2000-04

UNIVERSITY OF CALIFORNIA, SAN DIEGO

DEPARTMENT OF ECONOMICS

BOOTSTRAPPING THE INFORMATION MATRIX TEST

BY

CHRISTOPHER STOMBERG

AND

HALBERT WHITE

DISCUSSION PAPER 2000-04
APRIL 2000

Bootstrapping the Information Matrix Test

Christopher Stomberg

and

Halbert White

University of California, San Diego

Abstract: In this paper we provide considerable Monte Carlo evidence on the finite sample performance of several alternative forms of White's [1982] IM test. Using linear regression and probit models, we extend the range of previous analysis in a manner that reveals new patterns in the behavior of the asymptotic version of the IM test – particularly with respect to curse of dimensionality effects. We also explore the potential of parametric and nonparametric bootstrap methods for reducing the size bias that characterizes the asymptotic IM test. The nonparametric bootstrap is of particular interest because of the weak conditions it imposes, but the results of our Monte Carlo experiments suggest that this technique is not without limitations. The parametric bootstrap demonstrates good size and power in reasonably small samples, but requires assumptions that may be auxiliary from the standpoint of a QMLE. We observe that the effects of violating one of these auxiliary assumptions has a non-trivial impact on the size of IM tests that employ this technique.

Keywords: information matrix testing, specification testing, misspecification, QMLE, nonparametric bootstrap, parametric bootstrap, Monte Carlo, White test, probit model, linear regression model.

Bootstrapping the Information Matrix Test

Christopher Stomberg

and

Halbert White

University of California, San Diego

1 Introduction

The rapid increase in computational power available to researchers has had a tremendous impact on the efficiency with which econometric modeling can be performed. Where it was once time-consuming and expensive to evaluate multiple model specifications, it is now quite feasible to evaluate several competitive specifications in a single study. In an environment rich with potential models, it becomes ever more important to have a single, reliable specification test that can be used to aid the process of model selection and specification. This statistic would encapsulate several of the most important aspects of specification testing into one number, and serve as a thumbnail sketch of the quality of model specification. Though such a statistic would not obviate the need for more in-depth specification testing, it would serve as a good filter to narrow the field of candidates.

The information matrix (IM) testing procedure introduced by White in 1982 offers a conceptually appealing way to perform just this kind of omnibus specification testing on quasi-maximum likelihood models. It is a statistical test of the hypothesis that the information matrix equality holds, that is, of the hypothesis

$$H_0 : E(\nabla^2 \log f(X,q_0)) = E(\nabla \log f(X,q_0) \cdot \nabla' \log f(X,q_0)),$$

where $\log f(\cdot,q_0)$ is the log-likelihood for the random variable X, q_0 is probability limit of the associated (quasi-) maximum likelihood estimator, and $\nabla \log f$, and $\nabla^2 \log f$ are the gradient (score) vector and Hessian matrix of the log-likelihood respectively.

A variety of test statistics can be formed from subsets of the elements in the equality, or all of the elements can be used in a full IM test. This family of tests has the flexibility to check specifications on a wide range of econometrically important models, yet can also be used to focus on the individual aspects of specification that are relevant within each type of model. Perhaps one of the biggest advantages of the full IM test, though, is the fact that it directly or indirectly tests all of the major assumptions built into the maximum likelihood model that has been specified. Passing this test is an important signal that the fitted model adheres to the specification that has been set forth. These properties would seem to make the full IM test an ideal tool for preliminary model specification checks. Unfortunately, this test is not without some serious deficiencies.

The main issue with the IM test is that, in practice, its observed rejection rate under the null hypothesis is often far from the nominal rate. Earlier Monte Carlo studies such as those done by Taylor [1987], Orme [1990], Chesher and Spady [1991], or Davidson and MacKinnon [1992] have consistently shown a dramatic upward size bias in the full IM test in small samples. In some of the worst cases, rejection rates near 100% have been observed where 5% is the nominal rate. Significant upward bias in the rejection rates has also been documented in samples with well over one thousand observations. This kind of performance has made the information matrix test virtually impossible to use in practice and has sparked a search for ways to fix the problems perceived with it. There remain, however, important aspects of the distribution of this test that still have not been well documented, and, with the possible exception of the application of the parametric bootstrap by Horowitz [1994], most attempts to fix the test have had somewhat limited applicability or success.

The principal problem with the full IM test statistic resides with the mismatch between the approximate chi-square distribution given by asymptotic theory and the sampling distribution observed in practice. There are essentially two ways to handle this problem. One can either modify the statistic so that in smaller samples it behaves more like the asymptotic approximation, or one can modify the distributional approximation so that it is more in line with the observed sampling distribution. Among those taking the first approach are Taylor [1987], Orme [1990], and Davidson and MacKinnon [1992]. Most examples of this approach are either limited in their applicability, or have met with limited success. Examples of the second approach include Taylor [1989], Chesher and Spady [1991], and the application of the bootstrap by Horowitz [1994]. These attempts to approximate the distribution better have generally led to broader success.

Perhaps the earliest response to the size problem was Taylor's [1987] attempt to reduce the dimension of the statistic as a response to the so called "curse of dimensionality". Taylor used a regression-based form of the IM test on a linear regression model, and removed elements from the score vector that normally appear in the artificial regression for the particular tests he used. This resulted in a statistic that is asymptotically equivalent to White's original formulation only under special circumstances, and also did not have markedly improved size performance.

Orme [1990] introduced a more sophisticated approach to modifying the information matrix test statistic. He explored the idea of plugging in expected values for some of the sample quantities appearing in the IM test statistic. His most successful ω_3 formulation involved application of the null hypothesis to allow the use of the negative outer product of the score vector in place of the Hessian matrix, and also substituted expected values for the gradient vector of the indicator vector. The result is asymtotically equivalent to the original statistic, but involves computations that may be more stable in small samples. The size performance of this statistic was found to be considerably improved over the original formulation with rejection rates closer to 10% under the null versus the nominal 5%. This size performance came at the cost of power. Horowitz's [1994] experiments showed that Orme's ω_3 version of the IM test has little power against certain alternative models.

Davidson and MacKinnon [1992] took a completely different approach to modifying the computation of the IM test. They start from Chesher's [1984] observation that, within regression models, the IM test is equivalent to a test of parameter heterogeneity. They then constructed a direct test for parameter heterogeneity that deviates computationally from the IM test, but maintains asymptotic equivalence. Though their version of the test is only valid for linear and nonlinear regression models, the Davidson-MacKinnon formulation results in very good size performance. The power of their statistic, however, has not been tested. The drawback of this approach is its limitation to linear and non-linear regression models, which constrains its generality.

The first attempt to improve the approximation of the sampling distribution of the IM test statistic was also limited in its range of applicability, though for different reasons. Taylor's [1989] approach involved mapping out the Monte Carlo distribution of the statistic, and then using this information to estimate an adjustment factor for the original chi-square distribution function. The adjustment factor is based on the method of extended rational approximants which is applied to the chi-square distribution function. Taylor actually applied this technique directly to the inverse cumulative distribution function, which results in the approximation

$$F^{-1}(a\,, q, n) = u_{a\,,q} \cdot G(a\,, q, n), \text{with } u_{a\,,q} = (\,c^{\,2}_{a\,,q})^{-1},$$

where G is the adjustment function consisting of a ratio of two polynomials of up to order three in their arguments. This function modifies the form of the original chi-square with a non-linear (cubic) function that explicitly accounts for changes in the nominal rejection rate, the number of elements in the test vector, and the sample size. In practice, the rational approximants in G need to be estimated, and to do this a grid of Monte Carlo estimates of the rejection rates are required over a grid of values in n, a, and q. Taylor's results are especially intriguing because the adjustment factor he estimates gives some indication of exactly how the sampling distribution deviates from the asymptotic distribution. The main drawback of this type of approximation technique is its reliance on the results of a large number of Monte Carlo simulations. It is particularly computer intensive because simulations are required not only for the particular sample size and number of parameters at hand, but for a whole grid of other possibilities as well.

Chesher and Spady [1991] formulated a correction factor for the chi-square distribution based on analytical techniques, obviating the need for heavy simulation. Their formulation starts by computing Cornish-Fisher expansions of the distribution of the IM statistic to obtain a second-order approximation of this distribution. This results in a reduction of the theoretical approximation error to O(n^{-1}) rather than the O($n^{-1/2}$) yielded by traditional central limit theory. This adjustment equation is a nonlinear combination of higher-order chi-squares and n^{-1} applied to the asymptotic c_q^2. The formula for their approximation can be written as:

$$F^{-1}(a,q,n) = u_{a,q} \cdot \left[1 + n^{-1} \cdot G(u_{a,q}, q)\right],$$

where $u_{a,q} = (c^2_{a,q})^{-1}$. The Chesher-Spady adjustment function G contains powers of two in both its arguments, so it is quite non-linear. Though the use of a second-order Cornish-Fisher expansion limits the order of n involved in the correction, the benefits of this approach for the IM test are significant. Chesher and Spady obtain distributional approximations very close to the observed performance in a variety of settings. As for Taylor's approach, a secondary benefit of this correction is that it helps identify the magnitude of errors committed by using the standard O(1) c^2 approximation. On the negative side, their approximation results in a non-monotonic distribution function for small n. In one example, the C-F modified distribution becomes non-monotonic in samples of under 173 observations. Another difficulty in using Cornish-Fisher expansions of the IM statistic is their cumbersome calculation.

The parametric bootstrap explored by Horowitz [1994] offers a computationally less cumbersome approach to approximating the sampling distribution of the

IM test statistic. Horowitz's simulation results for the tobit and probit regression cases he studied are remarkable. In fact, the rejection rates he observed are essentially identical to the nominal rates. This is a greater improvement than might naively be expected from a second-order improvement in approximation. The parametric bootstrap also has the clear advantage of simplicity over the other approximation techniques. One need not perform an extensive Monte Carlo study or calculate hundreds of cumulants in order to benefit from the parametric bootstrap. One simply needs a fully parameterized DGP to stand in as the null model from which a sampling distribution for the IM test statistic is generated. It is, however, a requirement of the parametric bootstrap that a fully parameterized DGP be available, which is what limits its general applicability for QMLE's.

Linear and nonlinear regression models represent an example of a family of quasi-maximum likelihood models that do not rely on specific parametric DGP assumptions in order for asymptotic normality to obtain. In this case, using the parametric bootstrap requires simulating errors from a given parametric distribution such as the normal distribution function. Doing this imposes structure that is not necessary to the functioning of the model or the IM test. From this standpoint, it is of interest to explore alternative bootstrap approximation techniques that do not require auxiliary conditions beyond those already needed for application of a central limit theorem to the QMLE. The nonparametric bootstrap is such a technique, and is a particular focus of our attention.

In this paper we investigate the potential of various methods, including the nonparametric bootstrap, to eliminate the size bias problem in information matrix testing. Further, we re-evaluate the existing Monte Carlo evidence on the finite sample performance of various forms of the IM test. Both the number of regressors and the range of sample sizes are greater here than in previous studies in order to more clearly map the performance of the various forms of the information matrix test. This provides a more complete picture of the problems that arise in obtaining the sampling distribution of the IM test statistics. In fact, we show that it is possible, in a variety of ordinary situations, for the rejection rate of the IM test to actually *rise* with *n* and *fall* with increases in the number of regressors - exactly the opposite of what one might expect given previous research.

