Asymptotically Exact Inference in Conditional Moment Inequality Models

Timothy B. Armstrong*
Yale University

December 10, 2014

Abstract

This paper derives the rate of convergence and asymptotic distribution for a class of Kolmogorov-Smirnov style test statistics for conditional moment inequality models for parameters on the boundary of the identified set under general conditions. Using these results, I propose tests that are more powerful than existing approaches for choosing critical values for this test statistic. I quantify the power improvement by showing that the new tests can detect alternatives that converge to points on the identified set at a faster rate than those detected by existing approaches. A monte carlo study confirms that the tests and the asymptotic approximations they use perform well in finite samples. In an application to a regression of prescription drug expenditures on income with interval data from the Health and Retirement Study, confidence regions based on the new tests are substantially tighter than those based on existing methods.

1 Introduction

Theoretical restrictions used for estimation of economic models often take the form of moment inequalities. Examples include models of consumer demand and strategic interac-

*email: timothy.armstrong@yale.edu. Thanks to Han Hong and Joe Romano for guidance and many useful discussions, and to Liran Einav, Azeem Shaikh, Tim Bresnahan, Guido Imbens, Raj Chetty, Whitney Newey, Victor Chernozhukov, Jerry Hausman, Andres Santos, Elie Tamer, Vicky Zinde-Walsh, Alberto Abadie, Karim Chalak, Xu Cheng, Stefan Hoderlein, Don Andrews, Peter Phillips, Taisuke Otsu, Ed Vytlacil, Xiaohong Chen, Yuichi Kitamura and participants at seminars at Stanford and MIT for helpful comments and criticism. All remaining errors are my own. This paper was written with generous support from a fellowship from the endowment in memory of B.F. Haley and E.S. Shaw through the Stanford Institute for Economic Policy Research.

tions between firms, bounds on treatment effects using instrumental variables restrictions, and various forms of censored and missing data (see, among many others, Manski, 1990; Manski and Tamer, 2002; Pakes, Porter, Ho, and Ishii, 2006; Ciliberto and Tamer, 2009; Chetty, 2010, and papers cited therein). For these models, the restriction often takes the form of moment inequalities conditional on some observed variable. That is, given a sample $(X_1, W_1), \ldots (X_n, W_n)$, we are interested in testing a null hypothesis of the form $E(m(W_i, \theta)|X_i) \geq 0$ with probability one, where the inequality is taken elementwise if $m(W_i, \theta)$ is a vector. Here, $m(W_i, \theta)$ is a known function of an observed random variable W_i, which may include X_i, and a parameter $\theta \in \mathbb{R}^{d_\theta}$, and the moment inequality defines the identified set $\Theta_0 \equiv \{\theta | E(m(W_i, \theta)|X_i) \geq 0 \text{ a.s.}\}$ of parameter values that cannot be ruled out by the data and the restrictions of the model.

In this paper, I consider inference in models defined by conditional moment inequalities. I focus on test statistics that exploit the equivalence between the null hypothesis $E(m(W_i, \theta)|X_i) \geq 0$ almost surely and $Em(W_i, \theta)I(s < X_i < s + t) \geq 0$ for all (s, t). Thus, we can use $\inf_{s,t} \frac{1}{n} \sum_{i=1}^{n} m(W_i, \theta)I(s < X_i < s + t)$, or the infimum of some weighted version of the unconditional moments indexed by (s, t). Following the terminology commonly used in the literature, I refer to these as Kolmogorov-Smirnov (KS) style test statistics. The main contribution of this paper is to derive the rate of convergence and asymptotic distribution of this test statistic for parameters on the boundary of the identified set under a general set of conditions.

While asymptotic distribution results are available for this statistic in some cases (Andrews and Shi, 2013; Kim, 2008), the existing results give only a conservative upper bound of $\sqrt{n}$ on the rate of convergence of this test statistic in a large class of important cases. For example, in the interval regression model, the asymptotic distribution of this test statistic for parameters on the boundary of the identified set and the proper scaling needed to achieve it have so far been unknown in the generic case (see Section 2 for the definition of this model). In these cases, results available in the literature do not give an asymptotic distribution result, but state only that the test statistic converges in probability to zero when scaled up by $\sqrt{n}$. This paper derives the scaling that leads to a nondegenerate asymptotic distribution and characterizes this distribution. Existing results can be used for conservative inference in these cases (along with tuning parameters to prevent the critical value from going to zero), but lose power relative to procedures that use the results derived in this paper to choose critical values based on the asymptotic distribution of the test statistic on the boundary of the identified set.

To quantify this power improvement, I show that using the asymptotic distributions derived in this paper gives power against sequences of parameter values that approach points on the boundary of the identified set at a faster rate than those detected using root-n convergence to a degenerate distribution. Since local power results have not been available for the conservative approach based on root-n approximations in this setting, making this comparison involves deriving new local power results for the existing tests in addition to the new tests. The increase in power is substantial. In the leading case considered in Section 3, I find that the methods developed in this paper give power against local alternatives that approach the identified set at a $n^{-2/(d_X+4)}$ rate (where d_X is the dimension of the conditioning variable), while using conservative $\sqrt{n}$ approximations only gives power against $n^{-1/(d_X+2)}$ alternatives. The power improvements are not completely free, however, as the new tests require smoothness conditions not needed for existing approaches, and are shown to control a weaker notion of size (see the discussion at the end of Section 6). In another paper (Armstrong, 2011, 2014), I propose a modification of this test statistic that achieves a similar power improvement (up to a $\log n$ term) without sacrificing the robustness of the conservative approach (see also the more recent work of Armstrong and Chan 2012 and Chetverikov 2012).

Broadly speaking, the power improvement is related to the tradeoff between bias and variance for nonparametric kernel estimators (see, e.g. Pagan and Ullah, 1999, for an introduction to this topic). Under certain types of null hypotheses, the infimum in the test statistic is taken on a value of (s,t) with $t \to 0$ as the sample size increases. Here, t can be thought of as a bandwidth parameter that is chosen automatically by the test. The asymptotic approximations can be thought of as showing how t is chosen, which allows for less conservative critical values. See Section 2 for more intuition for these results.

To examine how well these asymptotic approximations describe sample sizes of practical importance, I perform a monte carlo study. Confidence regions based on the tests proposed in this paper have close to the nominal coverage in the monte carlos, and shrink to the identified set at a faster rate than those based on existing tests. In addition, I provide an empirical illustration examining the relationship between out of pocket prescription spending and income in a data set in which out of pocket prescription spending is sometimes missing or reported as an interval. Confidence regions for this application constructed using the methods in this paper are substantially tighter than those that use existing methods.

The rest of the paper is organized as follows. The rest of this section discusses the relation of these results to the rest of the literature, and introduces notation and definitions. Section 2

gives a nontechnical exposition of the results, and explains how to implement the procedures proposed in these papers. Together with the statements of the asymptotic distribution results in Section 3 and the local power results in Section 7, this provides a general picture of the results of the paper. Section 5 generalizes the asymptotic distribution results of Section 3, and Sections 4 and 6 deal with estimation of the asymptotic distribution for feasible inference. Section 8 presents monte carlo results. Section 9 presents the empirical illustration. Section 10 concludes. Proofs and other auxiliary material are in the supplementary appendix.

1.1 Related Literature

The results in this paper relate to recent work on testing conditional moment inequalities, including papers by Andrews and Shi (2013), Kim (2008), Khan and Tamer (2009), Chernozhukov, Lee, and Rosen (2009), Lee, Song, and Whang (2011), Ponomareva (2010), Menzel (2008) and Armstrong (2011). The results on the local power of asymptotically exact and conservative KS statistic based procedures derived in this paper are useful for comparing confidence regions based on KS statistics to other methods of inference on the identified set proposed in these papers. Armstrong (2011) derives local power results for some common alternatives to the KS statistics based on integrated moments considered in this paper (the confidence regions considered in that paper satisfy the stronger criterion of containing the entire identified set, rather than individual points, with a prespecified probability).

Out of these existing approaches to inference on conditional moment inequalities, the papers that are most closely related to this one are those by Andrews and Shi (2013) and Kim (2008), both of which consider statistics based on integrating the conditional inequality. As discussed above, the main contributions of the present paper relative to these papers are (1) deriving the rate of convergence and nondegenerate asymptotic distribution of this statistic for parameters on the boundary of the identified set in the common case where the results in these papers reduce to a statement that the statistic converges to zero at a root-n scaling and (2) deriving local power results that show how much power is gained by using critical values based on these new results. Armstrong (2011, 2014) uses a statistic similar to the one considered here, but proposes an increasing sequence of weightings ruled out by the papers above (and the present paper). This leads to almost the same power improvement as the methods in this paper even when conservative critical values are used. This approach has been further explored by Armstrong and Chan (2012) and Chetverikov (2012) (both of these papers were first circulated after the first draft of the present paper).

Khan and Tamer (2009) propose a statistic similar to the one considered here for a model

defined by conditional moment inequalities, but consider point estimates and confidence intervals based on these estimates under conditions that lead to point identification. Galichon and Henry (2009) propose a similar statistic for a class of partially identified models under a different setup. Statistics based on integrating conditional moments have been used widely in other contexts as well, and go back at least to Bierens (1982).

The literature on models defined by finitely many unconditional moment inequalities is more developed, but still recent. Papers in this literature include Andrews, Berry, and Jia (2004), Andrews and Jia (2008), Andrews and Guggenberger (2009), Andrews and Soares (2010), Chernozhukov, Hong, and Tamer (2007), Romano and Shaikh (2010), Romano and Shaikh (2008), Bugni (2010), Beresteanu and Molinari (2008), Moon and Schorfheide (2009), Imbens and Manski (2004) and Stoye (2009) and many others.

1.2 Notation

I use the following notation in the rest of the paper. For observations $(X_1, W_1), \ldots, (X_n, W_n)$ and a measurable function h on the sample space, $E_n h(X_i, W_i) \equiv \frac{1}{n}\sum_{i=1}^{n} h(X_i, W_i)$ denotes the sample mean. I use double subscripts to denote elements of vector observations so that $X_{i,j}$ denotes the jth component of the ith observation X_i. Inequalities on Euclidean space refer to the partial ordering of elementwise inequality. For a vector valued function $h : \mathbb{R}^\ell \to \mathbb{R}^m$, the infimum of h over a set T is defined to be the vector consisting of the infimum of each element: $\inf_{t\in T} h(t) \equiv (\inf_{t\in T} h_1(t), \ldots, \inf_{t\in T} h_m(t))$. I use $a \wedge b$ to denote the elementwise minimum and $a \vee b$ to denote the elementwise maximum of a and b. The notation $\lceil x \rceil$ denotes the least integer greater than or equal to x.

2 Overview of Results

This section gives a description of the main results at an intuitive level, and gives step-by-step instructions for one of the tests proposed in this paper. Section 2.1 defines the terms "asymptotically exact" and "asymptotically conservative" for the purposes if this paper, and explains how the results in this paper lead to asymptotically exact inference. Section 2.2 describes the asymptotic distribution result, and explains why the situations that lead to it are important in practice. Section 2.3 describes the reason for the power improvement. Section 2.4 gives instructions for implementing the test.

2.1 Asymptotically Exact vs Conservative Inference

Throughout this paper, I use the terms asymptotically exact and asymptotically conservative to refer to the behavior of tests for a fixed parameter value under a fixed probability distribution.

Definition 1. *For a probability distribution P and a parameter θ with θ satisfying the null hypothesis under P, a test is called asymptotically exact for (θ, P) if the probability of rejecting θ converges to the nominal level as the number of observations increases to infinity under P. A test is called asymptotically conservative for (θ, P) if the probability of falsely rejecting θ is asymptotically strictly less than the nominal level under P.*

Note that this definition depends on the data generating process and parameter being tested, and contrasts with a definition where a test is conservative only if the size of the test is less than the nominal size taken as the supremum of the probability of rejection over a composite null of all possible values of θ and P such that θ is in the identified set under P. This facilitates discussion of results like the ones in this paper (and other papers that deal with issues related to moment selection) that characterize the behavior of tests for different values of θ in the identified set.

As described above, the asymptotic distribution results used by Andrews and Shi (2013) and Kim (2008) reduce to a statement that $\sqrt{n}T_n(\theta) \xrightarrow{p} 0$ for certain data generating processes and parameter values on the identified set, where $T_n(\theta)$ is the test statistic described above. For such (θ, P), the procedures in those papers are asymptotically equivalent to rejecting when $\sqrt{n}T_n(\theta)$ is greater than some user specified parameter η, which leads to the procedure rejecting with probability approaching one and therefore being asymptotically conservative at such (θ, P) according to the above definition. The present paper derives asymptotic distribution results of the form $n^\delta T_n(\theta) \xrightarrow{d} Z$ for a nondegenerate limiting variable Z, where n^δ is a scaling with $\delta > 1/2$. Comparing $n^\delta T_n(\theta)$ to a critical value c_α derived from such an approximation then leads to asymptotically exact inference, and an increase in power at nearby alternatives relative to the asymptotically conservative procedure, since $T_n(\theta)$ is compared to $n^{-\delta}c_\alpha$ rather than $n^{-1/2}\eta$.

2.2 Asymptotic Distribution

The asymptotic distributions derived in this paper arise when the conditional moment inequality binds only on a probability zero set. This leads to a faster than root-n rate of

convergence to an asymptotic distribution that depends entirely on moments that are close to, but not quite binding.

