal{X}_n\right] \text{ for all } i\}.$$

In this section, we consider the implications of strengthening Assumption 1 as follows.

Assumption 1'. *All the conditions in Assumption 1 hold, and the cumulative distribution*

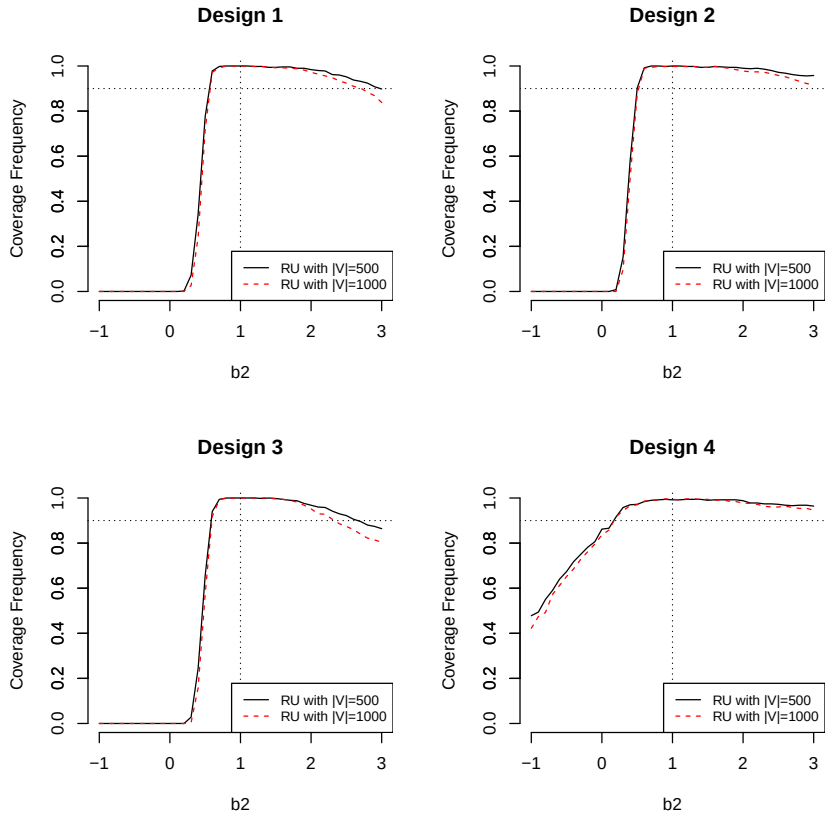


Figure 12: A comparison using 500 and 1000 values of v with $K = 5$ and $n = 1000$. RU stands for Rosen and Ura.

function of U_i given $\mathcal{X}_n$ is strictly increasing on its support for all i.

In addition to the inequalities in the representation of $\mathcal{B}_n^*$ above, Assumption 1' implies

$$\mathbb{E}\left[(2Y_i - 1) \mid \mathcal{X}_n\right] = 0 \iff X_i\beta = 0. \tag{G.1}$$

This holds under the implications given in Manski (1985, display (1) on page 315) and also those in Manski (1975, e.g., page 210 and Section 2.3.1) specialized to binary response.

This strengthening of Assumption 1 can lead to smaller sets of conditionally observationally equivalent parameters than the set $\mathcal{B}_n^*$ when there exist values of X_i such that $X_i\beta = 0$.[7] Under the standard support restrictions invoked in the literature under which point identification is achieved, such values of X_i are realized with probability zero, but they can occur with positive probability when X_i are discretely distributed. However, as we now show, this has little bearing on testing the hypothesis that $\beta = b$. To this end, let $\mathcal{B}_n^{**}$ denote the set of

[7]This distinction was pointed out by an anonymous referee.

conditionally observationally equivalent parameters under Assumption 1' above.[8] Following reasoning similar to the identification analysis of Komarova (2013, Proposition 5.1) with X having a discrete distribution, the set $\mathcal{B}_n^{**}$ can indeed be strictly smaller than $\mathcal{B}_n^*$. That is, when there are values of X_i such that $\mathbb{E}\left[(2Y_i - 1) \mid \mathcal{X}_n\right] = 0$, there can be values of b that belong to $\mathcal{B}_n^* \setminus \mathcal{B}_n^{**}$. Such parameter values b can thus be distinguished from the true parameter value β from perfect knowledge of the joint distribution of $(Y_1, ..., Y_n)$ conditional on $\mathcal{X}_n$. This ideal is however not obtained from sample data for any fixed n. Indeed, for any $b \in \mathcal{B}_n^*$, Theorem 9 establishes that under Assumption 1' any test of (2.2) that achieves finite sample size control rejects $H_0 : \beta = b$ with probability no greater than α, even if $b \notin \mathcal{B}_n^{**}$.