Our Monte Carlo experiments for the bootstrap demonstrate that a straightforward application of the nonparametric bootstrap is not necessarily as capable as the parametric bootstrap at eliminating size bias in the IM test. This is to be expected because of the weak structure imposed by the nonparametric bootstrap. In fact, the Monte Carlo experiments reported in this paper focus on situations where the parametric bootstrap appears in its most favorable light. We touch briefly upon IM testing in environments where the nonparametric bootstrap is the favored approach, and here the parametric bootstrap does less well. The results of these experiments suggest that this is an important area for further research.

This paper is organized as follows: Section 2 contains a brief overview of the IM test and, for concreteness and because of its first use in this context, how the nonparametric bootstrap applies. Section 3 presents Monte Carlo results that provide a baseline for the asymptotic forms of the information matrix test. Here we re-evaluate the performance of the IM test statistics and reveal properties that have not been well documented previously. This is followed in section 4 by a comparison of the performance of nonparametric and parametric bootstrap methods under the null hypothesis and under two alternatives. These results indicate the promise of the nonparametric bootstrap technique for correcting the size problems of the IM test, but also some of its limitations. The power experiments, though not exhaustive, suggest weaknesses in all forms of the IM test in small samples. The paper concludes with a discussion of directions for future research.

2 The IM Test Framework and the Bootstrap

One of the appealing aspects of the IM test is the consistent specification testing framework that it offers over a wide range of useful econometric models. Because it uses the information matrix equality as a measure of goodness of fit, it is as at home testing heteroskedasticity in a linear regression model as it is in testing overdispersion in a probit equation. Because of the documented weakness of the asymptotic approximation for this statistic in smaller finite samples it is a natural candidate for the application of the bootstrap which may provide a better level of approximation.

2.a The IM Test

The IM test exploits the fact that if a quasi-maximum likelihood estimator (QMLE) is correctly specified, the information matrix equality will hold. This equality states that under correct specification, the expected value of the matrix of second derivatives of the log-likelihood function is equivalent to the opposite of the expected value of the outer product of the vector of first derivatives. Deviations from correct specification cause these two matrices to diverge, so it is possible to use deviations from this equality as a signal of misspecification.

Following the notation of White [1994], the following information matrix equality holds for a QMLE under correct specification:

$$A_n^* = -\overline{B}_n^*,$$

where $A_n^* \equiv n^{-1}\sum_{t=1}^{n} E[\nabla' s_t^*]$, and $\overline{B}_n^* \equiv n^{-1}\sum_{t=1}^{n} E[s_t^* s_t^* ']$, and where $s_t^* = \nabla \log f_t(X_t, q_n^*)$.

The function $\log f_t$ is the log-likelihood function for the random variable X_t, and $q_n^* = \operatorname{plim} \hat{q}_n$, where $\hat{q}_n$ is the (quasi) maximum likelihood estimator. For simplicity, we assume the process X_t to be independent and identically distributed. The sample quantity, $\hat{A}_n + \hat{B}_n = 0$, accomplishes the matrix comparison that is the basis for the test. This comparison matrix is symmetric, which means that there are $p(p+1)/2$ unique elements in this expression. The actual test statistic is formed by vectorizing this matrix and then picking a set of unique elements from the resulting vector. For the linear regression normal error case, the full vector of indicators contains elements that are sensitive to conditional heteroskedsticity, as well as non-normal skewness and kurtosis. Section 3, which introduces the Monte Carlo experiments, presents specific examples of the terms entering the test statistic for both the linear regression and probit cases. See also White [1994, ch. 11] for other examples.

Since several directions are simultaneously tested, rejection of the null hypothesis does not give specific information about the exact cause of the rejection.

A representative summand in the IM test statistic is

$$\underset{q\times 1}{\hat{m}_t} = S \cdot vec(\nabla'\hat{s}_t + \hat{s}_t\hat{s}_t'),$$

where S is a selector matrix composed of columns of zeroes and ones arranged to pick out particular elements of the indicator vector. Setting S to be the identity matrix yields the full information matrix test. Careful selection of the indicators in the IM test vector with S can yield more precise information about the nature of misspecification. A familiar example of such a test is the White test for heteroskedasticity which selects only terms involving conditional variance.

The sample statistic formed from the vector $\hat{m}_t$ is:

$$\sqrt{n}\hat{M}_n = \sqrt{n}(\frac{1}{n}\sum_{t=1}^{n}\hat{m}_t).$$

Conditions ensuring the asymptotic normality of this statistic are given by White [1994] theorem 9.2. These include enough smoothness to guarantee that sufficient derivatives exist and that the moments used in the computation of the covariance matrix are bounded. Given this, a statistic that is asymptotically distributed c^2 under the null hypothesis can be formed as

$$\hat{\mathcal{M}}_n = n\hat{M}'_n \hat{J}_n^{-1} \hat{M}_n \overset{A}{\sim} c_q^2 . \qquad (1)$$

This is the information matrix test statistic.

The estimator of the covariance matrix, $\hat{J}_n$, can be cumbersome to compute because it involves the gradient of $\hat{m}_t$ which requires computation of third derivatives of the quasi log-likelihood function. Two important forms of this test statistic correspond to different choices for the estimator $\hat{J}_n$. A form of the covariance matrix corresponding to corollary 9.10 of White [1994] that is appropriate where serial dependence is absent is the following:

$$\hat{J}_n = n^{-1} \sum_{t=1}^{n} (\hat{m}_t - \nabla'_q \hat{M}_n \hat{A}_n^{-} \hat{s}_t)(\hat{m}_t - \nabla'_q \hat{M}_n \hat{A}_n^{-} \hat{s}_t)' .$$

This version, which involes the third derivatives of the likelihood function just mentioned, is the original form specified by White [1982].

The null hypothesis can be exploited to simplify the statistic by causing the third derivatives to vanish. A commonly used version of the covariance matrix under these conditions obtained by Lancaster [1983] and Chesher [1983], and following directly from corollary 9.11 in White [1994] is the following:

$$\hat{J}_n = n^{-1} \sum_{t=1}^{n} \hat{m}_t \hat{m}_t' - (n^{-1} \sum_{t=1}^{n} \hat{m}_t \hat{s}_t')[n^{-1} \sum_{t=1}^{n} \hat{s}_t \hat{s}_t']^{-1}(n^{-1} \sum_{t=1}^{n} \hat{s}_t \hat{m}_t') .$$

This variation of the test is often referred to as the outer product of the gradient (OPG) version; we refer to this as the Chesher-Lancaster form.

Note that both of these forms of the covariance matrix make use of the null hypothesis to do away with terms involving $E(m_t^*)$ that would otherwise need to appear in order to take account of the alternative, $E(m_t^*) \neq 0$. We study both the White form, and the Chesher-Lancaster forms of the statistic throughout.

A simple implementation often used for the computation of the test statistic defines the vector $\hat{x}_t$ for each of the two versions of the statistic as

$$\hat{x}_t = \hat{m}_t - \nabla_q' \hat{M}_n \hat{A}_n^{-} \hat{s}_t , \qquad \text{(2a)} \qquad \text{(White)}$$

$$\hat{x}_t = \{\hat{m}_t, \hat{s}_t\} . \qquad \text{(2b)} \qquad \text{(Chesher-Lancaster)}$$

The IM test statistic can be computed as:

$$\hat{\mathcal{M}}_n' = n \cdot R^2 \qquad (3)$$

where R^2 is the uncentered R^2 from an artificial regression of a vector of ones on the vector $\hat{x}_t$. Under conditions given by White [1994] in corollaries 9.10b and 9.11b

$\mathcal{M}_n' - \mathcal{M}_n \xrightarrow{P_0} 0$, so this simplified version is asymptotically distributed c_q^2 as well.

2.b The Bootstrap

The overriding problem with the IM test statistic is that asymptotic approximations poorly approximate the finite sampling distributions actually observed. Ideally, we would like to have full knowledge of the sampling distribution function behind a complicated statistic like this. Alternatively, we might want to repeatedly sample from the population in order to map out the sampling distribution. However, this type of population information is not available to the researcher except in circumstances of Monte Carlo simulation. The bootstrap is a technique that mimics the process of sampling repeatedly from the population, by instead re-sampling repeatedly from the sample data. Introduced by Efron [1979], with considerable refinement by Freedman [1981], Bickel and Freedman [1981], Hall [1986a,b,1987,1992] and others, this method essentially uses sample data to construct a numerical simulation of the sampling distribution of a statistic. Though the bootstrap is not essentially numerical in nature, its application is generally expedited using Monte Carlo sampling techniques. Depending on the circumstances of application, the bootstrap yields an approximation at least as good as an asymptotic approximation. The accuracy of the bootstrap approximation depends in part on the (asymptotic) pivotalness of the statistic it is applied to. A pivotal statistic has a distribution that is free of estimated parameters, whereas an asymptotically pivotal statistic has a limiting distribution that is free of estimated parameters. In many leading cases, where pivotal or asymptotically pivotal statistics exist, the bootstrap delivers faster convergence than standard asymptotic methods. The following discussion draws on Hall [1992].

To describe the set up, we start with a sample, $\mathcal{X} = \{X_1,...,X_n\}$, which consists of n repeated draws of the i.i.d. random variable X having the population distribution function F_0. F_0 is unknown to the researcher, but some of its characteristics are revealed through $\mathcal{X}$ which is governed by the sampling distribution function, F_1. A bootstrap resample, $\mathcal{X}^* = \{X_1^*,...,X_n^*\}$, can, in turn, be constructed by selecting n elements randomly with replacement from the sample $\mathcal{X}$. Each X_i^* has probability n^{-1} of being selected into this new artificial sample:

$$P(X_i^* = X_j | \mathcal{X}) = n^{-1}, \quad 1 \leq i,j \leq n.$$

The distribution of the resample is governed by the bootstrap distribution function, F_2, which can be estimated to an arbitrary degree of accuracy. This can be done by drawing repeated resamples and mapping the resulting distribution. This repeated resampling from F_1 makes it possible to generate a family of simulated samples that generally imitates the behavior of repeated sampling from the population distribution. It is this property that gives the bootstrap its appeal.

Consider a population parameter of interest, say $q(F_0)$. Because this is based on the population, $q(F_0)$ is unknown to the researcher. One must therefore construct an estimator of this parameter based on sample data; call this estimator

$q(F_1)$. For statistical inference, knowledge of the distribution of the sample parameter estimate about the population parameter is needed, this is denoted:

$$\mathcal{D}\{\sqrt{n}[q(F_1)-q(F_0)]\,|\,F_0\}\,.$$

$\mathcal{D}$ contains all of the information about the distribution that we might want, but it involves unknown quantities from the population distribution function *F₀*. The usual procedure at this point is to apply central limit theory to prove that $\mathcal{D}$ converges asymptotically to some simple and analytically tractable distribution such as the normal distribution. e.g.

$$\mathcal{D}\{\sqrt{n}[q(F_1)-q(F_0)]\,|\,F_0\}\xrightarrow{a.s.}N(0,D)\,,$$

where D is the asymptotic covariance matrix of $q(F_1)$.