To see why this case is typical in applications, consider an application of moment inequalities to regression with interval data. In the interval regression model, $E(W_i^*|X_i) = X_i'\beta$, and W_i^* is unobserved, but known to be between observed variables W_i^H and W_i^L, so that β satisfies the moment inequalities

$$E(W_i^L|X_i) \leq X_i'\beta \leq E(W_i^H|X_i).$$

Suppose that the distribution of X_i is absolutely continuous with respect to the Lebesgue measure. Then, to have one of these inequalities bind on a positive probability set, $E(W_i^L|X_i)$ or $E(W_i^H|X_i)$ will have to be linear on this set. Even if this is the case, this only means that the moment inequality will bind on this set for one value of β, and the moment inequality will typically not bind when applied to nearby values of β on the boundary of the identified set. Figures 1 and 2 illustrate this for the case where the conditioning variable is one dimensional. Here, the horizontal axis is the nonconstant part of x, and the vertical axis plots the conditional mean of the W_i^H along with regression functions corresponding to points in the identified set. Figure 1 shows a case where the KS statistic converges at a faster than root-n rate. In Figure 2, the parameter β_1 leads to convergence at exactly a root-n rate, but this is a knife edge case, since the KS statistic for testing β_2 will converge at a faster rate (note, however, that a formulation of the above interval regression model based on unconditional moments leads to the familiar root-n rate in all cases; see Bontemps, Magnac, and Maurin, 2012).

This paper derives asymptotic distributions under conditions that generalize these cases to arbitrary moment functions $m(W_i, \theta)$. In this broader setting, KS statistics converge at a faster than root-n rate on the boundary of the identified set under general conditions when the model is set identified and at least one conditioning variable is continuously distributed. See Armstrong (2011) for primitive conditions for a set of high-level conditions similar to the ones used in this paper for some of these models.

The rest of this section describes the results in the context of the interval regression example in a particular case. Consider deriving the rate of convergence and nondegenerate asymptotic distribution of the KS statistic for a parameter β like the one shown in Figure 1, but with X_i possibly containing more than one covariate. Since the lower bound never binds, it is intuitively clear that the KS statistic for the lower bound will converge to zero at a faster rate than the KS statistic for the upper bound, so consider the KS statistic for the

upper bound given by $\inf_{s,t} E_n Y_i I(s < X_i < s+t)$ where $Y_i = W_i^H - X_i'\beta$. If $E(W_i^H|X_i = x)$ is tangent to $x'\beta$ at a single point x_0, and $E(W_i^H|X_i = x)$ has a positive second derivative matrix V at this point, we will have $E(Y_i|X_i = x) \approx (x-x_0)'V(x-x_0)$ near x_0, so that, for s near x_0 and t close to zero, $EY_iI(s < X_i < s+t) \approx f_X(x_0) \int_{s_1}^{s_1+t_1} \cdots \int_{s_{d_X}}^{s_{d_X}+t_{d_X}} (x-x_0)'V(x-x_0)\, dx_{d_X} \cdots dx_1$ (here, if the regression contains a constant, the conditioning variable X_i is redefined to be the nonconstant part of the regressor, so that d_X refers to the dimension of the nonconstant part of X_i).

Since $EY_iI(s < X_i < s+t) = 0$ only when $Y_iI(s < X_i < s+t)$ is degenerate, the asymptotic behavior of the KS statistic should depend on indices (s,t) where the moment inequality is not quite binding, but close enough to binding that sampling error makes $E_nY_iI(s < X_i < s+t)$ negative some of the time. To determine on which indices (s,t) we should expect this to happen, split up $E_nY_iI(s < X_i < s+t)$ into a mean zero term and a drift term: $(E_n-E)Y_iI(s < X_i < s+t)+EY_iI(s < X_i < s+t)$. In order for this to be strictly negative some of the time, there must be non-negligible probability that the mean zero term is greater in absolute value than the drift term. That is, we must have $sd((E_n - E)Y_iI(s < X_i < s+t))$ of at least the same order of magnitude as $EY_iI(s < X_i < s+t)$. We have $sd((E_n - E)Y_iI(s < X_i < s+t)) = \mathcal{O}(\sqrt{\prod_i t_i}/\sqrt{n})$ for small t, and some calculations show that, for s close to x_0, $EY_iI(s < X_i < s+t) \approx f_X(x_0) \int_{s_1}^{s_1+t_1} \cdots \int_{s_{d_X}}^{s_{d_X}+t_{d_X}} (x-x_0)'V(x-x_0)\, dx_{d_X} \cdots dx_1 \geq C\|(s-x_0,t)\|^2 \prod_i t_i$ for some $C > 0$. Thus, we expect the asymptotic distribution to depend on (s,t) such that $\sqrt{\prod_i t_i}/\sqrt{n}$ is of the same or greater order of magnitude than $\|(s-x_0,t)\|^2 \prod_i t_i$, which corresponds to $\|(s-x_0,t)\|^2\sqrt{\prod_i t_i}$ less than or equal to $\mathcal{O}(1/\sqrt{n})$.

Assuming that $s-x_0$ and all elements of t are of the same order of magnitude (which turns out to be the case), this condition leads to $\|(s-x_0,t)\|^{2+d_X/2} \leq \mathcal{O}(1/\sqrt{n})$, and rearranging gives $\|(s-x_0,t)\| \leq \mathcal{O}(n^{-1/(d_X+4)})$. This leads to both $sd((E_n-E)Y_iI(s < X_i < s+t))$ (which behaves like $\sqrt{\prod_i t_i}/\sqrt{n}$) and $EY_iI(s < X_i < s+t)$ (which behaves like $\|(s-x_0,t)\|^2 \prod_i t_i$) being of order $\mathcal{O}(n^{-(d_X+2)/(d_X+4)})$.

Thus, we should expect that the values of (s,t) that matter for the asymptotic distribution of the KS statistic are those with $(s-x_0,t)$ of order $n^{-1/(d_X+4)}$, and that the KS statistic will converge in distribution to a nondegenerate limiting distribution when scaled up by $n^{-(d_X+2)/(d_X+4)}$. The results in this paper show this formally, and the proofs follow the above intuition, using additional arguments to show that the approximations hold uniformly over (s,t).

2.3 Local Power

To get an idea of the accuracy of the resulting confidence intervals, we can consider power against alternative parameter values β_n that approach the boundary of the identified set as the sample size increases. If our test detects all sequences β_n converging to the boundary of the identified set at a particular rate, this should be the rate at which the confidence region shrinks toward the identified set. To this end, let β be the parameter pictured in Figure 1, and let β_n be obtained by adding a scalar a_n to the intercept term of β (the results are similar for the slope parameters, but the intercept term leads to simpler calculatiosn).

In order for our test to reject with high probability, we need the test statistic to be greater in magnitude than the $\mathcal{O}(n^{-(d_X+2)/(d_X+4)})$ critical value. To see when this will happen, we can go through the calculations above, but with $Y_i = W_i^H - X_i'\beta_n = W_i^H - X_i'\beta - a_n$, rather than $W_i^H - X_i'\beta$. The calculations are similar, except that the drift term is now $E(W_i^H - X_i'\beta - a_n)I(s < X_i < s+t) \approx \|s - x_0, t\|^2 \prod_i t_i - a_n \prod_i t_i$. This expression is minimized when $s = x_0$, the components of t_i are equal and $a_n \approx \|t\|^2$. Plugging this back in, we see that the minimized drift term goes to zero at the same rate as $-a_n^{(2+d_X)/2}$. Thus, we should have high power when $a_n^{(2+d_X)/2}$ is large in magnitude relative to the $\mathcal{O}(n^{-(d_X+2)/(d_X+4)})$ critical value, which can be rearranged to give $a_n \geq \mathcal{O}(n^{-2/(d_X+4)})$.

Now consider a test using the critical value of Andrews and Shi (2013) or Kim (2008), which decreases at a slower $\mathcal{O}(n^{-1/2})$ rate. By the same calculations, we now compare the same $\mathcal{O}\left(a_n^{(2+d_X)/2}\right)$ drift term to a $\mathcal{O}(n^{-1/2})$ critical value, so that we obtain nontrivial power only when $a_n \geq \mathcal{O}(n^{1/(d_X+2)})$, which contrasts with the faster $\mathcal{O}(n^{-2/(d_X+4)})$ rate obtained by the new procedure introduced in the present paper. This is shown formally in Theorems 7.1 and 7.2.

2.4 Implementation of the Procedure

For convenience, I describe the implementation of one of the tests proposed in this paper, which uses these asymptotic distribution results to achieve the power improvements described above. Section 6 states formal conditions under which the test controls the probability of false rejection asymptotically, and gives a more detailed explanation for why the test works. See Section B of the supplementary appendix for a procedure that obtains critical values in a different way.

Let

$$T_n(\theta) = \inf_{s,t} E_n m(W_i, \theta) = (\inf_{s,t} E_n m_1(W_i, \theta), \ldots, \inf_{s,t} E_n m_{d_Y}(W_i, \theta)),$$

and let $S : \mathbb{R}^{d_Y} \to \mathbb{R}$ be a nonincreasing function of each component (so that $S(t)$ is positive and large in magnitude when the elements of t are negative and large in magnitude). The test compares $S(T_n(\theta))$ to a critical value based on subsampling, a generic resampling procedure for estimating the distribution of a test statistic. Since the asymptotic distribution and rate of convergence depend on the data generating process (with a $\sqrt{n}$ rate or $n^{(d_X+2)/(d_X+4)}$ in the two situations described in Section 2.2), the procedure uses a modification of a method for subsampling with unknown rates of convergence due to Bertail, Politis, and Romano (1999).

For a set of indices $\mathcal{S} \subseteq \{1, \ldots, n\}$, define $T_{\mathcal{S}}(\theta) = \inf_{s,t} \frac{1}{|\mathcal{S}|} \sum_{i \in \mathcal{S}} m(W_i, \theta)$, so that $S(T_{\mathcal{S}}(\theta))$ is the test statistic formed with the subsample $\mathcal{S}$. For a sequence τ_n, define

$$L_{n,b}(x|\tau) \equiv \frac{1}{\binom{n}{b}} \sum_{|\mathcal{S}|=b} I(\tau_b[S(T_{\mathcal{S}}(\theta)) - S(T_n(\theta))] \leq x)$$

and

$$\tilde{L}_{n,b}(x|\tau) \equiv \frac{1}{\binom{n}{b}} \sum_{|\mathcal{S}|=b} I(\tau_b S(T_{\mathcal{S}}(\theta)) \leq x).$$

Let $L_{n,b}(x|1) \equiv \frac{1}{\binom{n}{b}} \sum_{|\mathcal{S}|=b} I(S(T_{\mathcal{S}}(\theta)) - S(T_n(\theta)) \leq x)$, and let $L_{n,b}^{-1}(t|\tau) = \inf\{x | L_{n,b}(x|\tau) \geq t\}$ be the tth quantile of $L_{n,b}(x|\tau)$, and define $L_{n,b}^{-1}(t|1)$ similarly. $L_{n,b}(x|\tau)$ can be interpreted as a subsampling based estimate of the distribution of $\tau_n S(T_{\mathcal{S}}(\theta))$, computed under the assumption that τ_n is the rate of convergence of $S(T_{\mathcal{S}}(\theta))$.

With this notation, the test is defined as follows, for a nominal level α.

1. Let $b_1 = \lceil n^{\chi_1} \rceil$ and $b_2 = \lceil n^{\chi_2} \rceil$ for some $1 > \chi_1 > \chi_2 > 0$, and let $t_1, t_2, \ldots, t_{n_t} \in (0, 1)$. Let

$$\hat{\beta} = \frac{\frac{1}{n_t} \sum_{k=1}^{n_t} \left[L_{n,b_2}^{-1}(t_k|1) - L_{n,b_1}^{-1}(t_k|1) \right]}{\log b_1 - \log b_2}. \tag{1}$$

   Let $1 > \chi_a > 0$, and let c be a positive integer. Let $\hat{\beta}_a$ be defined the same way as $\hat{\beta}$, but with b_2 given by $\lceil n^{\chi_a} \rceil$ and b_1 given by c.

2. Let $\underline{\gamma}$ and $\overline{\gamma}$ be real numbers with $0 < \underline{\gamma} < \overline{\gamma} < \infty$, and define $\underline{\beta} = (d_X + \overline{\gamma})/(d_X + 2\overline{\gamma})$ and $\overline{\beta} = (d_X + \underline{\gamma})/(d_X + 2\underline{\gamma})$. Let $b = n^{\chi_3}$ for some $0 < \chi_3 < 1$, and let $\eta > 0$.

   (a) If $\hat{\beta}_a \geq \underline{\beta}$, reject if $n^{(\hat{\beta}\wedge\overline{\beta})\vee(1/2)} S(T_n(\theta)) > \tilde{L}_{n,b}(1-\alpha|b^{(\hat{\beta}\wedge\overline{\beta})\vee(1/2)})$.

   (b) If $\hat{\beta}_a < \underline{\beta}$, reject if $n^{1/2} S(T_n(\theta)) > \tilde{L}_{n,b}(1-\alpha|b^{1/2}) + \eta$.

3. Perform this test for each value of θ, and report $\mathcal{C} = \{\theta|\text{fail to reject } \theta\}$ as a confidence region for θ.

Theorem 6.1 gives conditions on θ and the data generating process such that this test is asymptotically exact or conservative. Under regularity conditions, the test is asymptotically exact in situations like the one described in Section 2.2, and achieves the power improvement described in Section 2.3. The quantities $\hat{\beta}$ and $\hat{\beta}_a$ in step 1 are estimates of the exponent in the rate of convergence. Step 2 uses a pre-test based on $\hat{\beta}_a$ to distinguish between the cases of root-n convergence and $n^{(d_X+2)/(d_X+4)}$ convergence described in Section 2.2, and other rates derived in Section 5, and uses a truncated version of $\hat{\beta}$. Section 6 describes the reasoning behind these choices in more detail.