Theorem 9. *Let $\mathcal{G}^\dagger$ denote the subset of $\mathcal{G}$ for which the cumulative distribution function of U_i given $\mathcal{X}_n$ is strictly increasing on its support for all i. Let Assumption 1' hold, let b be any element of $\mathcal{B}_n^*$, and let $\phi(b, Y_1, ..., Y_n; \mathcal{X}_n)$ be any rejection rule that achieves finite sample size control for $H_0 : \beta = b$, i.e.,*

$$\sup_{G \in \mathcal{G}^\dagger} P_{(b,G)}(\phi(b, Y_1, ..., Y_n; \mathcal{X}_n) = 1 \mid \mathcal{X}_n) \leq \alpha. \tag{G.2}$$

Then

$$P_{(\beta,G_0)}(\phi(b, Y_1, ..., Y_n; \mathcal{X}_n) = 1 \mid \mathcal{X}_n) \leq \alpha.$$

Proof of Theorem 9. Let ν be any number in $(0, 1)$. It suffices to show

$$P_{(\beta,G_0)}\left(\phi(b, Y_1, ..., Y_n; \mathcal{X}_n) = 1 \mid \mathcal{X}_n\right) \leq (1-\nu)^{-1} \sup_{G \in \mathcal{G}^\dagger} P_{(b,G)}(\phi(b, Y_1, ..., Y_n; \mathcal{X}_n) = 1 \mid \mathcal{X}_n), \tag{G.3}$$

since $(1-\nu)^{-1} \sup_{G \in \mathcal{G}^\dagger} P_{(b,G)}(\phi(b, Y_1, ..., Y_n; \mathcal{X}_n) = 1 \mid \mathcal{X}_n) \leq (1-\nu)^{-1}\alpha$ by (G.2) and $\nu \in (0, 1)$ can be chosen arbitrarily close to zero. By the definition of $\mathcal{B}_n^*$, there exists a conditional distribution of $(U_1, \cdots, U_n)$ given $\mathcal{X}_n$, say $\tilde{G}$, that satisfies all the requirements of Assumption 1 and is such that

$$P_{(\beta,G_0)}\left(Y_i = y_i \text{ for all } i \mid \mathcal{X}_n\right) = P_{(b,\tilde{G})}\left(Y_i = y_i \text{ for all } i \mid \mathcal{X}_n\right).$$

Define $\tilde{G}^\dagger$ as the mixture distribution

$$\tilde{G}^\dagger = (1 - \nu) \cdot \tilde{G} + \nu \cdot N(0, I),$$

[8] $\mathcal{B}_n^{**}$ is the set of $b \in \mathcal{B}$ for which there exist random variables $\{\tilde{Y}_i : i = 1, ..., n\}$ and $\{\tilde{U}_i : i = 1, ..., n\}$ on $(\Omega, \mathfrak{F}, \mathbb{P})$ such that all the conditions in Definition 1 hold and that the cumulative distribution function of $\tilde{U}_i$ given $\mathcal{X}_n$ is strictly increasing for all i. Assumption 1' could be modified to only require the conditional CDF of U_i to be strictly increasing in a neighborhood of zero and the ensuing analysis would go through unchanged.

where $N(0, I)$ is the n-dimensional standard normal distribution. This definition of $\tilde{G}^\dagger$ ensures that $\tilde{G}^\dagger \in \mathcal{G}^\dagger$. By the mixture structure, we have $P_{(b,\tilde{G}^\dagger)} = (1-\nu)P_{(b,\tilde{G})} + \nu P_{(b,N(0,I))}$. Thus

$$P_{(b,\tilde{G})} \leq (1-\nu)^{-1} P_{(b,\tilde{G}^\dagger)}.$$

Consequently,

$$\begin{aligned}
&P_{(\beta,G_0)}\left(\phi(b, Y_1, ..., Y_n; \mathcal{X}_n) = 1 \mid \mathcal{X}_n\right) \\
&= \sum_{(y_1,...,y_n)\in\{0,1\}^n} \phi(b, y_1, ..., y_n; \mathcal{X}_n) P_{(\beta,G_0)}(Y_i = y_i \text{ for all } i \mid \mathcal{X}_n) \\
&= \sum_{(y_1,...,y_n)\in\{0,1\}^n} \phi(b, y_1, ..., y_n; \mathcal{X}_n) P_{(b,\tilde{G})}\left(Y_i = y_i \text{ for all } i \mid \mathcal{X}_n\right) \\
&\leq (1-\nu)^{-1} \sum_{(y_1,...,y_n)\in\{0,1\}^n} \phi(b, y_1, ..., y_n; \mathcal{X}_n) P_{(b,\tilde{G}^\dagger)}\left(Y_i = y_i \text{ for all } i \mid \mathcal{X}_n\right),
\end{aligned}$$

which implies (G.3). □

The intuition behind Theorem 9 is as follows. For any $b \in \mathcal{B}_n^* \setminus \mathcal{B}_n^{**}$ there exists a distribution, $\tilde{G}$, of $U_1, ..., U_n$ conditional on $\mathcal{X}_n$ such that $P_{(b,\tilde{G})}$ matches the distribution of $Y_1, ..., Y_n$ given $\mathcal{X}_n$ and that satisfies Assumption 1. In particular, under this distribution $\mathbb{P}(U_i \geq 0 | \mathcal{X}_n) = 1/2$ for all i. A distribution $\tilde{G}^\dagger$ that satisfies the stronger restrictions of Assumption 1' can then be constructed by taking a mixture of $\tilde{G}$ with any continuous distribution on $\mathbb{R}^n$ with marginals that have median zero, such as the n-variate standard Gaussian distribution used in the proof. By putting arbitrarily small weight on the continuous component of the mixture, $\tilde{G}^\dagger$ can be constructed to make the rejection probability under $(b, \tilde{G}^\dagger)$ arbitrarily close to α, specifically no more than $(1-\nu)^{-1}\alpha$ for arbitrary choice of $\nu > 0$, while satisfying the stronger requirements of Assumption 1'.