Inference can then proceed based on the similarity between $\mathcal{D}$ and its limiting normal distribution. In contrast, the bootstrap proceeds by plugging in estimates of the components inside $\mathcal{D}$. For the population quantities, one plugs in estimates based on the sample, and for the sample quantities, one plugs in estimates based on the bootstrap resamples. This plug-in approach is perfectly feasible as both sample and bootstrap estimators are known quantities. Using asymptotic theory it is straightforward to prove under mild conditions that the bootstrap plug in estimator converges in distribution to the sample estimator:

$$\mathcal{D}\{\sqrt{n}[q(F_2)-q(F_1)]\,|\,F_1\}\xrightarrow{a.s.}\mathcal{D}\{\sqrt{n}[q(F_1)-q(F_0)]\,|\,F_0\}\,,$$

where F_2 stands for the bootstrap distribution, and $q(F_2)$ is an estimate drawn from this distribution. See Bickel and Freedman [1981] Theorem 2.1 for a proof of this for the sample mean.

Often, the researcher is only interested in particular aspects of the distribution of a statistic, which simplifies the situation. This can be treated as the solution t_0 to the problem

$$\{t_0 \mid E[f_t(F_0, F_1) \mid F_0] = 0, t \in T\},$$

where $f_t(F_0, F_1)$ is a function that appropriately embodies the aspect of the distribution we care about. For example, if we are interested in the sample bias, we would use

$$f_t(F_0, F_1) = q(F_1) - q(F_0) + t,$$

or if we are interested in a symmetric two-sided confidence interval we could use

$$f_t(F_0, F_1) = I\{q(F_1) - t \le q(F_0) \le q(F_1) + t\} - 0.95,$$

where *I{⋅}* is an indicator function equal to one if its argument is true, and zero otherwise. Taking the expected value of this function yields a probability estimate. Since the population quantities involved in this formula are unknown, the bootstrap can be applied to generate an analogous expression:

$$\{t_1 \mid E[f_t(F_1, F_2) \mid F_1] = 0, t \in T\}.$$

A procedure for estimating the bootstrap distribution F_2 is to repeatedly resample from the sample data B times, creating a set of resamples F_{2b}, $b=1,...,B$. The distributional behavior of the resample statistics will yield the information needed for inference. For example, in the case of the confidence interval calculations, we are interested in

$$f_t(F_1, F_2) = I\{q(F_2) - t_1 \leq q(F_1) \leq q(F_2) + t_1\} - 0.95 .$$

This is a neighborhood around the sample value that contains 95% of the bootstrap values. Numerically simulating the bootstrap distribution by resampling the data in this manner is referred to as the Monte Carlo method. The actual bootstrap distribution is defined only as $B \to \infty$, but can usually be well approximated with a reasonable number of simulations. Hall [1986b] has a discussion of appropriate settings for B in some leading cases.

The intuition behind this procedure is that the sample data reflects our best knowledge of the population distribution through the frequency distribution of its values. By repeatedly drawing at random from the sampling distribution one is simulating a situation in which repeated samples are drawn from the population. Infinite sampling from a well behaved population would give the observer perfect information about the population distribution function. In the same way, infinite sampling from the sampling distribution function can yield perfect information about the sampling distribution function. The key link for needed for inference is that the degree of similarity between the sampling distribution and the population distribution must be high.

Using Edgeworth expansion techniques one can prove for leading statistical estimators that the bootstrap is at least as accurate as the central limit approximation, and in many cases better. If the statistic in question is exactly pivotal, the bootstrap can yield arbitrarily close estimates of population quantities for the distribution. Typically test statistics in economics are at least asymptotically pivotal, in which case the bootstrap approximations are accurate to terms of $O(n^{-1})$. This is a considerably better approximation than the standard $O(n^{-1/2})$ afforded by central limit theory. Even if an asymptotic pivot is not available, the bootstrap can still yield an accuracy of $O(n^{-1/2})$. This result can be convenient in cases where asymptotic distributions are analytically intractable, and the bootstrap can be used to simulate the distribution of the statistic as well as could be expected from asymptotic theory. The limitation of Edgeworth expansions is their reliance on the existence and finiteness of high order population moments, as well as their analytical complexity. For an extended treatment of the bootstrap using Edgeworth expansions to analyze convergence properties see Hall [1992].

2.c Bootstrapping the IM test

The relatively weak assumptions needed to use the nonparametric bootstrap to develop confidence intervals can be advantageous in the context of QMLEs. With these estimators one may not need, or indeed have, the full MLE specification that is required to perform parametric bootstrap resampling. Generating bootstrap data using a parameterized error density function imposes additional structure that could be unwarranted. Not only is this structure unneccesary, it may well cause undesirable side-effects. The nonparametric bootstrap is appealing because it imposes no such additional structure.

One way to interpret the problem with the IM test procedure is that critical values drawn from asymptotic chi-square distribution are much too small. For this problem, we want to use the bootstrap to estimate an upper percentile for the IM test statistic's distribution. This can be laid out in terms of population quantities as the set E (an ellipse) such that

$$P\left(\sqrt{n}[M_n(F_1)-M_n(F_0)]\in E\right)=1-a \ ,$$

where a is the probability of a type I error. In other words, we want to find an ellipsoidal region such that $(1-a)\cdot 100\%$ of the sample statistics are are contained within it. Normalizing this expression with the appropriate covariance matrix, V_n, we get the equivalent expression in terms of the sphere, S, such that

$$P\left(\sqrt{n}V_n(F_1)^{-1/2}[M_n(F_1)-M_n(F_0)]\in S\right)=1-a \ .$$

This latter expression is our main interest because it is potentially an asymptotically pivotal statistic. Since this statistic generally converges to a c_q^2 which is indeed free of estimated parameters, we are assured of an asymptotic pivot in this case. It is this property that reduces errors in the approximation introduced by the presence of unknown parameters in the limit distribution. This delivers additional accuracy, to the order of O(n^{-1}).

The above formulation in terms of S can be translated into the following quadratic form where the parameter *t* is a measure the square of the radius of the sphere S:

$$f_t(F_0, F_1) = I\left\{n(M_n(F_1) - M_n(F_0))'V_n(F_1)^{-1}(M_n(F_1) - M_n(F_0)) \le t\right\} - (1 - a).$$

The expression inside the parenthesis is the sample IM test statistic. Under the null hypthesis, we would typically set $M_n(F_0)$ to zero which yields the form seen in equation (1). The parameter *t* is the sample estimate of the (*1-a*) percentile of the distribution of $\mathcal{M}_n$ and can be used as a critical value for hypothesis testing. The bootstrap plug-in estimator version of this statistic is:

$$f_t(F_1, F_2) = I\left\{n(M_n(F_2) - M_n(F_1))'V_n(F_2)^{-1}(M_n(F_2) - M_n(F_1)) \le t\right\} - (1 - a),$$

which can be re-written in less cumbersome notation as:

$$f_t(F_1, F_2) = I\left\{n(\hat{M}_n^* - \hat{M}_n)'\hat{V}_n^{*-1}(\hat{M}_n^* - \hat{M}_n) \le t\right\} - (1 - a).$$

Here, we are using the notation $q(F_0) = q^*, q(F_1) = \hat{q}$, and $q(F_2) = \hat{q}^*$, where asterisks indicate quantities estimated by bootstrap resampling. Finding the non-parametric bootstrap critical value involves solving

$$\{t \mid E[f_t(F_1, F_2) \mid F_1] = 0, t \in T\}$$

to get a value for *t*. Using averages to estimate expectations, we solve the following boostrap problem:

$$\{t \,|\left(\tfrac{1}{\mathrm{B}} \sum_{i=1}^{B} f_{ti}(F_1, F_2) \,|\, F_1 \right) = 0, t \in T\}\text{, or}$$

$$\{t \,|\left(\tfrac{1}{B} \sum_{i=1}^{B} I\left\{ n(\hat{M}^{*}_{ni} - \hat{M}_{n})' \hat{J}^{*-1}_{ni} (\hat{M}^{*}_{ni} - \hat{M}_{n}) \leq t \right\} - (1 - a) \right) = 0\}.$$

The nonparametric bootstrap resampling used to solve this equation is carried out by randomly sampling from the rows, the $(x_t, y_t)'s$, of the data matrix, with replacement. The QMLE is then estimated using the resampled data. The errors and data elements from this step are then used to compute the bootstrap indicator vector $\hat{M}^{*}_{n}$ and its covariance matrix, $\hat{J}^{*}_{n}$. From these, the centered bootstrap IM test statistic is computed as

$$\hat{\mathcal{M}}^{*}_{n} = n(\hat{M}^{*}_{n} - \hat{M}_{n})' \hat{J}^{*-1}_{n} (\hat{M}^{*}_{n} - \hat{M}_{n}). \tag{4}$$

Centering is critically important in the nonparametric bootstrap because of the possibility that one is resampling from a sample generated under the alternative hypothesis. Our goal in doing hypothesis testing using the nonparametric bootstrap is to recreate the distribution of the statistic under the null hypothesis, so alternatives (if present) must be purged from the bootstrap statistics. This affects the covariance estimator in addition to the mean of the statistic as explicitly established in equation 4.

The potential for the displacement of the mean of the bootstrap distribution is fairly obvious. Under the null hypothesis, $M_n^* = 0$, whereas under alternative hypotheses, $M_n^* \neq 0$, so $\hat{M}_n \xrightarrow{as} M_n^* \neq 0$. The nonparametric bootstrap yields $\hat{M}_n^* \xrightarrow{as} \hat{M}_n \neq 0$, so the center of the bootstrap distribution must be shifted back to the origin by subtracting the sample mean of the test indicators, which explains the appearance of $\hat{M}_n$ in the bootstrap statistic.

It must also be remembered that the use of $\hat{J}_n$ in the sample IM statistic is predicated on the correctness of the null hypothesis. Centering is needed in the nonparametric bootstrap to correct for the presence of possible alternatives affecting this covariance estimator. Define the following centered boostrap indicators for the IM statistic:

$$\hat{m}_t^* = \hat{m}_t^* - \hat{m}_t \,,$$

We have $E[\hat{m}_t^* \mid F_1] = 0$ under very general conditions, both under the null hypothesis and under alternatives. Versions of the bootstrap IM test covariance matrices that are valid under alternatives have the following forms:

White: $$\hat{J}_n^* = n^{-1} \sum_{t=1}^{n} (\hat{m}_t^* - \nabla_q' \hat{M}_n^* \hat{A}_n^{*-} \hat{s}_n^*)(\hat{m}_t^* - \nabla_q' \hat{M}_n^* \hat{A}_n^{*-} \hat{s}_n^*)' ,$$

CL: $$\hat{J}_n^* = n^{-1} \sum_{t=1}^{n} \hat{m}_t^* \hat{m}_t^{*\prime} - (n^{-1} \sum_{t=1}^{n} \hat{m}_t^* \hat{s}_t^{\prime *\prime})[n^{-1} \sum_{t=1}^{n} \hat{s}_t^* \hat{s}_t^{*\prime}]^{-1}(n^{-1} \sum_{t=1}^{n} \hat{s}_t^* \hat{m}_t^{*\prime}) .$$

Without the centering of $\hat{m}_t^*$, terms involving squares and crossproducts of the sample average of the indicators, $\hat{M}_n \neq 0$, would remain in $\hat{J}_n^*$. This would cause the bootstrap estimate of the covariance matrix under alternatives to be inflated relative to its estimate under the null hypothesis. The general impact of this is a reduction in the power of the procedure.