Since $\binom{n}{b}$ is large even for moderate choices of b, computing $L_{n,b}(x|\tau)$ can be computationally prohibitive. To overcome this, let B_n be a sequence tending to ∞ with n, and let $\mathcal{S}_1, \mathcal{S}_2, \ldots \mathcal{S}_{\binom{n}{b}}$ be the $\binom{n}{b}$ subsets of $\{1, \ldots n\}$ size b. Let $i_1, \ldots, i_{B_n}$ be drawn randomly from $1, \ldots \binom{n}{b}$ (with or without replacement). Then $L_{n,b}(x|\tau)$ can be replaced with $\frac{1}{B_n}\sum_{k=1}^{B_n} I(\tau_b[S(T_{\mathcal{S}_{i_k}}(\theta)) - S(T_n(\theta))] \leq x)$, and similarly for the other quantities (see Politis, Romano, and Wolf, 1999, Corollary 2.4.1). In forming a confidence region (step 3), it is important that the same replications $i_1, \ldots, i_{B_n}$ be used for each θ.

This procedure depends on several user defined parameters. For these, I recommend $S(t) = \max_{1\leq k\leq d_Y}(-t_k \vee 0)$, $\chi_1 = 1/2$, $\chi_2 = 1/3$, $\chi_a = 1/2$, $c = 5$, $n_t = 3$, $t_1 = .5$, $t_2 = .9$, $t_3 = .95$, $\underline{\beta} = [(d_X + 2)/(d_X + 4) + 1/2]/2$, $\overline{\beta} = (d_X + 1)/(d_X + 2)$ and $\eta = .001$, since these values have been tested in monte carlos and appear to perform well. However, these choices are not required for the validity of the procedure (although $S(t)$ must satisfy certain regularity conditions given in Theorem 6.1). One can also modify the procedure by replacing $\tilde{L}_{n,b}$ with $L_{n,b}$ in step 2, or by replacing the test in step 2b with one of the tests proposed by Andrews and Shi (2013) or Kim (2008).

3 Asymptotic Distribution of the KS Statistic

Given iid observations $(X_1, W_1), \ldots, (X_n, W_n)$, of random variables $X_i \in \mathbb{R}^{d_X}$, $W_i \in \mathbb{R}^{d_W}$, we wish to test the null hypothesis that $E(m(W_i, \theta)|X_i) \geq 0$ almost surely, where $m : \mathbb{R}^{d_W} \times \Theta \to \mathbb{R}^{d_Y}$ is a known measurable function and $\theta \in \Theta \subseteq \mathbb{R}^{d_\theta}$ is a fixed parameter value. I use the notation $\bar{m}(\theta, x)$ to denote $E(m(W_i, \theta)|X = x)$. In some cases when it is clear which parameter value is being tested, I will define $Y_i = m(W_i, \theta)$ for notational convenience. Defining Θ_0 to be the identified set of values of θ in Θ that satisfy $E(m(W_i, \theta)|X_i) \geq 0$ almost surely, these tests can then be inverted to obtain a confidence region that, for every $\theta_0 \in \Theta_0$, contains θ_0 with a prespecified probability (see Imbens and Manski, 2004). The tests considered here will be based on asymptotic approximations, so that these statements will only hold asymptotically.

The results in this paper allow for asymptotically exact inference using KS style statistics in cases where the $\sqrt{n}$ approximations for these statistics are degenerate. This includes the case described in the introduction in which one component of $E(m(W_i, \theta)|X_i)$ is tangent to zero at a single point and the rest are bounded away from zero. While this case captures the essential intuition for the results in this paper, I state the results in a slightly more general way in order to make them more broadly applicable. I allow each component of $E(m(W_i, \theta)|X)$ to be tangent to zero at finitely many points, which may be different for each component.

I consider KS style statistics that are a function of $\inf_{s,t} E_n m(W_i, \theta) I(s < X_i < s + t) = (\inf_{s,t} E_n m_1(W_i, \theta) I(s < X_i < s + t), \ldots, \inf_{s,t} E_n m_{d_Y}(W_i, \theta) I(s < X_i < s + t))$. Fixing some function $S : \mathbb{R}^{d_Y} \to \mathbb{R}_+$, we can then reject for large values of $S(\inf_{s,t} E_n m(W_i, \theta) I(s < X_i < s+t))$ (which correspond to more negative values of the components of $\inf_{s,t} E_n m(W_i, \theta) I(s < X_i < s+t)$ for typical choices of S). Although the moments $E_n m(W_i, \theta) I(s < X_i < s+t)$ are not weighted, but the results could be extended to allow for a weighting function $\omega(s, t)$, so that the infimum is over $\omega(s, t) E_n m(W_i, \theta) I(s < X_i < s + t)$ as long as $\omega(s, t)$ is smooth and bounded away from zero and infinity. The condition that the weight function be bounded uniformly in the sample size, which is also imposed by Andrews and Shi (2013) and Kim (2008), turns out to be important (see Armstrong, 2011).

I formalize the notion that θ is at a point in the identified set such that one or more of the components of $E(m(W_i, \theta)|X_i)$ is tangent to zero at a finite number of of points in the following assumption.

Assumption 3.1. *For some version of $E(m(W_i, \theta)|X_i)$, the conditional mean of each ele-*

ment of $m(W_i, \theta)$ *takes its minimum only on a finite set* $\{x | E(m_j(W_i, \theta)|X = x) = 0 \text{ some } j\} = \mathcal{X}_0 = \{x_1, \ldots, x_\ell\}$. *For each* k *from* 1 *to* ℓ, *let* $J(k)$ *be the set of indices* j *for which* $E(m_j(W_i, \theta)|X = x_k) = 0$. *Assume that there exist neighborhoods* $B(x_k)$ *of each* $x_k \in \mathcal{X}_0$ *such that, for each* k *from* 1 *to* ℓ, *the following assumptions hold.*

i.) $E(m_j(W_i, \theta)|X_i)$ *is bounded away from zero outside of* $\cup_{k=1}^{\ell} B(x_k)$ *for all* j *and, for* $j \notin J(k)$, $E(m_j(W_i, \theta)|X_i)$ *is bounded away from zero on* $B(x_k)$.

ii.) *For* $j \in J(k)$, $x \mapsto E(m_j(W_i, \theta)|X = x)$ *has continuous second derivatives inside of the closure of* $B(x_k)$ *and a positive definite second derivative matrix* $V_j(x_k)$ *at each* x_k.

iii.) X *has a continuous density* f_X *on* $B(x_k)$.

iv.) *Defining* $m_{J(k)}(W_i, \theta)$ *to have* j*th component* $m_j(W_i, \theta)$ *if* $j \in J(k)$ *and* 0 *otherwise,* $x \mapsto E(m_{J(k)}(W_i, \theta)m_{J(k)}(W_i, \theta)'|X_i = x)$ *is finite and continuous on* $B(x_k)$ *for some version of this conditional second moment matrix.*

Note that the assumption that X_i has a density at certain points means that the moment inequalities must be defined so that X_i does not contain a constant. Thus, the results stated below hold in the interval regression example with d_X equal to the number of nonconstant regressors.

Unless otherwise stated, I assume that the contact set $\mathcal{X}_0$ in Assumption 3.1 is nonempty. If Assumption 3.1 holds with $\mathcal{X}_0$ empty so that the conditional mean $\bar{m}(\theta, x)$ is bounded from below away from zero, θ will typically be on the interior of the identified set (as long as the conditional mean stays bounded away from zero when θ is moved a small amount). For such values of θ, KS statistics will converge at a faster rate (see Lemma C.5 in the supplementary appendix), leading to conservative inference even if the rates of convergence derived under Assumption 3.1, which are faster than $\sqrt{n}$, are used.

In addition to imposing that the minimum of the components of the conditional mean $\bar{m}(\theta, x)$ over x are taken on a probability zero set, Assumption 3.1 requires that this set be finite, and that $\bar{m}(\theta, x)$ behave quadratically in x near this set. I state results under this condition first, since it is easy to interpret as arising from a positive definite second derivative matrix at the minimum. Section 5 generalizes these results to other shapes of the conditional mean, and Section 6 proposes procedures that embed pre-tests for these conditions.

The next assumption is a regularity condition that bounds $m_j(W_i, \theta)$ by a nonrandom constant. This assumption will hold naturally in models based on quantile restrictions. In the interval regression example, it requires that the data have finite support.

Assumption 3.2. *For some nonrandom $\overline{Y} < \infty$, $|m_j(W_i, \theta)| \leq \overline{Y}$ with probability one for each j.*

Finally, I make the following assumption on the function S. Part of this assumption could be replaced by weaker smoothness conditions, but the assumption covers $x \mapsto \|x\|_- \equiv \|x \wedge 0\|$ for $\|t\| = (\sum_{k=1}^{d_Y} t_k^p)^{1/p}$ for any $p \geq 1$ or $\|t\| = \max_k |t_k|$, which should suffice for practical purposes.

Assumption 3.3. *$S : \mathbb{R}^{d_Y} \to \mathbb{R}_+$ is continuous and nonincreasing and satisfies $S(ax) = aS(x)$ for any nonnegative scalar a.*

The following theorem gives the asymptotic distribution and rate of convergence for $\inf_{s,t} E_n m(W_i, \theta) I(s < X_i < s + t)$ under these conditions. The asymptotic distribution of $S(\inf_{s,t} E_n m(W_i, \theta) I(s < X_i < s + t))$ under mild conditions on S then follows as an easy corollary.

Theorem 3.1. *Under Assumptions 3.1 and 3.2,*

$$n^{(d_X+2)/(d_X+4)} \inf_{s,t} E_n m(W_i, \theta) I(s < X_i < s + t) \xrightarrow{d} Z$$

where Z is a random vector on $\mathbb{R}^{d_Y}$ defined as follows. Let $\mathbb{G}_{P,x_k}(s,t)$, $k = 1, \ldots, \ell$ be independent mean zero Gaussian processes with sample paths in the space $C(\mathbb{R}^{2d_X}, \mathbb{R}^{d_Y})$ of continuous functions from $\mathbb{R}^{2d_X}$ to $\mathbb{R}^{d_Y}$ and covariance kernel

$$cov(\mathbb{G}_{P,x_k}(s,t), \mathbb{G}_{P,x_k}(s',t')) = E(m_{J(k)}(W_i,\theta) m_{J(k)}(W_i,\theta)' | X_i = x_k) f_X(x_k) \int_{s \vee s' < x < (s+t) \wedge (s'+t')} dx$$

where $m_{J(k)}(W_i, \theta)$ is defined to have jth element equal to $m_j(W_i, \theta)$ for $j \in J(k)$ and equal to zero for $j \notin J(k)$. For $k = 1, \ldots, \ell$, let $g_{P,x_k} : \mathbb{R}^{2d_X} \to \mathbb{R}^{d_Y}$ be defined by

$$g_{P,x_k,j}(s,t) = \frac{1}{2} f_X(x_k) \int_{s_1}^{s_1+t_1} \cdots \int_{s_{d_X}}^{s_{d_X}+t_{d_X}} x' V_j(x_k) x \, dx_{d_X} \cdots dx_1$$

for $j \in J(k)$ and $g_{x_k,j}(s,t) = 0$ for $j \notin J(k)$. Define Z to have jth element

$$Z_j = \min_{k \text{ s.t. } j \in J(k)} \inf_{(s,t) \in \mathbb{R}^{2d_X}} \mathbb{G}_{P,x_k,j}(s,t) + g_{P,x_k,j}(s,t).$$

The asymptotic distribution of $S(\inf_{s,t} E_n m(W_i, \theta) I(s < X_i < s+t))$ follows immediately from this theorem.

Corollary 3.1. *Under Assumptions 3.1, 3.2, and 3.3,*

$$n^{(d_X+2)/(d_X+4)} S(\inf_{s,t} E_n m(W_i, \theta) I(s < X_i < s+t)) \xrightarrow{d} S(Z)$$

for a random variable Z with the distribution given in Theorem 3.1.

These results will be useful for constructing asymptotically exact level α tests if the asymptotic distribution does not have an atom at the $1-\alpha$ quantile, and if the quantiles of the asymptotic distribution can be estimated. The next section treats estimation of the asymptotic distribution under Assumption 3.1, and shows that the distribution is indeed continuous. Since the asymptotic distribution and rate of convergence are different depending on the shape of the conditional mean, the tests in Section 4 need to be embeeded in a procedure with pre-tests to see whether Assumption 3.1 or some other condition best describes the data generating process. Section 5 extends Theorem 3.1 to other shapes of the conditional mean, and Section 6 uses these results to give conditions for the validity of the procedure in Section 2.4, which includes such a pre-test.

4 Inference

To ensure that the asymptotic distribution is continuous, we need to impose additional assumptions to rule out cases where components of $m(W_j, \theta)$ are degenerate. The next assumption rules out these cases.

Assumption 4.1. *For each k from 1 to ℓ, letting $j_{k,1}, \ldots, j_{k,|J(k)|}$ be the elements in $J(k)$, the matrix with q,rth element given by $E(m_{j_{k,q}}(W_i, \theta) m_{j_{k,r}}(W_i, \theta) | X_i = x_k)$ is invertible.*

This assumption simply says that the binding components of $m(W_i, \theta)$ have a nonsingular conditional covariance matrix at the point where they bind. A sufficient condition for this is for the conditional covariance matrix of $m(W_i, \theta)$ given X_i to be nonsingular at these points.

I also make the following assumption on the function S, which translates continuity of the distribution of Z to continuity of the distribution of $S(Z)$.

Assumption 4.2. *For any Lebesgue measure zero set A of strictly positive real numbers, $S^{-1}(A)$ has Lebesgue measure zero.*

Under these conditions, the asymptotic distribution in Theorem 3.1 is continuous. In addition to showing that the rate derived in that theorem is the exact rate of convergence

(since the distribution is not a point mass at zero or some other value), this shows that inference based on this asymptotic approximation will be asymptotically exact.