The mechanics of the nonparametric bootstrap IM test are straightforward. The statistic $\hat{\mathcal{M}}_{ni}^*$ is computed B times using different resamples each time. The results $(\hat{\mathcal{M}}_{n1}^*,...,\hat{\mathcal{M}}_{n\mathcal{B}}^*)$ are ordered, and the bootstrap statistic below which $(1-a)\times 100\%$ of the bootstrap observations occur is selected as a critical value. The sample IM test statistic is then compared with the bootstrap critical value as one would usually do with the asymptotic critical value. An equivalent procedure yields a bootstrap p-value by computing the quantity $\hat{P}^* = \frac{1}{B}\sum_{i=1}^{B} I(\hat{\mathcal{M}}_{ni}^* > \hat{\mathcal{M}}_n)$. Here one computes the share of bootstrap statistics exceeding the sample statistic. Rejection at the nominal 5% level occurs whenever the bootstrap p-value is less than $a=0.05$. This approach can be made considerably more computationally efficient if simulation is stopped, and acceptance of the null is recorded once the bootstrap statistic has exceeded the sample statistic more than $a \cdot B$ times.

3 Dimensions of the Size Problem - the Asymptotic IM Test

In this section we present some results based on Monte Carlo simulation of the original White and Chesher-Lancaster IM statistics using asymptotic critical values. It has long been a hallmark of simulation studies of the IM test that rejection rates climb as regressors are added to the model, and fall toward the asymptote as sample sizes are increased. Significantly, the Monte Carlo experiments carried out here suggest that this conclusion is more accidental than real. There is, in general, a considerably richer variety of idiosyncracies in the distribution of the IM test than has heretofore been realized. Before we can delve into the results, however, some obstacles to broadening the scope of the Monte Carlo simulations need to be addressed.

3.a Controlling Predictability

Keeping simulation results stable and comparable turns out to be challenging, particularly as additional regressors are added to the models in smaller samples. This is not surprising, and it can be combatted by exercising sufficient control over the *r*-squared of the underlying data generating process. The typical Monte Carlo simulation for linear regression or probit models involves generating a *Y* (or latent *Y**) sequence using an *X* matrix with *k* independent columns of random data with a fixed variance and zero mean plus an error term. The error is usually a zero-mean, fixed-variance sequence as well. With this kind of framework, simply adding more columns to the *X* matrix increases the *r*-squared underlying the model. This turns out to have important consequences, especially within the probit model, for the feasibility and comparability of the Monte Carlo simulations.

Consider the following model as an illustration of this issue:

$$\underset{n\times1}{y} = \underset{1\times1}{a} + \underset{n\times k}{X} \cdot \underset{k\times1}{b} + \underset{n\times1}{e}\text{, where } e \sim iid(0, s_e)\text{, and } X_i \sim (m_x, s_x),\ i = 1,\dots,k$$

where y is a (potentially latent) independent variable, $E[X_i \cdot X_j] = 0, \forall i \neq j$, and $E[X \cdot e] = 0$. For simplicity, assume the *X*'s are distributed identically, as they would typically be in a simulation experiment. The probability limit of the regression *r*-squared for the least squares regression applied to this model when *x* is observed has the simple form

$$R^2_{pred} = \frac{s_x^2 \sum_{i=1}^{k} b_i^2}{s_x^2 \sum_{i=1}^{k} b_i^2 + s_e^2}. \tag{5}$$

Without loss of generality, setting $b_i = s_e^2 = s_x^2 = 1$, allows simplification of the limiting *r*-squared equation to $R^2_{pred} = {}^{k}\!/_{(k+1)}$.

This is plotted in Figure 1. Starting out at 0.50, and after initially climbing rapidly, it asymptotes to the maximum of one. For the linear regression model, allowing the underlying *r*-squared to climb as regressors are added is mostly an issue of comparability. There are, in principle, no problems with this for information matrix testing because of the way it is standardized. Nevertheless, controlling *r*-

squared is probably worthwhile simply to eliminate the possibility that the signal to noise ratio in the model is affecting results.

The problems for the probit model run a little deeper. To understand the problems that arise in this case, consider the fact that the conditional probability $P[Y=1|X]=\Phi(X'b)$ (where Φ is the standard normal cumulative distribution function), requires a large positive argument to achieve a value close to one, and a large negative argument to generate a zero. This immediately causes convergence problems in the the maximum likelihood estimation process if the predictability of Y is high. This is due to the fact that extremely large parameter estimates would be called for to generate the needed conditional probabilities lying near zero or one. Assuming that satisfactory parameter estimates can be achieved, the entire likelihood function is still highly volatile in these cases. This happens because the gradient of the likelihood equation contains a conditional variance term, $\Phi(X'b)\cdot(1-\Phi(X'b))$, in the denominator, as does the IM statistic itself. This term tends to zero as Φ tends either to zero or one, which causes the gradient function, and IM statistic to explode as conditional probabilities approach their limits.

The simplest solution for this problem is to eliminate iterations which fail the necessary convergence or invertibility conditions. In the small samples where these problems are most apparent, this can become expensive and is fraught with potential biases. To give an example, with seven regressors (including the constant), the uncontrolled r-squared for the underlying model is 0.857. In this case, simulation shows that it can take over 240 attempts to deliver 100 sequences that pass the convergence and invertibility criteria. This means that nearly 60% of the runs are being eliminated, and most of those remaining obviously belong to a select group.

A simple alternative solution to this problem is to control the *r*-squared by inverting equation (5) and using this result to scale the variances in the models appropriately. The resulting relation was used to scale the variances of the simulated *x*'s to keep predictability constant within both the linear regression and probit models:

$$s_x^2 = \frac{R^2 \cdot s_e^2}{(1-R^2)\sum_{i=1}^{k} b_i^2} . \qquad (6)$$

Unless otherwise noted, the *r*-squared was set at 0.5 throughout the experiments.

We can now turn to the models used in the Monte Carlo experiments.

3.b Linear Regression Model

The data generating process (DGP) used in the Monte Carlo experiments, is the following:

$$Y_t = X_t' b + e_t , \qquad (7)$$

where $\underset{k\times 1}{X_t} = \left(1, \tilde{X}_t'\right)'$, $\underset{(k-1)\times 1}{\tilde{X}_t} \sim NID(0, s_x^2)$, s_x^2 is defined by equation (6), $e_t \sim NID(0,1)$, and $\underset{\mathrm{k}\times 1}{b} = (1,\ldots,1)'$. The indicator vector for the full IM test in the linear regression case is given by

$$m_t = \frac{1}{s^2} S \cdot vec\begin{bmatrix} \underset{k\times k}{x_t x_t'(u_t^2 - 1)} & \underset{k\times 1}{x_t(u_t^3 - 3u_t)} \\ \bullet & \underset{1\times 1}{u_t^4 - 5u_t^2 + 2} \end{bmatrix},$$

where $u_t = \frac{1}{s}(y_t - x_t' b)$. The matrix inside the brackets has *(k+2)(k+1)/2* unique elements where *k* is the number of regressors including the constant. This convention will be maintained throughout this paper. *S* is chosen in this case to select all unique elements while leaving out the element in the upper left corner which is identically one. The resulting vector is comprised of second, third, and fourth moments of the sample data. The upper left block is sensitive to conditional heteroskedasticity since it is based on the covariance between various crossproducts of the *X* matrix and the *u*'s. Selecting these indicators alone would result in the White test for heteroskedasticity. The upper right block contains a vector of elements which test for skewness, or skewness conditional on the regressors. Under the null hypothesis (conditionally normal error distribution), the conditional skewness of the distribution is equal to zero, so this vector has an expected value of zero. The bottom diagonal element is a measure of excess kurtosis. Under the null hypothesis, $E[u_t^4] = 3$, and since $E[u_t^2] = 1$, the whole term has expected value zero as well. Note that for each higher moment, the complexity of the relationship between the *X*'s and the standardized errors is reduced.

3.c Probit Model

The probit DGP is specified as follows:

$$\tilde{Y}_t = X_t' b + e_t \tag{8}$$

$$Y_t = \begin{Bmatrix} 1 \text{ if } \ \tilde{\mathrm{Y}}_{\mathrm{t}} > 0 \\ 0 \text{ otherwise} \end{Bmatrix}$$

where $\underset{k\times 1}{X_t} = \left(1, \tilde{X}_t'\right)'$, $\underset{(k-1)\times 1}{\tilde{X}_t} \sim NID(0, s_x^2)$, s_x^2 is defined by equation (6), $e_t \sim NID(0,1)$, and $\underset{\mathrm{k}\times 1}{b} = (1,\ldots,1)'$. The indicator vector for the Probit IM test statistic is given by

$$m_t = S \cdot vec\left\{ \underset{k\times k}{x_t x_t'} \cdot (-x_t' b) \cdot f(x_t' b) \cdot u_t \right\},$$

where $u_t = \dfrac{y_t - \Phi(x_t' b)}{\Phi(x_t' b)(1 - \Phi(x_t' b))}$, and f is the standard normal density function. The matrix inside the brackets has *(k+1)(k)/2-1* unique elements, where *k* is the number of regressors including the constant. *S* is the selector matrix described in the linear regression case. The standardized residual, u_t, appearing in this statistic bears more than a passing resemblance to the standardized residual in the linear regression version of the statistic. The test statistic in this case is sensitive to deviations in either the conditional mean or variance specifications. The dimension of this test is lower than in the linear regression case because the Bernoulli distribution is a single parameter family where the mean and variance are indexed by a single parameter.

For both sets of experiments, the values of the *X* matrix are fixed, so the error term is singled out as the key source of variability within the experiments. This follows the approach of Orme [1990], Chesher and Spady [1991], and Horowitz [1994]. Sample sizes were allowed to range from 50 to 10,000 to map out the small sample as well as the medium/large sample performance of the statistics. The dimension of the *X* matrix ranges from two to seven, which is the largest model for which the linear regression version of the test can be computed in a sample of fifty observations. We employ both the White (equation 2(a)) and Chesher-Lancaster (equation 2(b)) forms of the statistic in the experiments. Rather than relying on the $n \cdot R^2$ form for computation, the statistics here are computed directly based on equation (1). This allows the covariance matrix to be separately estimated, which is important for comparison with our subsequent nonparametric bootstrap experiments. For each draw from the Monte Carlo distribution, we run both versions of the test statistic on the same data in order to avoid errors induced by the sampling procedure.

3.d Monte Carlo Simulation Results

Tables 1 and 2 summarize the results of Monte Carlo experiments using asympototic critical values for the linear and probit regression models. It is immediately apparent that gross over-rejection of the null hypothesis is the general rule in these results. Also, as expected, adding regressors to the equations generally causes rejection rates to climb for a given sample size.

We first turn our attention to the linear regression model where, with a sample size of fifty and just two regressors, the Chesher-Lancaster form of the statistic incorrectly rejects the null hypothesis almost 65% of the time at the nominal 5% level. In the worst case, the rejection rate approaches 98% using this form of the statistic. In general, the Chesher-Lancaster statistic yields higher rejection rates than the White form, but the latter is a close second, with rejection rates topping 92% in the n=100, and k=7 case. The probit model fares little better with a top rejection rate near 98% observed at n=250 and k=7 using the Chesher-Lancaster statistic. In the n=50 and n=100 cases, the White form of the statistic *appears* to have much better size properties than the Chesher-Lancaster form, but for larger sample sizes demonstrates behavior that is more comparable. With either model, boosting the sample size up to n=10,000 brings the rejection rates near the nominal rates only for models with few regressors. Even in these large samples adding regressors to the models still has a very noticeable impact on the rejection rates.

Though adding regressors to the equations does generally cause false rejection rates to climb for a given sample size, this turns out to be only part of the story. As Figures 2 and 3 graphically illustrate, below a certain threshold sample size, things can work in exactly the opposite direction, which is to say that rejection rates can actually fall with increases in the number of regressors. The most dramatic result is a near zero rejection rate when n=50, k=7 for both of the linear regression cases, as well as the White form of the probit model case. Part of the problem here is actually a facet of the "curse of dimensionality" discussed in previous work. Fitting a model with six, or seven regressors using fifty observations is not an extraordinary thing to do, but the full IM test enters into marginal territory because of the large number of indicators appearing in the statistic. Figure 4 illustrates the growth of q (the number of test indicators) for the full IM test as the number of variables is increased. This reveals the first potential difficulty associated with using the full IM test on even moderately complex models in small samples. In the worst case examined, (linear regression with k=7 and n=50*)*, we are estimating 35 IM test elements using only fifty data observations. A linear model with nine regressors would generate a situation with more parameters to estimate than there are data points to estimate them with. It is no great surprise then, that unpleasant things happen in the region where q is close to n.

What may also be surprising is the near complete breakdown of the IM statistic in the cases where we estimate models with seven regressors on a sample of fifty observations. It turns out that this is due to numerical bounds on the statistic that are incompatible with the asymptotic distribution. As discussed earlier, the IM statistic is closely approximated by $n \cdot R^2$ resulting from an artificial regression. Because the R^2 statistic is bounded on the interval [0,1], the resulting IM test statistic is bounded on [0,n]. For large n, this upper bound presents no real constraint on the distribution of the statistic. However, where n is small and k is large, the statistic can be bounded above by a number that is near or even below the asymptotic critical value. This clearly makes it difficult to reject the null hypothesis.

Figure 4 plots the critcal values associated with the asymptotic IM test. Consider the linear model case where k=7, and n=50. The 95% asymptotic critical value in this case is 49.8, yet the statistic is bounded above by 50. This means that in order for the sampling distribution of the IM test to have the proper rejection properties, 5% of the realizations from it would have to fall within 0.02 of its upper bound at 50. This is very unlikely, and explains why the test is breaking down in this case. The asymptotic distribution is distinctly inappropriate for these very small sample cases.

Nevertheless there appears to be another factor abetting under-rejection in both models. This is most evident in the results for the White version of the statistic, though there is some indication that both forms of the statistic suffer in a similar manner. Rejection rates for the White version of the IM test actually fall as attributes are added to the probit model with n=50. With 100 observations, the rejection rates peak at .419 with five regressors, then begin to fall again. Rejection rates for the linear regression IM test (White version) with n=50 also peak at k=5 before falling rapidly. This happens well before numerical bounds should become an issue for these statistics. Associated with this anomaly are rejection rates that actually rise with n before beginning their descent toward asymptotic levels. A close look at the results in Horowitz [1994 p. 406] reveals a hint of similar complexity. Some of those probit model results also exhibit rejection rates that increase slightly with *n*. The results found here suggest that this was not an accident of sampling.

A closer look at selections from the sampling distribution of the IM test statistic helps to illuminate what is going on. Though it appears that the White form of the statistic is uniquely beset with rejection rates that rise with *n*, both forms of the statistic actually reveal very similar distributional behavior. Taking a slice from Table 1, Figures 5 and 6 plot histograms of the full IM statistic in models with seven regressors. The vertical lines in the plots indicate where the asymptotic critical value is located, and for reference, the asymptotic chi-square distribution has also been superimposed. In the upper left panels for the linear model, we see a clear illustration of the distribution bumping up against the numerical boundary at *n*=50. Some observations exceed 50 due to the fact that the form of test statistic employed here is only asymptotically bounded. It appears that the White form of the statistic is more sensitive to this boundary than the Chesher-Lancaster form. The probit model results, on the other hand, appear less affected by the numerical boundary problem in this example, presumably because fewer indicators are involved.

In all cases, as sample size is allowed to grow, the sampling distribution of the IM statistic spreads out and shifts right dramatically. Comparison of the panels for *n*=50 and *n*=100 gives clear indication that, even in the probit model cases, numerical bounds are probably playing an important role in shortening the right tail of the distribution. The spreading and shifting of the sample distribution continues in some cases well beyond where numerical bounds would appear to have any impact. In the probit examples, the distribution actually persists in spreading out to the right in samples well beyond 500 observations. It is the large rightward shift of mass observed in these histograms that is responsible for causing some of the rejection rates to climb with *n*. The spreading of the distribution, though, is causing an off-setting effect that puts more mass in the left tail of the distribution. The balance between spread and shift varies across the different cases leading to different apparent behavior in the rejection rates, but the underlying phenomenon appears to be similar in all of the IM test examples shown here.

To quantify spread and shift, Figures 7 and 8 portray the median and standard deviation of the sampling distribution of the IM test, focusing on cases with less than 1,000 observations. Superimposed on these charts for reference are the 50th and 95th percentiles from the asymptotic chi-square distribution. The White version of the statistic demonstrates universally lower median values than the Chesher-Lancaster version, but this is only pronounced within the probit model. In virtually every case, the 50th percentile of the sample statistic actually exceeds the 95th percentile of the asymptotic distribution at some sample size. Only the models with two or three regressors demonstrate median IM statistics approaching the asymptotic level within this range of sample sizes. Models with more regressors all exhibit medians that increase over some range of sample sizes. This gives clear evidence that a rightward shift of mass with increases in *n* is universal to these experiments given a sufficient number of regressors.

While the median values from the sampling distribution of the IM test generally begin to drift downward with more than 300-400 observations, the standard deviations of these distributions appear to grow more persistently. Figure 8 demonstrates that this is particularly true for the probit model examples where standard deviations appear to be growing even at the top end of the range of sample sizes. The disparity between the two forms of the statistic is also more dramatic in this example. The distribution of the Chesher-Lancaster form of the probit model IM statistic exhibits considerably larger increases in spread than the White form. These increases in spread put enough extra mass in the left tail to counteract an overall shift to the right with the result that rejection rates are relatively unaffected. The White form of the statistic spreads less, so the rightward shift, which is smaller, but also more sustained in larger samples, has an observable upward impact on rejection rates over a wider range of *n*. The differences between these two forms of the IM test appear to be a matter of the degree to which they are affected by factors causing the sampling distribution to move around. The evidence presented here suggests that the underlying factors are nevertheless similar.

The distinction between the CL and White forms of the IM test lies solely in the covariance matrix estimators being employed. These covariance matrices are based on asymptotic formulas that obviously do not account for the full variability of the IM test vector in finite samples. To generate the observed over-rejection, they must be generally too small in magnitude. However, the fact that the overall IM test distribution is both closer to the origin, and more compact in very small samples suggests covariance matrix estimates that may actually tend to be larger (though not as large as they need to be) in these cases. It is possible that the cause of this is inefficient estimation of the covariance matrix, which would cause the estimates to be inflated. As sample sizes increase, the estimates would become more efficient (and therefore compact) which offers an explanation for the increasing medians and spreads. Beyond this point, it is probably the properties of the test statistic vector itself that drive the patterns that we observe. The relative performance of the two forms of the IM test in the probit model also supports this hypothesis. While the Chesher-Lancaster form of the covariance matrix eliminates complex terms that may be harder to estimate with precision in small samples, the tradeoff is that it offers less good overall approximation. This could explain why this form of the test plateaus earlier in the median and spread statistics, but at a higher level than the White form that takes considerably longer to level off.

In either case, the ideal covariance matrix estimator would reflect the true variability of the test vector in all cases of sample size and model dimensionality. But asymptotic formulations should not be expected to ever meet this goal since, by construction, they do not involve any dependence on sample size. As we have seen, there appears to be substantial evidence of this type of dependence in the sampling distribution of the IM statistic. Potential alternative candidate covariance estimators would involve dependence on sample size to account for the patterns of small sample performance observed here. They would also need to take account of the exaggerated increases in the spread of the distribution that occurs with increases in the number of regressors. Finding such an estimator is an interesting challenge for subsequent research. For now we maintain our focus on the ability of bootstrap methods to handle this situation.

4 Bootstrap Monte Carlo Results

Rather than attempt to fix the problematic covariance matrix estimators directly, the bootstrap approach, whether parametric or nonparametric, takes the estimators as they stand, and builds potentially better approximations of the resulting IM test statistic distributions. As we shall see, the estimation problems that presumably underlie the anomalies of the IM test in small samples also appear to have a role in how well the bootstrap performs. The nonparametric bootstrap procedure turns out to offer dramatic size improvements over the asymptotic formulations, but is still not reliable in small samples. Even the parametric bootstrap reveals weak performance in the smallest samples. We now examine the Monte Carlo evidence on the bootstrap IM test in detail.

4.a Size Performance of the Bootstrap IM Test

Our simulations are based on precisely the same simulated sample data and models as the tests involving asymptotic critical values presented in the previous section. For each iteration, a sample is drawn from the null distribution, and a sample statistic is computed. Within each Monte Carlo iteration, the nonparametric bootstrap was used to compute a p-value for the sample statistic. The number of bootstrap iterations was set at B=100, which we found to be a robust, and efficient number of iterations. Levels for B ranging up to 5000 were tested with no appreciable impact on the results. Similarly we report rejection rates for the parametric bootstrap version of the IM test for each Monte Carlo iteration.

Tables 3 and 4 summarize the results of the bootstrap IM tests under the null hypothesis. Overall, the nonparametric bootstrap betters the performance of the asymptotic test procedure considerably by approaching the nominal rejection rate far more quickly. This is one of the key predictions of bootstrap approximation theory. Rejection rates are close to the nominal rate once sample size approaches between 500 and 1,000 observations for the linear model, and between 1,000 and 2,500 observations for the probit model. In contrast, using standard asymptotic critical values, even with 10,000 observations is not enough to guarantee proper size performance in all cases. Another feature of these results is that, unlike the IM test using asymptotic critical values, the rejection rates using the nonparametric bootstrap never deviate far from the nominal rate. Even though over-rejection is common, it is usually slight. The largest deviation from the nominal rejection rate appears in the probit model results where the Chesher-Lancaster form the test rejected the null hypothesis 18.6% of the time. This is well below peak rejection rates for the asymptotic test procedure.

This may not be as positive as it appears, however, since the majority of the size bias for the nonparametric bootstrap is downward. The downward size bias in these small sample bootstrap IM tests also grows as regressors are added. The probit model results appear to be most affected by this. In a sample of fifty observations, the rejection rate for the White form of the IM test starts at 0.103, which then falls to 0.005 with three regressors, and zero for the remainder of the probit models at this sample size. Significant under-rejection remains observable, though slight, in samples of up to 500 observations. The Chesher-Lancaster version of the test appears a little less affected by this problem, but still exhibits significant under rejection in samples of up to 250 observations.

By comparison, the performance of the parametric bootstrap in this base case appears extremely good, with rejection rates hovering near the nominal rate across the board. Only with the Chesher-Lancaster form of the IM test in the probit model case do we observe any significant under-rejection problems. Overall, our simulations confirm the results of Horowitz, but over a wider range of models and samples. The essentially perfect nominal coverage of this procedure stands out because it is not a general result of bootstrap theory for non-pivotal statistics. These results, therefore, probably owe much to the correspondence of the model estimated (the error density function in particular) with the underlying DGP for the Monte Carlo simulations. In this situation, the parametric bootstrap happens to be generating resampled data from almost exactly the same distribution as that which originally generated the data. Deviations from the original DGP are solely due to errors in estimated parameters, which are not generally large. As we will see later, we cannot always guarantee pleasant situations like this, and this does have an impact on the quality of the results.