Theorem 4.1. *Under Assumptions 3.1, 3.2, and 4.1, the asymptotic distribution in Theorem 3.1 is continuous. If Assumptions 3.3 and 4.2 hold as well, the asymptotic distribution in Corollary 3.1 is continuous.*

Thus, an asymptotically exact test of $E(m(W_i, \theta)|X_i) \geq 0$ can be obtained by comparing the quantiles of $S(\inf_{s,t} E_n m(W_i, \theta) I(s < X_i < s + t))$ to the quantiles of any consistent estimate of the distribution of $S(Z)$. I propose two methods for estimating this distribution. The first is a generic subsampling procedure, and is described below. The second method uses the fact that the distribution of Z in Theorem 3.1 depends on the data generating process only through finite dimensional parameters to simulate an estimate of the asymptotic distribution, and is covered in Section B of the appendix.

For the subsampling based estimate, let $\tau_b = b^{(d_X+2)/(d_X+4)}$. For this choice of τ_b and some sequence $b = b_n$ with $b \to \infty$ and $b/n \to 0$, we use $L_{n,b}(\cdot|\tau_b)$ or $\tilde{L}_{n,b}(\cdot|\tau_b)$ to estimate the distribution of $n^{(d_X+2)/(d_X+4)} S(T_n(\theta))$, where $L_{n,b}$ and $\tilde{L}_{n,b}$ are given in Section 2.4. Thus, letting $L_{n,b}^{-1}$ and $\tilde{L}_{n,b}^{-1}$ be as defined in Section 2.4, we reject if $n^{(d_X+2)/(d_X+4))} S(T_n(\theta)) > L_{n,b}^{-1}(1 - \alpha | b^{(d_X+2)/(d_X+4)})$ (or if $n^{(d_X+2)/(d_X+4))} S(T_n(\theta)) > \tilde{L}_{n,b}^{-1}(1 - \alpha | b^{(d_X+2)/(d_X+4)})$). The following theorem states that this procedure is asymptotically exact. The result follows immediately from general results for subsampling in Politis, Romano, and Wolf (1999).

Theorem 4.2. *Under Assumptions 3.1, 3.2, 3.3, 4.1 and 4.2, the probability of rejecting using the subsampling procedure described above with nominal level α converges to α as long as $b \to \infty$ and $b/n \to 0$.*

To extend this method to conditions other than Assumption 3.1, one needs a pre-testing procedure to determine whether Assumption 3.1 or some other condition best describes the shape of the conditional mean. This is incorporated in the test described in Section 2.4, which is treated in detail in Section 6. Before describing these results, I extend the results of Section 3 to other shapes of the conditional mean. These results are needed for the tests in Section 6, which rely on the rate of convergence being sufficiently well behaved if it is in a certain range.

5 Other Shapes of the Conditional Mean

Assumption 3.1 states that the components of the conditional mean $\bar{m}(\theta, x)$ are minimized on a finite set and have strictly positive second derivative matrices at the minimum. More generally, if the conditional mean is less smooth, or does not take an interior minimum, $\bar{m}(\theta, x)$ could be minimized on a finite set, but behave differently near the minimum. Another possibility is that the minimizing set could have zero probability, while containing infinitely many elements (for example, an infinite countable set, or a lower dimensional set when $d_X > 1$).

In this section, I derive the asymptotic distribution and rate of convergence of KS statistics under a broader class of shapes of the conditional mean $\bar{m}(\theta, x)$. I replace part (ii) of Assumption 3.1 with the following assumption.

Assumption 5.1. *For $j \in J(k)$, $\bar{m}_j(\theta, x) = E(m_j(W_i, \theta)|X = x)$ is continuous on $B(x_k)$ and satisfies*

$$\sup_{\|x - x_k\| \leq \delta} \left\| \frac{\bar{m}_j(\theta, x) - \bar{m}_j(\theta, x_k)}{\|x - x_k\|^{\gamma(j,k)}} - \psi_{j,k}\left(\frac{x - x_k}{\|x - x_k\|}\right) \right\| \overset{\delta \to 0}{\to} 0$$

for some $\gamma(j,k) > 0$ and some function $\psi_{j,k} : \{t \in \mathbb{R}^{d_X} | \|t\| = 1\} \to \mathbb{R}$ with $\overline{\psi} \geq \psi_{j,k}(t) \geq \underline{\psi}$ for some $\overline{\psi} < \infty$ and $\underline{\psi} > 0$. For future reference, define $\gamma = \max_{j,k} \gamma(j,k)$ and $\tilde{J}(k) = \{j \in J(k) | \gamma(j,k) = \gamma\}$.

When Assumption 5.1 holds, the rate of convergence will be determined by γ, and the asymptotic distribution will depend on the local behavior of the objective function for j and k with $j \in \tilde{J}(k)$.

Under Assumption 3.1, Assumption 5.1 will hold with $\gamma = 2$ and $\psi_{j,k}(t) = \frac{1}{2} t V_j(x_k) t$ (this holds by a second order Taylor expansion, as described in the appendix). For $\gamma = 1$, Assumption 5.1 states that $\bar{m}_j(\theta, x)$ has a directional derivative for every direction, with the approximation error going to zero uniformly in the direction of the derivative. More generally, Assumption 5.1 states that $\bar{m}_j(\theta, x)$ increases like $\|x - x_k\|^\gamma$ near elements x_k in the minimizing set $\mathcal{X}_0$. For $d_X = 1$, this follows from simple conditions on the higher derivatives of the conditional mean with respect to x. With enough derivatives, the first derivative that is nonzero uniformly on the support of X_i determines γ. I state this formally in the next theorem. For higher dimensions, Assumption 5.1 requires additional conditions to rule out contact sets of dimension less than d_X, but greater than 1.

Theorem 5.1. *Suppose $\bar{m}(\theta, x)$ has p bounded derivatives, $d_X = 1$ and $supp(X_i) = [\underline{x}, \overline{x}]$. Then, if $\min_j \inf_x \bar{m}_j(\theta, x) = 0$, either Assumption 5.1 holds, with the contact set $\mathcal{X}_0$ possibly containing the boundary points $\underline{x}$ and $\overline{x}$, for $\gamma = r$ for some integer $r < p$, or, for some x_0 on the support of X_i and some finite B, $\bar{m}_j(\theta, x) \leq B|x - x_0|^p$ for some j.*

Theorem 5.1 states that, with $d_X = 1$ and p bounded derivatives, either Assumption 5.1 holds for γ some integer less than p, or, for some j, $\bar{m}_j(\theta, x)$ is less than or equal to the function $B|x-x_0|^p$. In the latter case, adding the nonnegative variable $B|X_i-x_0|^p-\bar{m}(\theta, X_i)$ to $m_j(W_i, \theta)$ would make Assumption 5.1 hold for $\gamma = p$, so the rate of convergence for the KS statistic must be at least as slow as the rate of convergence when Assumption 3.1 holds with $\gamma = p$. This classification of the possible rates of convergence is used in the subsampling based estimates of the rate of convergence described in Sections 2.4 and 6.

Under Assumption 3.1 with part (ii) replaced by Assumption 5.1, the following modified version of Theorem 3.1, with a different rate of convergence and limiting distribution, will hold.

Theorem 5.2. *Under Assumption 3.1, with part (ii) replaced by Assumption 5.1, and Assumption 3.2,*

$$n^{(d_X+\gamma)/(d_X+2\gamma)} \inf_{s,t} E_n m(W_i, \theta) I(s < X_i < s + t) \xrightarrow{d} Z$$

where Z is the random vector on $\mathbb{R}^{d_Y}$ defined as in Theorem 3.1, but with $J(k)$ replaced by $\tilde{J}(k)$ and $g_{P,x_k,j}(s, t)$ defined as

$$g_{P,x_k,j}(s,t) = f_X(x_k) \int_{s_1}^{s_1+t_1} \cdots \int_{s_{d_X}}^{s_{d_X}+t_{d_X}} \psi_{j,k}\left(\frac{x}{\|x\|}\right) \|x\|^\gamma \, dx_{d_X} \cdots dx_1$$

for $j \in \tilde{J}(k)$. If Assumption 3.3 holds as well, then

$$n^{(d_X+\gamma)/(d_X+2\gamma)} S(\inf_{s,t} E_n m(W_i, \theta) I(s < X_i < s + t)) \xrightarrow{d} S(Z).$$

If Assumption 4.1 holds as well, Z has a continuous distribution. If Assumptions 3.3, 4.1 and 4.2 hold, $S(Z)$ has a continuous distribution.

Theorem 5.2 can be used once Assumption 5.1 is known to hold for some γ (which, in the case where $d_X = 1$, holds under the conditions of Theorem 5.1), as long as γ can be estimated. The procedure described in Section 2.4 uses an estimated rate of convergence

based on subsampling, and a detailed derivation of this procedure is given in the next section. Section B.2 of the appendix provides an alternative procedure based on estimating the second derivative matrix of the conditional mean.

6 Testing Rate of Convergence Conditions

This section gives a derivation of the procedure described in Section 2.4, and gives a formal result with conditions under which the procedure is asymptotically exact or conservative. See Section B.2 of the appendix for an alternative approach based on estimation of the second derivative.

The procedure uses pre-tests for the rate of convergence, which mostly follow Bertail, Politis, and Romano (1999) (see also Chapter 8 of Politis, Romano, and Wolf, 1999), but with some modifications to accomodate the possibility that the statistic may not converge at a polynomial rate if the rate is slow enough. The results in Section 5 are used to give primitive conditions under which the rate of convergence will be well behaved so that these results can be applied.

Let $L_{n,b}(x|\tau)$, $L_{n,b}(x|1)$, $L^{-1}_{n,b}(x|\tau)$ and $L^{-1}_{n,b}(x|1)$ be defined as in Section 2.4. Note that $\tau_b L^{-1}_{n,b}(t|1) = L^{-1}_{n,b}(t|\tau)$. If τ_n is the true rate of convergence, $L^{-1}_{n,b_1}(t|\tau)$ and $L^{-1}_{n,b_2}(t|\tau)$ both approximate the tth quantile of the asymptotic distribution. Thus, if $\tau_n = n^\beta$ for some β, $b_1^\beta L^{-1}_{n,b_1}(t|1)$ and $b_2^\beta L^{-1}_{n,b_2}(t|1)$ should be approximately equal, so that an estimator for β can be formed by choosing $\hat{\beta}_t$ to set these quantities equal. Some calculation gives

$$\hat{\beta}_t = (\log L^{-1}_{n,b_2}(t|1)) - \log L^{-1}_{n,b_1}(t|1))/(\log b_1 - \log b_2).$$

The rate estimate $\hat{\beta}$ defined in (1) averages these over a finite number of quantiles t, and is one of the estimators proposed by Bertail, Politis, and Romano (1999).

The results in Bertail, Politis, and Romano (1999) show that subsampling with the estimated rate of convergence $n^{\hat{\beta}}$ is valid as long as the true rate of convergence is n^β for some $\beta > 0$. However, this will not always be the case for the estimators considered in this paper. For example, under the conditions of Theorem 5.1, the rate of convergence will either be $n^{(1+\gamma)/(1+2\gamma)}$ for some $\gamma < p$ (here, $d_X = 1$), or the rate of convergence will be at least as slow as $n^{(1+p)/(1+2p)}$. In the latter case, Theorem 5.1 does not guarantee that the rate of convergence is of the form n^β. Even if Assumption 5.1 holds for some γ for θ on the boundary of the identified set, the rate of convergence will be faster for θ on the interior of the identified set, where trying not to be conservative typically has little payoff in terms of

power against parameters outside of the identified set.

The procedure in Section 2.4 uses truncation to remedy these issues. The estimated rate of convergence is truncated above at $\overline{\beta} < 1$, so that the test will be conservative on the interior of the identified set. If the rate of convergence is estimated to be slower than $\underline{\beta}$, the test reverts to a conservative $\sqrt{n}$ rate, which handles the case where the statistic may oscillate between slower rates. In cases where the true exponent is between $\underline{\beta}$ and $\overline{\beta}$, the procedure is asymptotically exact. Note that the theorem below allows the contact set $\mathcal{X}_0$ to be a positive probability set, a countable set with zero probability, or some other set with infinitely many elements. As long as condition (ii) in the theorem holds, the contact set need not be finite.

Theorem 6.1. *Suppose that Assumptions 3.2, 3.3 and 4.2 hold, and that S is convex and that $E(m(W_i,\theta)m(W_i,\theta)'|X_i = x)$ is continuous and strictly positive definite. Suppose that, for some $\overline{\gamma}$, either of the following conditions holds:*

i.) Assumptions 3.1 and 4.1 hold with part (ii) of Assumption 3.1 replaced by Assumption 5.1 for some $\gamma \leq \overline{\gamma}$, where the set $\mathcal{X}_0 = \{x|\bar{m}_j(\theta,x) = 0 \text{ some } j\}$ may be empty

or

ii.) for some $x_0 \in \mathcal{X}_0$ such that X_i has a continuous density in a neighborhood of x_0 and $B < \infty$, $\bar{m}_j(\theta,x) \leq B\|x - x_0\|^\gamma$ for some $\gamma > \overline{\gamma}$ and some j.

Under these conditions, the test in Section 2.4 is asymptotically level α. If Assumption 3.1 holds with part (ii) of Assumption 3.1 replaced by Assumption 5.1 for some $\underline{\gamma} < \gamma < \overline{\gamma}$ and $\mathcal{X}_0$ nonempty, this test will be asymptotically exact level α.

In the one dimensional case, the conditions of Theorem 6.1 follow immediately from smoothness assumptions on the conditional mean by Theorem 5.1. The following theorem states this formally (the proof is immediate from Theorem 5.1). The condition that the minimum not be taken on the boundary of the support of X_i could be removed by extending Theorem 5.2 to allow $\mathcal{X}_0$ to include boundary points, or the result can be used as stated with a pre-test for this condition.

Theorem 6.2. *Suppose that $d_X = 1$, Assumptions 3.2, 3.3 and 4.2 hold, and that S is convex and $E(m(W_i,\theta)m(W_i,\theta)'|X_i = x)$ is continuous and strictly positive definite. Suppose that $supp(X_i) = [\underline{x},\overline{x}]$ and that $\bar{m}(\theta,x)$ is bounded away from zero near $\underline{x}$ and $\overline{x}$ and has p bounded derivatives. Then the conditions of Theorem 6.1 hold for any $\overline{\gamma} < p$.*

The recommendation $\overline{\beta} = (d_X + 1)/(d_X + 2)$ given in Section 2.4 corresponds to $\overline{\gamma} = 1$ (a single directional derivative). The recommendation $\underline{\beta} = [(d_X + 2)/(d_X + 4) + 1/2]/2$ corresponds to $\underline{\beta}$ halfway between the rate for two derivatives and the exponent 1/2 for the conservative rate (however, the number of derivatives p needed to justify this choice in Theorem 6.2 is greater than 2).

It should be noted that Theorems 6.1 and 6.2 require stronger conditions, and show a weaker notion of coverage, compared to the results of Andrews and Shi (2013) for the more conservative approach considered in that paper. Theorems 6.1 and 6.2 place smoothness conditions on the conditional mean, while the approach of Andrews and Shi (2013) does not require such conditions. Since the shape of the conditional mean plays an integral role in the asymptotic distributions derived in this paper, it seems likely that some smoothness conditions along the lines of those used in these theorems are indeed needed for the conclusions regarding the validity of these tests to hold. Thus, the power improvements for this procedure, which are shown in the next section, likely come at a cost of additional assumptions.

Regarding the notion of coverage shown by these theorems, these theorems show that the probability of false rejection is asymptotically less than or equal to the the nominal level for certain data generating processes P and parameter values θ in the identified set under P. However, this leaves open the possibility that there may be sequences (θ_n, P_n) under which the rejection probability is not controlled, even though (θ_n, P_n) satisfies the conditions of the above theorems for each n. For example, one might worry that, even though the above procedure works well when $E[m(W_i, \theta)|X_i = x] = \|X_i - x_0\|^\gamma$ for any given γ, there are sequences γ_n under which the test overrejects. This issue of uniformity in the underlying distribution is often a concern in situations such as the present one, where the asymptotic distribution changes dramatically with the data generating process (see, e.g., Andrews and Guggenberger, 2010; Romano and Shaikh, 2012).

Although more conservative, the procedures of Andrews and Shi (2013) are known to be valid uniformly over relatively broad classes of data generating processes. While the uniform validity of the procedures in the present paper is left for future research, stronger conditions are needed even for asymptotic control of the rejection probability for a given (θ, P). Thus, one should excercise caution in interpreting confidence regions based on this procedure. On the other hand, the tests in Andrews and Shi (2013) use a critical value that is, asymptotically, determined entirely by a certain tuning parameter (the infinitesimal uniformity factor, in the terminology of that paper) under the data generating processes

considered here. Since the results in the present paper give a nondegenerate approximation to how the test statistic behaves under the null in such situations, one may be more confident in excluding a parameter value from a confidence region if one of the tests in the present paper rejects as well.

7 Local Alternatives

Consider local alternatives of the form $\theta_n = \theta_0 + a_n$ for some fixed θ_0 such that $m(W_i, \theta_0)$ satisfies Assumption 3.1 and $a_n \to 0$. Throughout this section, I restrict attention to the conditions in Section 3, which corresponds to the more general setup in Section 5 with $\gamma = 2$. To translate the a_n rate of convergence to θ_0 to a rate of convergence for the sequence of conditional means, I make the following assumptions. As before, define $\bar{m}(\theta, x) = E(m(W_i, \theta)|X_i = x)$.

Assumption 7.1. *For each $x_k \in \mathcal{X}_0$, $\bar{m}(\theta, x)$ has a derivative as a function of θ in a neighborhood of (θ_0, x_k), denoted $\bar{m}_\theta(\theta, x)$, that is continuous as a function of (θ, x) at (θ_0, x_k) and, for any neighborhood of x_k, there is a neighborhood of θ_0 such that $\bar{m}_j(\theta, x)$ is bounded away from zero for θ in the given neighborhood of θ_0 and x outside of the given neighborhood of x_k for $j \in J(k)$ and for all x for $j \notin J(k)$.*

Assumption 7.2. *For each $x_k \in \mathcal{X}_0$ and $j \in J(k)$, $E\{[m_j(W_i, \theta) - m_j(W_i, \theta_0)]^2|X_i = x\}$ converges to zero uniformly in x in some neighborhood of x_k as $\theta \to \theta_0$.*

I also make the following assumption, which extends Assumption 3.2 to a neighborhood of θ_0.

Assumption 7.3. *For some fixed $\overline{Y} < \infty$ and θ in a some neighborhood of θ_0, $|m(W_i, \theta)| \leq \overline{Y}$ with probability one.*

In the interval regression example, these conditions are satisfied as long as Assumption 3.1 holds at θ_0 and the data have finite support. These conditions are also likely to hold in a variety of models once Assumption 3.1 holds at θ_0.

The following theorem derives the behavior of the test statistic under local alternatives relative to critical values based on the results in this paper.

Theorem 7.1. *Let θ_0 be such that $E(m(W_i, \theta_0)|X_i) \geq 0$ almost surely and Assumptions 3.1, 7.1, 7.2, and 7.3 are satisfied for θ_0. Let $a \in \mathbb{R}^{d_\theta}$ and let $a_n = an^{-2/(d_X+4)}$. Let $Z(a)$*

be a random variable defined the same way as Z in Theorem 3.1, but with the functions $g_{P,x_k,j}(s,t)$ replaced by the functions

$$g_{P,x_k,j,a}(s,t) = \frac{1}{2} f_X(x_k) \int_{s<x<s+t} x' V_j(x_k) x \, dx + \bar{m}_{\theta,j}(\theta_0, x_k) a f_X(x_k) \prod_i t_i$$

for $j \in J(k)$ for each k where $\bar{m}_{\theta,j}$ is the jth row of the derivative matrix $\bar{m}_\theta$. Then

$$n^{(d_X+2)/(d_X+4)} \inf_{s,t} E_n m(W_i, \theta + a_n) I(s < X_i < s+t) \xrightarrow{d} Z(a).$$

An immediate consequence of this theorem is that an asymptotically exact test gives power against $n^{-2/(d_X+4)}$ alternatives (as long as $\bar{m}_{\theta,j}(\theta_0, x_k)a$ is negative for each j or negative enough for at least one j), but not against alternatives that converge strictly faster (while this follows immediately from Theorem 7.1 only if critical values are based directly on the asymptotic distribution under θ_0, it can be shown using standard arguments from the subsampling literature that this holds for the subsampling based critical values as well). The dependence on the dimension of X_i is a result of the curse of dimensionality. With a fixed amount of "smoothness," the speed at which local alternatives can converge to the null space and still be detected is decreasing in the dimension of X_i.

Note that the minimax optimal rate for nonparametric testing in the supremum norm with two derivatives is $(n/\log n)^{-2/(d_X+4)}$ (see, e.g., Lepski and Tsybakov, 2000), so the $n^{-2/(d_X+4)}$ rate derived here is faster than this rate by a $\log n$ factor. This does not contradict the minimax rates since (1) the tests in this paper have not been shown to control size uniformly over underlying distributions in this smoothness class, and require more smoothness even for pointwise validity and (2) the local alternatives considered here differ from those used to derive minimax rates (here, the conditional moment restriction is violated near the contact set $\mathcal{X}_0$, and the fact that the conditional mean is bounded away from zero away from this set makes it easier to "find" this set; this is not the case when one considers minimax rates).

Now consider power against local alternatives of this form, with a possibly different sequence a_n, using the conservative estimate that $\sqrt{n} \inf_{s,t} E_n m(W_i, \theta) I(s < X_i < s+t) \xrightarrow{p} 0$ for $\theta \in \Theta_0$. That is, we fix some $\eta > 0$ and reject if $\sqrt{n} S(\inf_{s,t} E_n m(W_i, \theta_0 + a_n) I(s < X_i < s+t)) > \eta$. The following theorem shows that this test will reject only when a_n approaches the boundary of the identified set at a slower rate.

Theorem 7.2. *Let θ_0 be such that $E(m(W_i, \theta_0)|X_i) \geq 0$ almost surely and Assumptions 3.1,*

7.1, 7.2, and 7.3 are satisfied for θ_0. Let $a \in \mathbb{R}^{d_\theta}$ and let $a_n = an^{-1/(d_X+2)}$. Then, for each j,

$$\begin{aligned}&\sqrt{n}\inf_{s,t} E_n m_j(W_i, \theta_0 + a_n) I(s < X < s+t) \\ &\xrightarrow{p} \min_{k \text{ s.t. } j \in J(k)} \inf_{s,t} f_X(x_k) \int_{s<x<s+t} \left(\frac{1}{2} x' V x + \overline{m}_{\theta,j}(\theta_0, x_k) a \right) dx.\end{aligned}$$

The $n^{-1/(d_X+2)}$ rate is slower than the $n^{-2/(d_X+4)}$ rate for detecting local alternatives with the asymptotically exact test. As with the asymptotically exact tests, the conservative tests do worse against this form of local alternative as the dimension of the conditioning variable X_i increases.

8 Monte Carlo

I perform a monte carlo study to examine the finite sample behavior of the tests I propose, and to see how well the asymptotic results in this paper describe the finite sample behavior of KS statistics. First, I simulate the distribution of KS statistics for various sample sizes under parameter values and data generating processes that satisfy Assumption 3.1, and for data generating processes that lead to a $\sqrt{n}$ rate of convergence. As predicted by Theorem 3.1, for the data generating process that satisfies Assumption 3.1, the distribution of the KS statistic is roughly stable across sample sizes when scaled up by $n^{(d_X+2)/(d_X+4)}$. For the data generating process that leads to $\sqrt{n}$ convergence, scaling by $\sqrt{n}$ gives a distribution that is stable across sample sizes. Next, I examine the size and power of KS statistic based tests using the asymptotic distributions derived in this paper. I include procedures that test between the conditions leading to $\sqrt{n}$ convergence and the faster rates derived in this paper using the subsampling estimates of the rate of convergence described in Sections 2.4 and 6, as well as infeasible procedures that use prior knowledge of the correct rate of convergence to estimate the asymptotic distribution.

8.1 Monte Carlo Designs

Throughout this section, I consider two monte carlo designs for a mean regression model with missing data. In this model, the latent variable W_i^* satisfies $E(W_i^*|X_i) = \theta_1 + \theta_2 X_i$, but W_i^* is unobserved, and can only be bounded by the observed variables $W_i^H = \overline{w} I(W_i^* \text{ missing}) + W_i^* I(W_i^* \text{ observed})$ and $W_i^L = \underline{w} I(W_i^* \text{ missing}) + W_i^* I(W_i^* \text{ observed})$ are observed, where

$[\underline{w}, \overline{w}]$ is an interval known to contain W_i^*. The identified set Θ_0 is the set of values of (θ_1, θ_2) such that the moment inequalities $E(W_i^H - \theta_1 - \theta_2 X_i | X_i) \geq 0$ and $E(\theta_1 + \theta_2 X_i - W_i^L | X_i) \geq 0$ hold with probability one. For both designs, I draw X_i from a uniform distribution on $(-1, 1)$ (here, $d_X = 1$). Conditional on X_i, I draw U_i from an independent uniform $(-1, 1)$ distribution, and set $W_i^* = \theta_{1,*} + \theta_{2,*} X_i + U_i$, where $\theta_{1,*} = 0$ and $\theta_{2,*} = .1$. I then set W_i^* to be missing with probability $p^*(X_i)$ for some function p^* that differs across designs. I set $[\underline{w}, \overline{w}] = [-.1 - 1, .1 + 1] = [-1.1, 1.1]$, the unconditional support of W_i^*. Note that, while the data are generated using a particular value of θ in the identified set and a censoring process that satisfies the missing at random assumption (that the probability of data missing conditional on (X_i, W_i^*) does not depend on W_i^*), the data generating process is consistent with forms of endogenous censoring that do not satisfy this assumption. The identified set contains all values of θ for which the data generating process is consistent with the latent variable model for θ and some, possibly endogenous, censoring mechanism.

The shape of the conditional moment inequalities as a function of X_i depends on p^*. For Design 1, I set $p^*(x) = (0.9481x^4 + 1.0667x^3 - 0.6222x^2 - 0.6519x + 0.3889) \wedge 1$. The coefficients of this quartic polynomial were chosen to make $p^*(x)$ smooth, but somewhat wiggly, so that the quadratic approximation to the resulting conditional moments used in Theorem 3.1 will not be good over the entire support of X_i. The resulting conditional means of the bounds on W_i^* are $E(W_i^L | X_i = x) = (1 - p^*(x))(\theta_{1,*} + \theta_{2,*} x) + p^*(x)\underline{w}$ and $E(W_i^H | X_i = x) = (1 - p^*(x))(\theta_{1,*} + \theta_{2,*} x) + p^*(x)\overline{w}$. In the monte carlo study, I examine the distribution of the KS statistic for the upper inequality at $(\theta_{1,D1}, \theta_{2,D1}) \equiv (1.05, .1)$, a parameter value on the boundary of the identified set for which Assumption 3.1 holds, along with confidence intervals for the intercept parameter θ_1 with the slope parameter θ_2 fixed at .1. For the confidence regions, I also restrict attention to the moment inequality corresponding to W_i^H, so that the confidence regions are for the one sided model with only this conditional moment inequality (this also makes the choice of the function S largely irrelevant; throughout the monte carlos, I take $S(t) = |t \wedge 0|$). Figure 3 plots the conditional means of W_i^H and W_i^L, along with the regression line corresponding to $\theta = (1.05, .1)$. The confidence intervals for the slope parameter invert a family of tests corresponding to values of θ that move this regression line vertically.