One of the important features for samples of less than 1,000 observations is that rejection rates for the nonparametric bootstrap IM test exhibit a mirror image of the problems encontered when using asymptotic critical values. Figure 9 illustrates the patterns in rejection rates that are observed in the nonparametric bootstrap IM test as sample sizes are increased. One striking difference between these results and those employing asymptotic critical values is that the the largest humps and deviations from the nominal rate appear for models with the fewest regressors. In fact, models with greater numbers of regressors never appear to exhibit meaningful over-rejection.

This result is linked to the fact that the same phenomena that plague the asymptotic test results are also causing difficulties for the nonparametric bootstrap. When using asymptotic critical values, the problem was mainly over-rejection (other than in a few exceptional cases) caused by a sample statistic that is much too large and variable, compared to the predictions given by asymptotics. By the same token, the under-rejections here are caused by nonparametric bootstrap statistics that are too large and variable compared to the sampling distribution we observe. The bootstrap replicates the tendency of the IM test statistic to be overblown in small samples, especially with a larger number of regressors. The resulting distribution is spread out, and in small samples, shifted relative to the sampling distribution. This results in *under*-rejection when this spread and shifted distribution is used for inference.

Figures 10, 11, 12, and 13 provide Q-Q plots comparing the observed sampling distribution of the IM test with the two alternative bootstrap (approximating) distributions for a selection of the simulated models. In these plots the bottom axis represents the cumulative probability of the sampling distribution, and the left-hand axis represents the cumulative probability of the approximating distribution. A perfect fit between sampling distribution and approximation is represented by the 45 degree line. Though complete correspondence is optimal, it is the fit between the distributions in the upper tail that matters most for hypothesis testing. The data for these charts is based on 1,000 Monte Carlo IM test iterations per simulation with a single draw from each bootstrap distribution per iteration. This yields 1,000 simulated data points for each distribution displayed in these charts. For reference, figures 14, 15, 16, and 17 feature histogram plots of the same data.

In smaller samples, the lack of correspondence between the nonparametric bootstrap distribution and the sampling distribution is obvious. For example, in the upper left panel of Figure 10, (the Chesher-Lancaster form of the IM statistic in the linear model), the nonparametric bootstrap distribution crosses the 45 degree line at the 70th percentile from above and left, and then stays well below this line to the upper right corner. Since it is above the 45 degree line to the left of this point, it has excessive mass in that region of the distribution. Conversely, in the upper ranges of the sampling distribution, the nonparametric bootstrap distribution has accumulated too little mass because of its longer tail. The deviation in the upper right corner turns out to be relatively small, which leads to rejection rates that are not too far from nominal. However, boosting the number of regressors changes the situation radically in this case. Setting k=4 leads to a nonparametric bootstrap distribution whose mass lies well to the right of the sampling distribution. This is clearly seen in the upper middle panel of Figure 14. This is responsible for the zero rate of rejection that is observed in Table 3. Raising k to 6 simply exacerbates this problem, and rejection of the null hypothesis is essentially precluded at this point. It should be noted that the nonparametric bootstrap very frequently generated unstable estimates of the IM statistic in all of the n=50 cases when k exceeded 5. The resulting distributions are very adversely affected by this, despite elimination of the offending cases.

As sample sizes increase, the nonparametric bootstrap distributions tend to adhere more closely to the corresponding sampling distributions, as expected. Convergence appears to happen quickest in the crucial right-tail area, which aids the good size performance of the procedure. In both the linear and probit model cases, the White version of the statistic exhibited a more distorted bootstrap distribution than the Chesher-Lancaster version. Comparing the top panels (n=50) of Figures 10 and 11, or Figures 12 and 13 shows the White form to be somewhat less biased, but as sample sizes increase to 250 or more, this form then clearly exhibits a greater bias. This does not translate into worse performance in the rejection rate statistics because the White form nevertheless has reasonably good fit with the upper tail of the sampling distribution. This is actually a general tendency for both forms of the statistic; even in the n=1,000 case. The histograms show a modal point in the bootstrap distribution that is slightly to the left of the where the sampling distribution puts it, combined with a right tail that fits the shape of the sampling distribution quite well. This combination yields reasonably good inference in this case, but also leaves room for improvement. This would be particularly true if these bootstrap distributions were used for purposes that employed more than just their upper tails. It is of relatively little consequence in our hypothesis testing environment when the nonparametric bootstrap renders a poor approximation of the lower 95% of the sampling distribution. Such deviations would be considerably more critical if we were interested in simulating sample moments with the nonparametric bootstrap.

The excellent size performance of the parametric bootstrap is reflected in the very close correspondence between the bootstrap distribution and the sampling distribution along the 45 degree line virtually everywhere. This is to be expected, since there is such a high level of correspondence between the model that is used to generate the bootstrap data, and the underlying Monte Carlo DGP. There are, nevertheless, a few noticeable deviations in the parametric bootstrap distribution for the probit model IM tests. In the Chesher-Lancaster version of the test, with n=100 for the larger models, the parametric bootstrap distribution has a slight tendency to put too much mass in the right tail. This causes quite noticeable underrejection, which is as low as 0.006 when k=6. Conversely, the parametric bootstrap distribution of the White version of the statistic has a slight tendency to put too much mass in the left tail when n=50. This leads to significant over rejection, but the magnitude of this deviation is relatively small.

Putting the key features of the bootstrap distributions into sharper focus, Figures 18 and 19 summarize their medians, while figures 20 and 21 summarize their standard deviations. The medians of the nonparametric bootstrap follow those of the sampling distribution reasonably well in many cases, but as the model becomes more complex, a larger sample is required for this to occur. The tendency of the bootstrap distribution to be shifted a little to the left of the sample distribution is most evident in the White version of the test applied to the linear model and the Chesher-Lancaster version of the statistic applied to the probit model. In these cases, the bootstrap medians are consistently below the sample medians except in some small samples. This obvious leftward shift in the center of the distribution is capturing the bulges in the Q-Q plots of the bootstrap distribution and does not necessarily affect rejection rates adversely.

The standard deviation plots of the IM test statistics offer a good illustration of the fact that these nonparametric bootstrap distributions can yield seriously biased moment estimates while still delivering reasonable size performance for hypothesis testing. The standard deviation of the nonparametric bootstrap IM statistic is greater than that of the sample statistic in almost all cases, except for those with few regressors and close to 1,000 observations. In some cases, the spread is more than double that of the sample statistic, and in many cases, a very sizeable amount of excess variation is still evident when n=1,000. The plots also display a characteristic hump that reaches its maximum with a few hundred observations before vanishing. This appears to be an exaggeration of a similar tendency in the sample statistic. All of the plots also reveal the tendency of the nonparametric bootstrap distribution to blow up in the n=50 case when there were a relatively large number of regressors.

Despite the problems encountered, these findings demonstrate that the nonparametric bootstrap does indeed offer substantial improvement over the use of asymptotic critical values for the IM test. This technique delivers nominal rejection performance for the IM test far quicker than standard asymptotic procedures. The rate of improvement is of the order that should be expected from the asymptotic refinements that the nonparametric bootstrap offers. The relatively poor performance of this technique in the smallest samples is not generally surprising, given how far the distribution of the sample statistic wanders from its asymptote in these cases. Here, the better asymptotic properties of the nonparametric bootstrap have little applicability.

The clear success of the parametric bootstrap relative to the nonparametric bootstrap in these cases owes considerably to the design of the experiments. By specifying the correct functional error distribution from the start, the range of outcomes for the parametric bootstrap is automatically limited to regions that include the proper null hypothesis. This is not true for the nonparametric bootstrap, which is not limited in its outcomes in this manner, and must, therefore, rely more heavily on asymptotics to achieve proper coverage.

The excessive variation of the nonparametric bootstrap IM statistics also suggests that there might be a robustness problem. Because the nonparametric bootstrap directly incorporates outlier data points into the resulting distribution, they could well be leveraging the results. Moreover, outliers may have particular importance in the case of the IM statistic where estimation of high sample moments can magnify their effects. The long tailed distributions found in the smaller samples could well be due to this type of problem. It is likely that alternative non-parametric or semi-parametric techniques that incorporate some smoothing of the bootstrap distribution would have potential benefits, if this is indeed the case.

4.b Power Against a Heteroskedastic Alternative

Size performance is only half of the story of any test statistic. We wish not only to know that a test will not generate too many or too few false positives, but that it will also properly flag the alternative hypothesis that it is set up to detect. In the search for improvements in size performance, this aspect of IM testing has frequently been overlooked in previous research. Here, by testing for a common alternative model, we explore the ability of the parametric and nonparametric bootstraps to match their size improvements over the asymptotic procedure with improvements in power. In general, we find that greater sample sizes are required to achieve reasonable sensitivity to the heteroskedastic alternatives we specified. This turns out not to be a defect of the procedure, but a limitation of the underlying

IM statistic. Using Monte Carlo critical value estimates we can assess the practical limits to the power of the IM test using any of these procedures in small samples.

In our power experiments, the full IM test is run against a data generating process (DGP) with a heteroskedastic specification. The terms involving conditional skewness and kurtosis in the linear regression version of the statistic are not strictly needed for testing heteroskedasticity, but they are included in the test statistic to maintain its omnibus testing nature. This slightly handicaps the statistic's power against a heteroskedastic alternative. All parameters for these Monte Carlo experiments are identical to those laid out in the previous section, with modifications only to incorporate heteroskedasticity.

The heteroskedastic linear regression data generating process is of the form:

$$Y_t = X_t' b + n_{e,t}^{1/2} \cdot e_t \qquad (9)$$

where $\underset{k\times1}{X_t} = \left(1, \tilde{X}_t'\right)'$, $\tilde{X}_t \sim NID(0, s_x^2)$, s_x^2 is defined by equation (6), $n_{e,t} = (X_t' b)^2$, $e_t \sim NID(0,1)$, and $\underset{k\times1}{b} = (1,\ldots,1)'$. The heteroskedasticity function $n_{e,t}$ gives an equal weight to each of the regressors. It has an expected value of one regardless of the number of regressors because the variance of *X* is determined by equation (6). This formulation retains the constant signal to noise property of the original null-hypothesis models. In the Monte Carlo simulations, we fit linear regression models that ignore the heteroskedasticity in the DGP and are thus misspecified.

The probit DGP is of the form:

$$\tilde{Y}_t = \tilde{X}_t'\boldsymbol{b} + \boldsymbol{n}_{e,t}^{1/2} \cdot \boldsymbol{e}_t \tag{10}$$

$$Y_t = \begin{Bmatrix} 1 \text{ if } \ \tilde{Y}_t > 0 \\ 0 \text{ otherwise} \end{Bmatrix},$$

where $\underset{k\times 1}{X_t} = \left(1, \tilde{X}_t'\right)'$, $\tilde{X}_t \sim NID(0, s_x^2)$, s_x^2 is defined by equation (6), $\boldsymbol{n}_{e,t} = (X_t'\boldsymbol{b})^2$, $\boldsymbol{e}_t \sim NID(0,1)$, and $\underset{\text{k}\times 1}{\boldsymbol{b}} = (1,\ldots,1)'$. This DGP also retains the constant signal-noise property of the null hypothesis model. In the Monte Carlo simulations we fit probit models that ignore the heteroskedasticity in the DGP and are thus misspecified in a manner analogous to the linear models above.