For Design 2, I set $p^*(x) = [(|x - .5| \vee .25) - .15] \wedge .7)$. Figure 4 plots the resulting conditional means. For this design, I examine the distribution of the KS statistic for the upper inequality at $(\theta_{1,D2}, \theta_{2,D2}) = (1.1, .9)$, which leads to a positive probability contact set for the upper moment inequality and a $n^{1/2}$ rate of convergence to a nondegenerate

distribution. The regression line corresponding to this parameter is plotted in Figure 4 as well. For this design, I form confidence intervals for the slope parameter θ_1 with θ_2 fixed at .9, using the KS statistic for the moment inequality for W_i^H.

The confidence intervals reported in this section are computed by inverting the tests on a grid of parameter values. I use a grid with meshwidth .01 that covers the area of the parameter space with distance to the boundary of the identified set no more than 1.

8.2 Distribution of the KS Statistic

To examine how well Theorem 3.1 describes the finite sample distribution of KS statistics under Assumption 3.1, I simulate from Design 1 for a range of sample sizes and form the KS statistic for testing $(\theta_{1,D1}, \theta_{2,D1})$. Since Assumption 3.1 holds for testing this value of θ under this data generating process, Theorem 3.1 predicts that the distribution of the KS statistic scaled up by $n^{(d_X+2)/(d_X+4)} = n^{3/5}$ should be similar across the sample sizes. The performance of this asymptotic prediction in finite samples is examined in Figure 5, which plots histograms of the scaled KS statistic $n^{3/5}S(T_n(\theta))$ for the sample sizes $n \in \{100, 500, 1000, 2000, 5000\}$. The scaled distributions appear roughly stable across sample sizes, as predicted.

In contrast, under Design 2, the KS statistic for testing $(\theta_{1,D2}, \theta_{2,D2})$ will converge at a $n^{1/2}$ rate to a nondegenerate distribution. Thus, asymptotic approximation suggests that, in this case, scaling by $n^{1/2}$ will give a distribution that is roughly stable across sample sizes. Figure 6 plots histograms of the scaled statistic $n^{1/2}S(T_n(\theta))$ for this case. The scaling suggested by asymptotic approximations appears to give a distribution that is stable across sample sizes here as well.

8.3 Finite Sample Performance of the Tests

I now turn to the finite sample performance of confidence regions for the identified set based on critical values formed using the asymptotic approximations derived in this paper, along with possibly conservative confidence regions that use the $n^{1/2}$ approximation. The critical values use subsampling with different assumed rates of convergence. I report results for the tests based on subsampling estimates of the rate of convergence described in Sections 2.4 and 6, tests that use the conservative rate $n^{1/2}$, and infeasible tests that use a $n^{3/5}$ rate under Design 1, and a $n^{1/2}$ rate under Design 2. The implementation details are as follows. For the critical values using the conservative rate of convergence, I estimate the .9 and .95 quantiles of the distribution of the KS statistic at each value of θ using subsampling, and

add the correction factor .001 to prevent the critical value from going to zero. The critical values using estimated rates of convergence are computed as described in Section 2.4, with the recommended tuning parameters given in that section. All subsampling estimates use 1000 subsample draws.

Table 1 reports the coverage probabilities for $(\theta_{1,D1}, \theta_{2,D1})$ under Design 1. As discussed above, under Design 1, $(\theta_{1,D1}, \theta_{2,D1})$ is on the boundary of the identified set and satisfies Assumption 3.1. As predicted, the tests that subsample with the $n^{1/2}$ rate are conservative. The nominal 95% confidence regions that use the $n^{1/2}$ rate cover $(\theta_{1,D1}, \theta_{2,D1})$ with probability at least .99 for all of the sample sizes. Subsampling with the exact $n^{3/5}$ rate of convergence, an infeasible procedure that uses prior knowledge that Assumption 3.1 holds under $(\theta_{1,D1}, \theta_{2,D1})$ for this data generating process, gives confidence regions that cover $(\theta_{1,D1}, \theta_{2,D1})$ with probability much closer to the nominal coverage. The subsampling tests with the estimated rate of convergence also perform well, attaining close to the nominal coverage.

Table 2 reports coverage probabilities for testing $(\theta_{1,D2}, \theta_{2,D2})$ under Design 2. In this case, subsampling with a $n^{1/2}$ rate gives an asymptotically exact test of $(\theta_{1,D2}, \theta_{2,D2})$, so we should expect the coverage probabilities for the tests that use the $n^{1/2}$ rate of convergence to be close to the nominal coverage probabilities, rather than being conservative. The coverage probabilities for the $n^{1/2}$ rate are generally less conservative here than for Design 1, as the asymptotic approximations predict, although the coverage is considerably greater than the nominal coverage, even with 5000 observations. In this case, the infeasible procedure is identical to the conservative test, since the exact rate of convergence is $n^{1/2}$. The confidence regions that use subsampling with the estimated rate contain $(\theta_{1,D2}, \theta_{2,D2})$ with probability close to the nominal coverage (although undercoverage is somewhat severe in the $n = 100$ case), but are generally more liberal than their nominal level.

Tables 3 and 4 summarize the portion of the parameter space outside of the identified set covered by confidence intervals for the intercept parameter θ_1 with θ_2 fixed at $\theta_{2,D1}$ for Design 1 and $\theta_{2,D2}$ for Design 2. The entries in each table report the upper endpoint of one of the confidence regions minus the upper endpoint of the identified set for the slope parameter, averaged over the monte carlo draws. As discussed above, the true upper endpoint of the identified set for θ_1 under Design 1 with θ_2 fixed at $\theta_{2,D1}$ is $\theta_{1,D1}$, and the true upper endpoint of the identified set for θ_1 under Design 2 with θ_2 fixed at $\theta_{2,D2}$ is $\theta_{1,D2}$, so, letting $\hat{u}_{1-\alpha}$ be the greatest value of θ_1 such that $(\theta_1, \theta_{2,D1})$ is not rejected, Table 3 reports averages of $\hat{u}_{1-\alpha} - \theta_{2,D1}$, and similarly for Table 4 and Design 2.

The results of Section 7 suggest that, for the results for Design 1 reported in Table 3,

the difference between the upper endpoint of the confidence region and the upper endpoint of the identified set should decrease at a $n^{2/5}$ rate for the critical values that use or estimate the exact rate of convergence (the first and third rows), and a $n^{1/3}$ rate for subsampling with the conservative rate and adding .001 to the critical value (the second row). This appears roughly consistent with the values reported in these tables. The conservative confidence regions start out slightly larger, and then converge more slowly. For Design 2, the KS statistic converges at a $n^{1/2}$ rate on the boundary of the identified set for θ_1 for θ_2 fixed at $\theta_{2,D2}$, and arguments in Andrews and Shi (2013) show that $n^{1/2}$ approximation to the KS statistic give power against sequences of alternatives that approach the identified set at a $n^{1/2}$ rate. The confidence regions do appear to shrink to the identified set at approximately this rate over most sample sizes, although the decrease in the width of the confidence region is larger than predicted for smaller sample sizes, perhaps reflecting additional power improvements as the subsampling procedures find the binding moments.

9 Illustrative Empirical Application

As an illustrative empirical application, I apply the methods in this paper to regressions of out of pocket prescription drug spending on income using data from the Health and Retirement Study (HRS). In this survey, respondents who did not report point values for these and other variables were asked whether the variables were within a series of brackets, giving point values for some observations and intervals of different sizes for others. The income variable used here is taken from the RAND contribution to the HRS, which adds up reported income from different sources elicited in the original survey. For illustrative purposes, I focus on the subset of respondents who report point values for income, so that only prescription drug spending, the dependent variable, is interval valued. The resulting confidence regions are valid under any potentially endogenous process governing the size of the reported interval for prescription expenditures, but require that income be missing or interval reported at random. I use the 1996 wave of the survey and restrict attention to women with no more than \$15,000 of yearly income who report using prescription medications. This results in a data set with 636 observations. Of these observations, 54 have prescription expenditures reported as an interval of nonzero width with finite endpoints, and an additional 7 have no information on prescription expenditures.

To describe the setup formally, let X_i and W_i^* be income and prescription drug expenditures for the ith observation. We observe (X_i, W_i^L, W_i^H), where $[W_i^L, W_i^H]$ is an interval that

contains W_i^*. For observations where no interval is reported for prescription drug spending, I set $W_i^L = 0$ and $W_i^H = \infty$. I estimate an interval median regression model where the median $q_{1/2}(W_i^*|X_i)$ of W_i^* given X_i is assumed to follow a linear regression model $q_{1/2}(W_i^*|X_i) = \theta_1 + \theta_2 X_i$. This leads to the conditional moment inequalities $E(m(W_i, \theta)|X_i) \geq 0$ almost surely, where $m(W_i, \theta) = (I(\theta_1 + \theta_2 X_i \leq W_i^H) - 1/2, 1/2 - I(\theta_1 + \theta_2 X_i \leq W_i^L))$ and $W_i = (X_i, W_i^L, W_i^H)$.

Figure 7 shows the data graphically. The horizontal axis measures income, while the vertical axis measures out of pocket prescription drug expenditures. Observations for which prescription expenditures are reported as a point value are plotted as points. For observations where a nontrivial interval is reported, a plus symbol marks the upper endpoint, and an x marks the lower endpoint. For observations where no information on prescription expenditures is obtained in the survey, a circle is placed on the x axis at the value of income reported for that observation. The vertical axis is truncated at \$15,000, leading to 5 observations not being shown (these observations are still used in forming the confidence regions reported below).

I form 95% confidence intervals by inverting level .05 tests using the KS statistics described in this paper with critical values calculated using the conservative rate of convergence $n^{1/2}$, and rates of convergence estimated using the methods described in Sections 2.4 and 6. The rest of the implementation details are the same as for the monte carlos in Section 8.

For comparison, I also compute point estimates and confidence regions using the least absolute deviations (LAD) estimator (Koenker and Bassett, 1978) for the median regression model with only the observations for which a point value for spending was reported. These are valid under the additional assumption that the decision to report an interval or missing value is independent of spending conditional on income. The confidence regions use Wald tests based on the asymptotic variance estimates computed by Stata. These asymptotic variance estimates are based on formulas in Koenker and Bassett (1982) and require additional assumptions on the data generating process, but I use these rather than more robust standard errors in order to provide a comparison to an alternative procedure using default options in a standard statistical package.

Figure 8 plots the outline of the 95% confidence region for θ using the pre-tests and rate of convergence estimates described above, while Figure 9 plots the outline of the 95% confidence region using the conservative approximation. Figure 12 plots the outline of the 95% confidence region from estimating a median regression model on the subset of the data with point values reported for spending. Table 5 reports the corresponding confidence

intervals for the components of θ. For the confidence regions based on KS tests, I use the projections of the confidence region for θ onto each component. For the confidence regions based on median regression with point observations, the 95% confidence regions use the limiting normal approximation for each component of θ separately.

The results show a sizeable increase in statistical power from using the estimated rates of convergence. With the conservative tests, the 95% confidence region estimates that a \$1,000 increase in income is associated with at least a \$3 increase in out of pocket prescription spending at the median. With the tests that use the estimated rates of convergence, the 95% confidence region bounds the increase in out of pocket prescription spending associated with a \$1,000 increase in income from below by \$11.30.

The 95% confidence region based on median regression using observations reported as points overlaps with both moment inequality based confidence regions, but gives a different picture of which parameter values can be ruled out by the data. The upper bound for the increase in spending associated with a \$1,000 increase in income is \$24.40 using LAD, compared to \$37.20 and \$34.70 using KS statistics with all observations and the conservative and estimated rates respectively. The corresponding lower bound is \$10 using LAD with point observations, substantially larger than the lower bound of \$3 using the conservative procedure, but actually smaller than the \$11.30 lower bound under the estimated rate.

Note also that these tests could, but do not, provide evidence against the assumptions required for LAD on the point reported values. If the LAD 95% confidence region did not overlap with one of the moment inequality 95% confidence regions, there would be no parameter value consistent with this assumption at the .1 level (for any parameter value, we can reject the joint null of both models holding using Bonferroni's inequality and the results of the .05 level tests). This type of test will not necessarily have power if the interval reporting at random assumption for the dependent variable does not hold, so it should not be taken as evidence that the more robust interval regression assumptions can be replaced with LAD methods.

It is worth noting that, while the identified set for this model is convex, neither of the confidence regions in Figures 8 or 9 are convex. The shape of the confidence region in Figures 8, which uses the new procedures proposed in this paper, is particularly irregular. The nonconvexity of these regions likely arises from the fact that the pre-tests and subsampling procedures use different critical values for different values of θ, so that moving θ in a particular direction may first cause the test to reject (as the test statistic gets larger), then fail to reject (as the critical value gets larger) and then reject again (as the test statistic increases enough

to overcome the larger critical value). If convexity of the confidence region is desired, one can report the convex hull of the original confidence region. This is done for the confidence regions in this section in Figures 10 and 11. The resulting confidence region, by construction, has at least the same coverage probability as the original confidence region (although it may be more conservative).

10 Conclusion

This paper derives the asymptotic distribution of a class of Kolmogorov-Smirnov style test statistics for conditional moment inequality models under a general set of conditions. I show how to use these results to form valid tests that are more powerful than existing approaches based on this statistic. Local power results for the new tests and existing tests are derived, which quantify this power improvement. While the increase in power comes at a cost of robustness to smoothness conditions, a complementary paper (Armstrong, 2011, 2014) proposes methods for inference that achieve almost the same power improvement while still being robust to failure of smoothness conditions.

References

Adler, R. J. (1990): "An Introduction to Continuity, Extrema, and Related Topics for General Gaussian Processes," *Lecture Notes-Monograph Series*, 12, i–155.

Andrews, D. W., S. Berry, and P. Jia (2004): "Confidence regions for parameters in discrete games with multiple equilibria, with an application to discount chain store location," .

Andrews, D. W., and P. Guggenberger (2009): "Validity of Subsampling and ?plug-in Asymptotic? Inference for Parameters Defined by Moment Inequalities," *Econometric Theory*, 25(03), 669–709.

——— (2010): "Asymptotic Size and a Problem with Subsampling and with the m out of n Bootstrap," *Econometric Theory*, 26(02), 426–468.

Andrews, D. W. K., and P. Jia (2008): "Inference for Parameters Defined by Moment Inequalities: A Recommended Moment Selection Procedure," *SSRN eLibrary*.

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

Andrews, D. W. K., and G. Soares (2010): "Inference for Parameters Defined by Moment Inequalities Using Generalized Moment Selection," *Econometrica*, 78(1), 119–157.

Armstrong, T. (2011): "Weighted KS Statistics for Inference on Conditional Moment Inequalities," *Unpublished Manuscript*.

Armstrong, T., and H. P. Chan (2012): "Multiscale Adaptive Inference on Conditional Moment Inequalities," *Unpublished Manuscript*.

Armstrong, T. B. (2014): "Weighted KS statistics for inference on conditional moment inequalities," *Journal of Econometrics*, 181(2), 92–116.

Beresteanu, A., and F. Molinari (2008): "Asymptotic Properties for a Class of Partially Identified Models," *Econometrica*, 76(4), 763–814.

Bertail, P., D. N. Politis, and J. P. Romano (1999): "On Subsampling Estimators with Unknown Rate of Convergence," *Journal of the American Statistical Association*, 94(446), 569–579.

Bierens, H. J. (1982): "Consistent model specification tests," *Journal of Econometrics*, 20(1), 105–134.

Bontemps, C., T. Magnac, and E. Maurin (2012): "Set Identified Linear Models," *Econometrica*, 80(3), 1129–1155.

Bugni, F. A. (2010): "Bootstrap Inference in Partially Identified Models Defined by Moment Inequalities: Coverage of the Identified Set," *Econometrica*, 78(2), 735–753.

Chernozhukov, V., H. Hong, and E. Tamer (2007): "Estimation and Confidence Regions for Parameter Sets in Econometric Models," *Econometrica*, 75(5), 1243–1284.

Chernozhukov, V., S. Lee, and A. M. Rosen (2009): "Intersection bounds: estimation and inference," *Arxiv preprint arXiv:0907.3503*.

Chetty, R. (2010): "Bounds on Elasticities with Optimization Frictions: A Synthesis of Micro and Macro Evidence on Labor Supply," *NBER Working Paper*.

Chetverikov, D. (2012): "Adaptive Test of Conditional Moment Inequalities," *Unpublished Manuscript*.

Ciliberto, F., and E. Tamer (2009): "Market structure and multiple equilibria in airline markets," *Econometrica*, 77(6), 1791–1828.

Davydov, Y. A., M. A. Lifshits, and N. V. Smorodina (1998): *Local Properties of Distributions of Stochastic Functionals*. American Mathematical Society.

Galichon, A., and M. Henry (2009): "A test of non-identifying restrictions and confidence regions for partially identified parameters," *Journal of Econometrics*, 152(2), 186–196.

Ichimura, H., and P. E. Todd (2007): "Chapter 74 Implementing Nonparametric and Semiparametric Estimators," vol. Volume 6, Part 2, pp. 5369–5468. Elsevier.

Imbens, G. W., and C. F. Manski (2004): "Confidence Intervals for Partially Identified Parameters," *Econometrica*, 72(6), 1845–1857.

Khan, S., and E. Tamer (2009): "Inference on endogenously censored regression models using conditional moment inequalities," *Journal of Econometrics*, 152(2), 104–119.

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

Kim, K. i. (2008): "Set estimation and inference with models characterized by conditional moment inequalities," .

Koenker, R., and G. Bassett (1978): "Regression Quantiles," *Econometrica*, 46(1), 33–50.

——— (1982): "Robust Tests for Heteroscedasticity Based on Regression Quantiles," *Econometrica*, 50(1), 43–61.

Kosorok, M. R. (2008): *Introduction to Empirical Processes and Semiparametric Inference*.

Lee, S., K. Song, and Y. Whang (2011): "Testing functional inequalities," *Unpublished Manuscript*.

Lehmann, E. L., and J. P. Romano (2005): *Testing statistical hypotheses*. Springer.

Lepski, O., and A. Tsybakov (2000): "Asymptotically exact nonparametric hypothesis testing in sup-norm and at a fixed point," *Probability Theory and Related Fields*, 117(1), 17–48.

Manski, C. F. (1990): "Nonparametric Bounds on Treatment Effects," *The American Economic Review*, 80(2), 319–323.

Manski, C. F., and E. Tamer (2002): "Inference on Regressions with Interval Data on a Regressor or Outcome," *Econometrica*, 70(2), 519–546.

Menzel, K. (2008): "Estimation and Inference with Many Moment Inequalities," *Preprint, Massachussetts Institute of Technology*.

Moon, H. R., and F. Schorfheide (2009): "Bayesian and Frequentist Inference in Partially Identified Models," *National Bureau of Economic Research Working Paper Series*, No. 14882.

Pagan, A., and A. Ullah (1999): *Nonparametric econometrics*. Cambridge University Press.

Pakes, A., J. Porter, K. Ho, and J. Ishii (2006): "Moment Inequalities and Their Application," *Unpublished Manuscript*.

Pitt, L. D., and L. T. Tran (1979): "Local Sample Path Properties of Gaussian Fields," *The Annals of Probability*, 7(3), 477–493.

Politis, D. N., J. P. Romano, and M. Wolf (1999): *Subsampling*. Springer.

Pollard, D. (1984): *Convergence of stochastic processes*. Springer, New York, NY.

Ponomareva, M. (2010): "Inference in Models Defined by Conditional Moment Inequalities with Continuous Covariates," .

Romano, J. P., and A. M. Shaikh (2008): "Inference for identifiable parameters in partially identified econometric models," *Journal of Statistical Planning and Inference*, 138(9), 2786–2807.

Romano, J. P., and A. M. Shaikh (2010): "Inference for the Identified Set in Partially Identified Econometric Models," *Econometrica*, 78(1), 169–211.

Romano, J. P., and A. M. Shaikh (2012): "On the uniform asymptotic validity of subsampling and the bootstrap," *The Annals of Statistics*, 40(6), 2798–2822.

Stoye, J. (2009): "More on Confidence Intervals for Partially Identified Parameters," *Econometrica*, 77(4), 1299–1315.

van der Vaart, A. W., and J. A. Wellner (1996): *Weak convergence and empirical processes*. Springer.

Wright, J. H. (2003): "Detecting Lack of Identification in Gmm," *Econometric Theory*, 19(02), 322–330.

Figure 1: Case with faster than root-n convergence of KS statistic

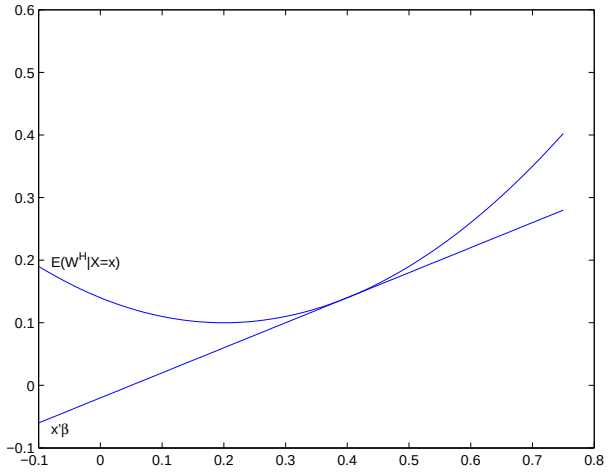


Figure 2: Cases with root-n convergence of KS statistic (β_1) and faster rates (β_2)

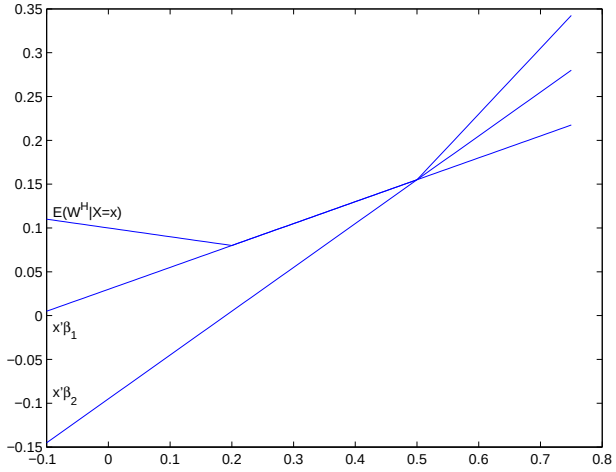


Figure 3: Conditional Means of W_i^H and W_i^L for Design 1

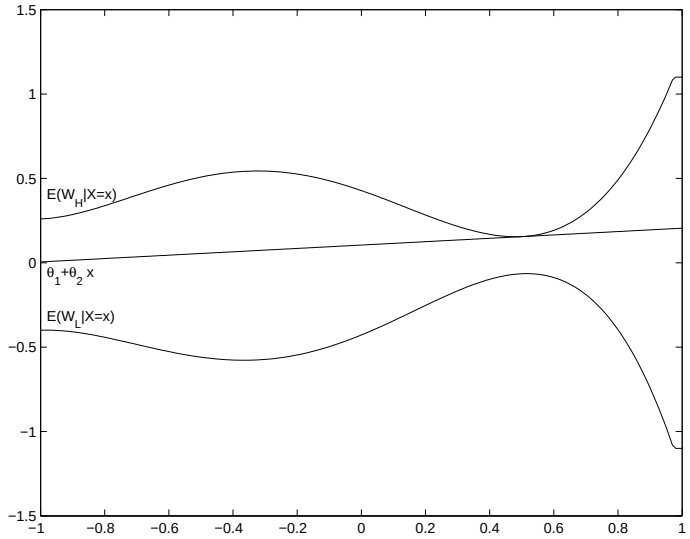


Figure 4: Conditional Means of W_i^H and W_i^L for Design 2

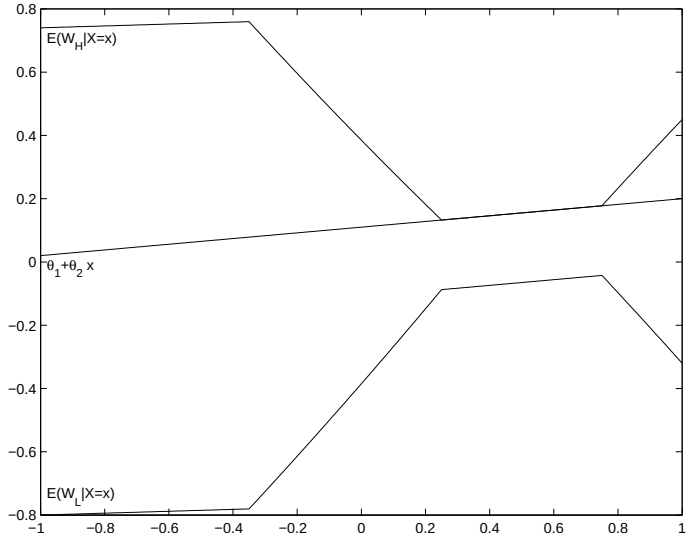


Figure 5: Histograms for $n^{3/5}S(T_n(\theta))$ for Design 1 ($n^{3/5}$ Convergence)

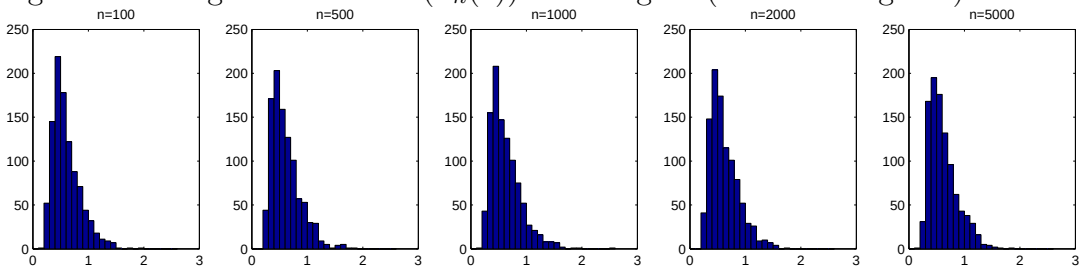


Figure 6: Histograms for $n^{1/2}S(T_n(\theta))$ for Design 2 ($n^{1/2}$ Convergence)