Additional time is required for running the Monte Carlo simulations under the alternative because nearly all *B* bootstrap iterations must be run to establish rejection. This can take up to almost twenty times longer than the experiments run under the null hypothesis. Because of this limitation, *n*=2,500 is the largest sample size explored here. This cutoff incorporates settings at which the nonparametric bootstrap appeared to have near nominal size performance in most cases.

4.c Size Adjusted Power of the Sample Statistic

One way to eliminate the size bias of the IM test is to calibrate critical values using Monte Carlo estimates of the sampling distribution under the null hypothesis. These estimates are obviously not available to a researcher under normal conditions, but they enable some benchmarking of the test's sensitivity. By doing this we can observe the degree to which the sampling distribution of the IM statistic is shifted under alternatives. This tells us, first of all, whether there is any possibility of detecting the alternative with the IM test given proper critical values. It also provides a solid benchmark against which to compare the various testing procedures. Since the Monte Carlo critical values make use of population information not available in the sample, they represent a type of "gold standard". Any procedure whose performance approaches the performance of the test using Monte Carlo critical values is about as good as can be expected.

The Monte Carlo critical value estimates themselves also give us another window on the sampling distribution of the IM statistic. Figures 22 and 23 plot these estimates of the IM test critical values in comparison with the asymptotic critical values. These plots neatly illustrate the inadequacy of the asymptotic critical values for the IM test. They also clearly mirror, in exaggerated form, the pattern of the medians observed in figures 18 and 19. The steep upward kink near the left axis of these charts illustrates the impact of the IM statistic's numerical boundaries on the upper tail of the distribution.

We turn now, in Table 5, to the power of the IM test against our heteroskedastic alternative using Monte Carlo critical values. The main feature of these results is that the deviation in the IM statistic generated by our choice of alternatives is not enough to be detected reliably in small samples. In the linear regression model, a sample of at least 250 observations may be needed to guarantee good power against this particular alternative. The best rate of rejection for any of the cases with 50 observations, for example, is around 30%, and in three cases is below 10%. These rates of rejection improve quickly as sample size increases, reaching essentially unit power with a sample of 500 observations. The Chesher-Lancaster form of the IM statistic appears to be generally more sensitive than the White form in this model. This is probably attributable to the impact of elements that are eliminated from the CL form of the covariance matrix that do not vanish under alternatives.

In the probit model, on the other hand, it is the White form of the IM statistic that appears to have greater sensitivity. This is also most likely explained by the effect of the alternative on the Chesher-Lancaster form of the covariance matrix. In this model larger samples are required than in the linear regression case before reasonable sensitivity to the alternative is observed. Only with 2,500 sample observations is a rejection rate of over 90% observed across all experiments. In some cases, the power of the IM test in the probit model also appears to deteriorate rapidly with increases in the number of regressors. For example, with n=1000 the Chesher-Lancaster form of the statistic rejected the null hypothesis 94% of the time with two regressors, yet it rejected only 30% of the time with seven regressors. The White form of the statistic is considerably less susceptible to this type of degradation.

An important observation here is that there is no form of the full IM test that is particularly powerful for this heteroskedastic alternative in samples with less than 250 observations (closer to 1,000 observations for the probit model). Because we are employing Monte Carlo critical values in these experiments, we should not expect to be able to improve on this performance with any bootstrap-based procedure. This illustrates the limitations of simply revising critical value estimates. While this approach can, under the correct conditions, handle size problems, it does not directly address the underlying sensitivity of the test statistic. The bootstrap procedures will have little power if the sampling distribution of the IM test statistic itself is not changed by a particular alternative.

4.d Power of the Bootstrap procedure

Evidence of this appears in Tables 6 and 7, which compare the power without size adjustment of the asymptotic, nonparametric bootstrap, and parametric bootstrap IM test procedures. As suggested by the results obtained when using Monte Carlo estimates of the IM test critical values, none of these procedures exhibits large power in samples smaller than 250 observations. The unadjusted power of the asymptotic procedure looks impressive, until compared to its size under the null. Tables 8 and 9 directly compare the unadjusted power of these tests with the size performance seen in Tables 1 through 4. Based on this measure, the asymptotic procedure actually generates a fairly large margin of additional rejections under the alternative when the models are less complex. The difficulty is to make use of this fact. The instances where the asymptotic critical values might appear usable are obvious here in Monte Carlo simulations, but are essentially unknown in real applications. This severely limits the usefulness of the asymptotic procedure despite its potential for power in certain circumstances. Only as sample sizes approach 2,500 does the power of the asymptotic IM test procedure begin to appear reliable (given these experimental conditions).

The relatively good size properties of the nonparametric bootstrap IM test, on the other hand, belie troubles with the power of this procedure. This version of the IM test often exhibits rejection rates that are actually lower under the alternative hypothesis than under the null hypothesis. This type of performance occurs in samples with up to 500 observations in the linear regression case, and with up to 1,000 observations in the probit model case. The performance of the nonparametric bootstrap IM test procedure also quickly deteriorates as regressors are added. In the probit model case with n=1,000, and k=2, both White and Chesher-Lancaster versions of the statistic reject the null hypothesis 68% of the time. This rejection rate falls precipitously to 7.6% at k=5, and then continues to drop to 0.9% at k=7. As with the asymptotic procedure, it is only at a sample size of 2,500 observations that the power of the nonparametric bootstrap IM test greatly exceeds its size over most settings of k. Even at this sample size, though, rejection rates fall by over 50% in the probit model examples when k increases from 6 to 7. The linear model bootstrap IM tests appear to be somewhat less dramatically affected by additional regressors, but still only achieve relatively reliable power against this alternative with a sample of 2,500 observations.

The parametric bootstrap IM test offers an example where even excellent size performance is not a guarantee of univerally good power. However, this procedure performs about as well as can be expected given the results obtained using Monte Carlo critical values. For example, in the probit case with n=50, it yields rejection rates of under 15% for the Chesher-Lancaster form of the statistic in all cases. This is in line with the results in Table 5. Also in line with those results, this procedure generally demonstrates good power with rejection rates that climb toward 100%. This happens with between 250 and 500 observations in the linear model cases, and closer to 1,000 observations in the probit model cases. The White form of the IM test has generally more power against the probit model alternative, and appears more sensitive to the addition of regressors in the linear model cases. In this example, it is clear that sampling directly from the parametric error distribution of the null model is a performance enhancing proposition when this is a desirable thing to do.

Despite efforts to center and rescale the nonparametrically bootstrapped IM test statistic to account for the effects of the alternative, it is apparent that the distribution of this statistic still lies far from that of the null hypothesis. This may be related to the problems that hamper the size of the test in small samples, but the power of the test using either form of the covariance matrix is remarkably similar.

This indicates that the estimation benefits of using the Chesher-Lancaster form of the covariance matrix over the White form appear relatively unimportant for the nonparametric bootstrap. The poor power performance observed here is also considerably more prevalent in moderately large samples. This suggests that these problems are not just a matter of estimation precision, but are also due to effects of the alternative hypothesis on the bootstrap distribution. Evidently, this is not controlled by the centering and rescaling that has been done.

These results illustrate the importance of directly testing the power of a statistical procedure. In contrast to its good size perfromance, the nonparametric bootstrap offers only minor improvements over the asymptotic procedure under the alternative. The performance of the nonparametric bootstrap in this situation hints at the difficulties that i.i.d. resampling from a distribution containing the alternative hypothesis can yield regardless of the precautions taken. The experiments involving Monte Carlo critical value estimates also demonstrate that, even in the best of circumstances, the IM test is hard pressed in smaller samples to detect deviations from the null hypothesis. Bootstrap procedures for correcting critical values work well at times, but are nevertheless limited by the sensitivity of the underlying statistic. To improve the power of the test beyond this point, one must turn attention toward refining the sensitivity of the statistic itself.

4.e Spurious Sensitivity to a Non-Normal Alternative

So far, we have focused on situations that are most favorable to the parametric bootstrap. These are cases where there is no ignored misspecification that is not also part of the IM test. The real strength of the nonparametric bootstrap is that it is applicable even in cases where ignored misspecification is acceptable. The parametric bootstrap is not well suited to this environment because it requires full parametric specification of the null hypothesis model - something that may not be needed, or even exist. In this section we will explore a straightforward example of ignored misspecification in the linear regression model that favors the use of the nonparametric bootstrap for IM testing.

The full IM statistic that has been the focus of our experiments contains skewness and kurtosis terms that directly test the normality of the error distribution of the linear regression model when applied in that case. Such normality is not needed for the consistency of the linear regression parameter estimates or for valid inference. In fact, the White test for heteroskedasticity, which is a subset of the full IM test, makes no use of normality assumptions whatsoever. In this environment, one of the potential issues raised by the use of the parametric bootstrap in IM testing is the unnecessary additional structure imposed by using a parametric error density function. To the extent that the structure of this parametric model deviates from the true structure of the underlying DGP, a bias could be induced. In larger samples, as central limit theory would suggest, this problem should eventually vanish, but perhaps not quickly enough to eliminate it as a source of substantial problems in finite samples. Such effects would detract from the seemingly ideal performance of the parametric bootstrap technique for IM testing. The nonparametric bootstrap, on the other hand, imposes no parametric model on the error distribution, and should thus be immune from such issues.

To illustrate these assertions in a conservative example, a *homo*skedastic linear regression DGP with non-normal errors was implemented. The White test, a constrained IM test involving only elements that test heteroskedasticity, is then run on this data. The homoskedastic model should pass the White test despite the errors being non-normal. The DGP is of the following form:

$$Y_t = X_t' b + e_t \,, \tag{11}$$

where $\underset{k\times 1}{X_t} = \left(1, \tilde{X}_t'\right)'$, $\tilde{X}_t \sim NID(0, s_x^2)$, s_x^2 is defined by equation (6), $e_t \sim UID(0,1)$, and b = (1,…,1)'. The indicator vector for the White test employs a selector matrix, *S,* that picks off only the elements of the full IM test that are directly sensitive to heteroskedasticity:

$$\underset{white}{m_t} = \frac{1}{s^2} \cdot vec(\underset{k\times k}{X_t X_t'}(u_t^2 - 1)) \,.$$

The error distribution is uniform with a zero mean and unit variance. The needed half-interval to guarantee unit variance for the uniform is $0.5\sqrt{12}$. The device for controlling the variance of the *X* matrix previously described was also used here so as to keep the signal-noise ratio constant in all experiments. As with the power experiments, the sample sizes were restricted to *n*=2,500 or less to reduce computational loads. The parametric bootstrap uses an incorrect normal error model in place of the true uniform.

Table 10 summarizes the result of these experiments. The heteroskedasticity test using asymptotic critical values exhibits dramatically lower rejection rates than found for the full IM test in the normal linear regression model. The Chesher-Lancaster and White forms of the test exhibit opposite tendencies in the smallest samples. Whereas the Chesher-Lancaster form tends to over-reject significantly, the White form tends to under-reject signifcantly. Also, while the Chesher-Lancaster form tends toward the nominal rate in moderate-sized samples, the White form still exhibits a tendency toward under-rejection even in samples with 2,500 observations. Differences observed here are probably explained by the impact that the non-normality has on the parts of the covariance matrix that have been eliminated (using the null hypothesis) in the Chesher-Lancaster form of the statistic. The fact that these two forms behave very differently serves to underscore the importance of keeping the null-hypothesis assumptions used to construct the covariance estimators in line with the model being tested.