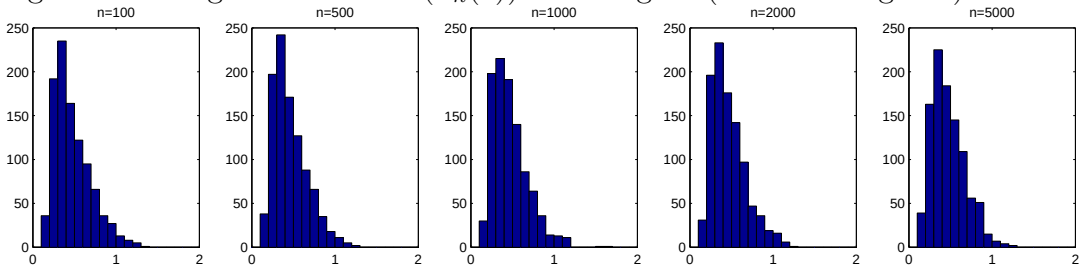


Figure 7: Data for Empirical Illustration

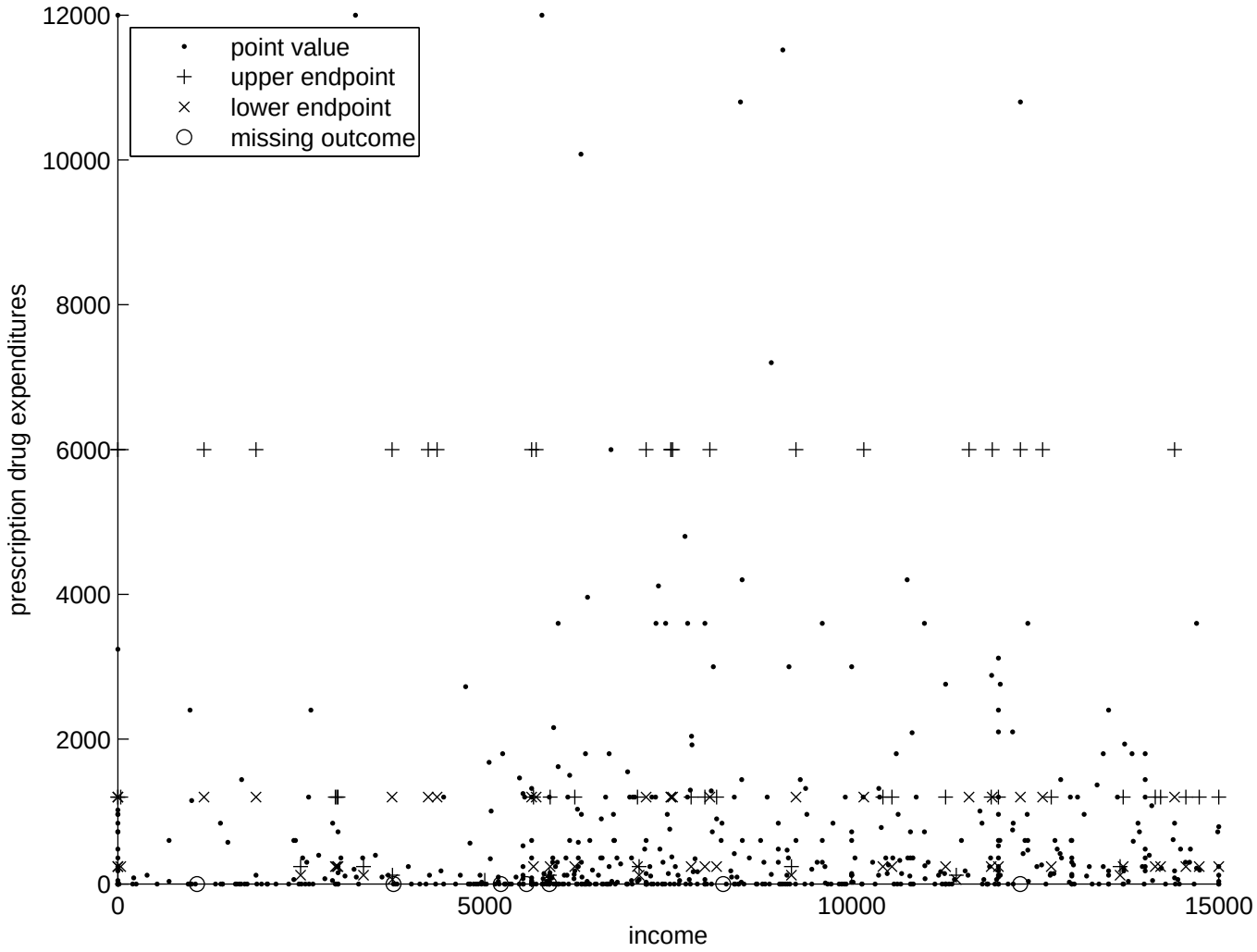


Figure 8: 95% Confidence Region Using Estimated Rate

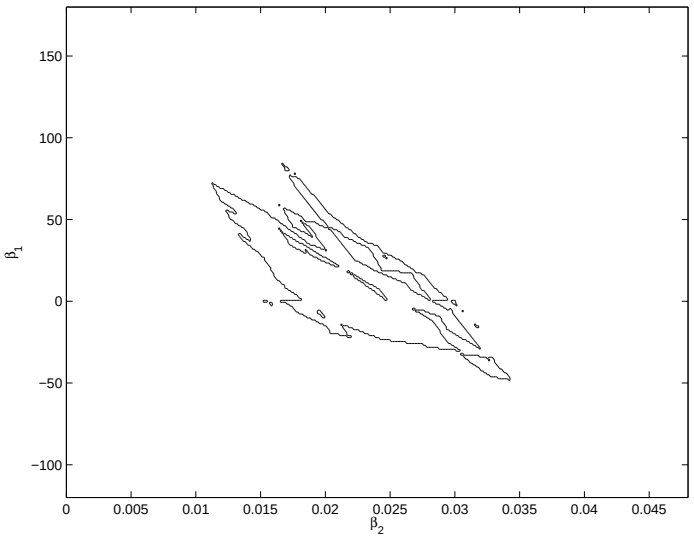


Figure 9: 95% Confidence Region Using Conservative Rate

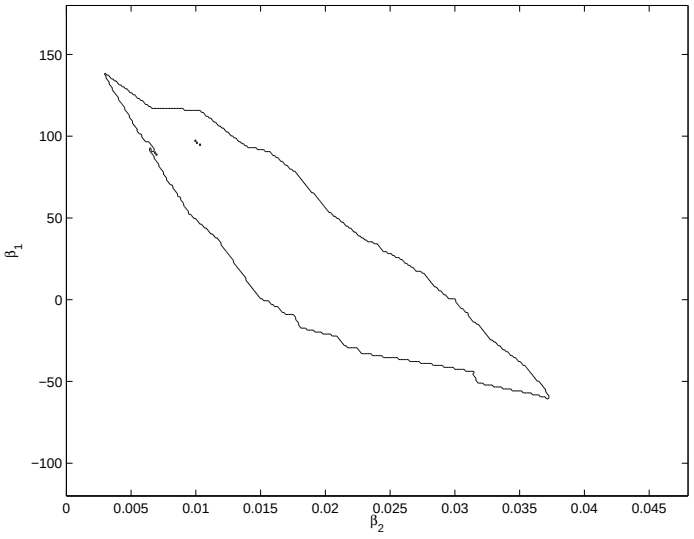


Figure 10: Convex Hull of 95% Confidence Region Using Estimated Rate

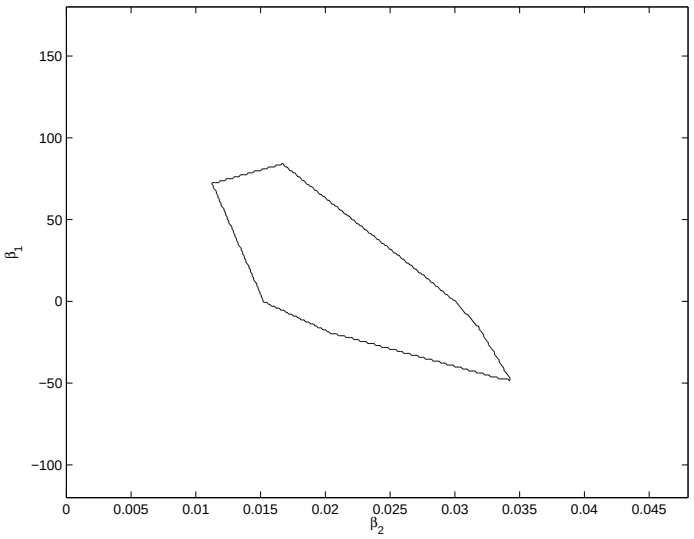


Figure 11: Convex Hull of 95% Confidence Region Using Conservative Rate

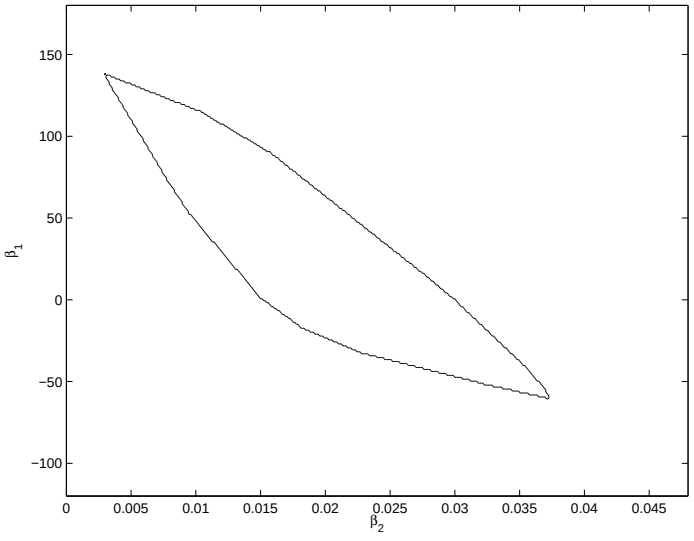
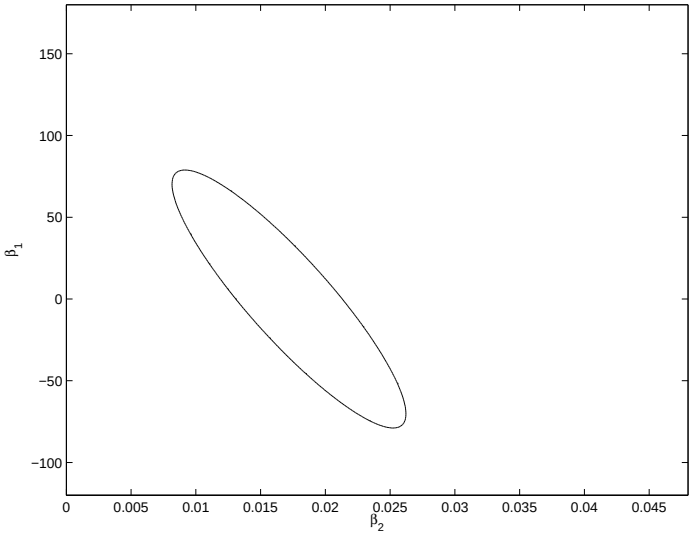


Figure 12: 95% Confidence Region Using LAD with Points



	$n = 100$	$n = 500$	$n = 1000$	$n = 2000$	$n = 5000$
nominal 90% coverage					
estimated rate	0.873	0.890	0.897	0.889	0.879
conservative rate ($n^{1/2}$)	0.991	0.987	0.987	0.995	0.996
(infeasible) exact rate ($n^{3/5}$)	0.921	0.909	0.905	0.903	0.890
nominal 95% coverage					
estimated rate	0.940	0.943	0.954	0.947	0.934
conservative rate ($n^{1/2}$)	0.998	1.000	0.998	1.000	0.999
(infeasible) exact rate ($n^{3/5}$)	0.976	0.965	0.949	0.956	0.953

Table 1: Coverage Probabilities for Design 1

	$n = 100$	$n = 500$	$n = 1000$	$n = 2000$	$n = 5000$
nominal 90% coverage					
estimated rate	0.780	0.910	0.928	0.925	0.924
conservative rate ($n^{1/2}$)	0.949	0.947	0.938	0.932	0.924
(infeasible) exact rate ($n^{1/2}$)	0.949	0.947	0.938	0.932	0.924
nominal 95% coverage					
estimated rate	0.885	0.945	0.966	0.971	0.979
conservative rate ($n^{1/2}$)	0.991	0.982	0.975	0.974	0.979
(infeasible) exact rate ($n^{1/2}$)	0.991	0.982	0.975	0.974	0.979

Table 2: Coverage Probabilities for Design 2

	$n = 100$	$n = 500$	$n = 1000$	$n = 2000$	$n = 5000$
nominal 90% coverage					
estimated rate	0.26	0.13	0.08	0.06	0.03
conservative rate ($n^{1/2}$)	0.33	0.17	0.12	0.09	0.06
(infeasible) exact rate ($n^{3/5}$)	0.21	0.10	0.07	0.05	0.03
nominal 95% coverage					
estimated rate	0.35	0.17	0.11	0.07	0.05
conservative rate ($n^{1/2}$)	0.39	0.22	0.15	0.11	0.07
(infeasible) exact rate ($n^{3/5}$)	0.29	0.13	0.09	0.06	0.04

Table 3: Mean of $\hat{u}_{1-\alpha} - \theta_{1,D1}$ for Design 1

	$n = 100$	$n = 500$	$n = 1000$	$n = 2000$	$n = 5000$
nominal 90% coverage					
estimated rate	0.11	0.08	0.06	0.04	0.02
conservative rate ($n^{1/2}$)	0.20	0.09	0.06	0.04	0.02
(infeasible) exact rate ($n^{1/2}$)	0.20	0.09	0.06	0.04	0.02
nominal 95% coverage					
estimated rate	0.18	0.10	0.07	0.05	0.03
conservative rate ($n^{1/2}$)	0.27	0.11	0.08	0.05	0.03
(infeasible) exact rate ($n^{1/2}$)	0.27	0.11	0.08	0.05	0.03

Table 4: Mean of $\hat{u}_{1-\alpha} - \theta_{2,D2}$ for Design 2

	θ_1	θ_2
Estimated Rate	$[-48, 84]$	$[0.0113, 0.0342]$
Conservative Rate	$[-60, 138]$	$[0.0030, 0.0372]$
LAD with Points	$[-63, 63]$	$[0.0100, 0.0244]$

Table 5: 95% Confidence Intervals for Components of θ