The nonparametric bootstrap results are largely similar to those obtained for the full IM test, though there are no instances here where it significantly over-rejects. There is also a more prolonged tendency to under-reject the null-hypothesis; with slightly sub-nominal rejection rates observable even in samples of 2,500. Overall, there appears to be some sensitivity in these results to the error model being employed in the simulations. Nonetheless because of the limitations of the nonparametric bootstrap distribution in general, it is somewhat difficult to draw firm conclusions from these results.

The parametric bootstrap, on the other hand, performs radically differently in this example. Both forms of this test reveal problems arising from the normality assumption that has been imposed. In this case, it appears to cause gross under-rejection in a large array of cases. As expected, this problem begins to vanish as sample size increases, but the dimensionality of the model still has a noticeable impact on the results in larger samples. Even in a sample with 2,500 observations, the White version of the statistic yields a rejection rate of only 0.2% when k=7. The dramatic under-rejection is attributable to the longer tails generated by the incorrectly applied normal errors in the parametric bootstrap procedure. This would also impact the power of the parametric bootstrap test procedure in this instance.

The results of this section demonstrate that there are few places in hypothesis testing where auxiliary hypotheses are completely benign. It is crucial that the assumptions underlying the testing framework are in agreement with those of the model itself. This is particularly true of specification tests. In a complex test, like the IM test, the impact of various assumptions is not always obvious. The safe choice almost always relies on the smallest number assumptions, which is a primary motivation for using the nonparametric bootstrap on the IM test. As the results for that technique illustrate, however, this is not without its costs in terms of both precision and sample requirements. On the other hand, the parametric bootstrap, which works well when its auxiliary assumptions fit the underlying

DGP, can deviate from nominal performance dramatically when these assumptions do not fit. Though this problem vanishes as samples grow larger, this is not where the bootstrap is of greatest use since the asymptotic critical values also begin to work reasonably well in these regions. The problems for the parametric bootstrap highlighted here are of greatest concern in precisely the small sample instances where the bootstrap is most likely to be used.

5 Summary and Concluding Remarks

The key property of an omnibus specification test is its general applicability. Such a test needs to be appropriate in a wide range of circumstances without modification if it is to have wide acceptance. There are two facets of the applicability problem. One is theoretical, and the other is empirical. On one hand, the test needs to be theoretically appropriate for a wide variety of models while requiring only weak assumptions. On the other hand, the test needs to demonstrate that it performs well in practice under a wide variety of realistic conditions involving limited samples. We have explored some aspects of both facets of the information matrix test in this paper.

First, we have provided considerable Monte Carlo evidence on the finite sample performance of a variety of forms of the IM test. By carefully controlling the signal to noise ratio of the models, we have extended the range of analysis in a way that has revealed new patterns in the behavior of the IM test using asymptotic critical values. We have also explored the behavior of the parametric and nonparametric bootstrap methods for providing valid inference for the IM test. The results of our Monte Carlo experiments suggest that the nonparametric bootstrap has merit, but is not without limitations. The parametric bootstrap demonstrates good size and power performance in reasonably small samples, but also exhibits sensitivity to spurious alternatives. None of the procedures, demonstrates great power in the smallest samples we examined. Based on our results using Monte

Carlo estimates of critical values, this is due to fundamental limitations in the sensitivity of the full IM statistic in these examples.

From a practical standpoint, our results suggest that in smaller samples it is prudent to focus specification testing tightly on a small number of misspecification indicators rather than mechanically performing the full IM test. This deviates from the omnibus testing framework, but may yield far better inference. The ideal form of the full IM test that demonstrates good size performance, is sensitive in all of the proper directions in small samples, and does not add spurious structure to the modeling framework has yet to be found.

There may well be ways to modify the nonparametric bootstrap IM test to enhance its performance in small samples. Attacking the robustness problem directly by trimming the bootstrap distribution would have the desired effect of reducing the spread of the bootstrap distribution without affecting its location. The downside of using a trimmed sample is that the level of trimming arbitrarily determines the spread of the resulting bootstrap distribution. Finding the correct level of trimming required to achieve the proper spread introduces another unknown into the estimation process. Applying smoothers to the nonparametric bootstrap distribution is another approach that could potentially downweight the impact of outliers. Though common smoothers, like the kernel estimator, are often not well estimated in very small samples, they could offer improvements in medium sized samples. Semi-parametric techniques like this could bridge the gap between the excessive structure of the parametric bootstrap, and the minimalist approach of the nonparametric bootstrap.

A continuing theme throughout this paper is that, particularly in very small samples, it is extremely difficult to get proper inference for the IM test no matter which technique is being used. There are no clear winners in terms of small sample performance. This appears to be a fundamental property of the IM test statistic itself. The combination of high sample moments and large numbers of elements appearing in the statistic are the main contributors to this problem. The impact these two factors have on the covariance matrix estimator in particular, and the effect this estimator has on the sampling distribution of the IM statistic is significant. This is perhaps one direction where future efforts should be concentrated. Finding a more stable covariance matrix estimator, as well as one which better reflects the small sample variability of the statistic, would enhance the performance of all aspects of IM testing.

Throughout this paper we have maintained an assumption of independence in the data. No studies to date have explored the complexities that dependence adds to the IM test. Bootstrap techniques for dependent data create opportunities to relax the i.i.d. assumption, and it is a modification that must be explored before the bootstrap IM test will have applicability to the true breadth of economic data. The bootstrap applied in the case of dependence, however, is more complicated because the resampled distributions need to remain faithful to the dependence structure in the data. Several potential nonparametric bootstrap techniques are available for doing this including the moving blocks bootstrap of Kunsch [1989], and the stationary bootstrap of Politis and Romano [1994]. A variety of dependent data parametric bootstrap techniques are also available, but the spurious structure problem is considerably greater in the dependent case than we have seen here, as there is substantial potential for errors in the specification of the dependence structure to contaminate a parametrically bootstrapped IM test statistic.

There are other areas of information matrix testing which have yet been little explored, but which may yield interesting results, and useful statistics for practitioners in a variety of areas. One example is the dynamic information matrix test outlined by White [1987,1994]. This is a test of correct dynamic specification based on testing the first order martingale difference sequence properties of the model. It applies to standard time series models as well as models of second moments such as ARCH and GARCH. To date no simulation studies have been done on the finite sample properties of this test, and it is uncertain whether these tests suffer the same difficulties explored here. Investigating the dynamic information matrix test is a promising direction for further research.

6 References

Bickel, Peter J., and David A. Freedman [1981]: "Some Asymptotic Theory for the Bootstrap," *Annals of Statistics*, 9, 1196-1217.

Chesher, Andrew [1983]: "The Information Matrix Test: Simplified Calculation Via a Score Test Interpretation," *Economics Letters*, 13, 45-48.

Chesher, Andrew and Richard Spady [1991]: "Asymptotic Expansions of the Information Matrix Test Statistic," *Econometrica*, 59, 787-815.

Davidson R, and J.G. MacKinnon [1992]: "A New Form of the Information Matrix Test," *Econometrica*, 60, 145-57.

Efron, Bradley [1979]: "Bootstrap Methods: Another Look at the Jacknife," *Annals of Statistics*, 7, 1-26.

Efron, Bradley, and Gail Gong [1983]: "A Leisurely Look at the Bootstrap, the Jackknife, and Cross-Validation," *The American Statistician*, 37, 36-48.

Efron, Bradley, and R. Tibshirani [1986]: "Booststrap Methods for Standard Errors, Confidence Intervals, and Other Measures of Statistical Accuracy," *Statistical Science*, 1, 54-77.

Efron, Bradley, and Robert J. Tibshirani [1993]: *An Introduction to the Bootstrap*, Chapman and Hall, New York.

Freedman, D.A. [1981]: "Bootstrapping Regression Models," *Annals of Statistics*, 9, 1218-28.

Greene, William [1993]: *Econometric Analysis*, MacMillan, New York.

Hall, Peter [1986a]: "On the Bootstrap and Confidence Intervals," *Annals of Statistics*, 14, 1431-1452.

Hall, Peter [1986b]: "On the Number of Bootstrap Simulations Required to Construct a Confidence Interval," *Annals of Statistics*, 14, 1453-1462.

Hall, Peter [1987]: "On the Bootstrap and Likelihood Based Confidence Regions," *Biometrika*, 74, 481-93.

Hall, Peter [1992]: *The Bootstrap and Edgeworth Expansion*, Springer-Verlag, New York.

Horowitz, Joel L. [1994]: "Bootstrap-Based Critical Values for the Information Matrix Test," *Journal of Econometrics*, 61, 395-411.

Kunsch, H.R. [1989]: "The Jackknife and the Bootstrap for General Stationary Observations," *Annals of Statistics*, 17, 1217-1241.

Lancaster, Tony [1984]: "The Covariance Matrix of the Information Matrix Test," *Econometrics*, 52, 1051-53.

Orme, Chris [1990]: "The Small-Sample Performance of the Information-Matrix Test," *Journal of Econometrics*, 46, 309-311.

Politis, Dmitris and Joseph Romano [1994]: "The Stationary Bootstrap," *Journal of the American Statistical Association*, 89, 1303-1313.

Taylor, Larry [1987]: "The Size Bias of White's Information Matrix Test," *Economics Letters*, 24, 63-67.

White, Halbert [1980]: "A Heteroskedasticity Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity," *Econometrica*, 48, 817-838.

White, Halbert [1982]: "Maximum Likelihood Estimation of Misspecified Models," *Econometrica,* 50, 1-25.

White, Halbert [1987]: "Specification Testing in Dynamic Models," in: T.F. Bewley, ed., *Advances in Econometrics: Fifth World Congress, Vol. I,* Cambridge University Press, New York, 1-58.

White, Halbert [1994]: *Estimation, Inference, and Specification Analysis*, Cambridge University Press, New York.

Wooldridge, Jefferey [1991 a]: "On the Application of Robust Regression-Based Diagnostics to Models of Conditional Means and Conditional Variances," *Journal of Econometrics*, 47, 5-46.

Wooldridge, Jefferey [1991 b]: "Specification Testing and Quasi-Maximum Likelihood Estimation," *Journal of Econometrics*, 48, 29-55.

Table 1

Size of the Asymptotic IM Test - Linear Regression Model

Monte Carlo results with full vector of indicators

Empirical rejection rates under the null hypothesis

		Number of Regressors					
Chesher-Lancaster		**2**	**3**	**4**	**5**	**6**	**7**
n	*50*	*0.648*	*0.816*	*0.876*	*0.948*	*0.888*	*0.043*
	100	*0.516*	*0.683*	*0.841*	*0.917*	*0.959*	*0.978*
	250	*0.334*	*0.481*	*0.618*	*0.791*	*0.887*	*0.955*
	500	*0.229*	*0.349*	*0.434*	*0.625*	*0.734*	*0.838*
	1,000	*0.155*	*0.209*	*0.297*	*0.415*	*0.515*	*0.679*
	2,500	*0.102*	*0.112*	*0.178*	*0.230*	*0.281*	*0.371*
	5,000	*0.076*	*0.101*	*0.148*	*0.138*	*0.187*	*0.247*
	10,000	*0.065*	*0.074*	*0.086*	*0.082*	*0.134*	*0.137*

		Number of Regressors					
White		**2**	**3**	**4**	**5**	**6**	**7**
n	*50*	*0.546*	*0.688*	*0.708*	*0.798*	*0.454*	*0.000*
	100	*0.448*	*0.579*	*0.712*	*0.820*	*0.880*	*0.921*
	250	*0.306*	*0.415*	*0.528*	*0.698*	*0.811*	*0.889*
	500	*0.214*	*0.325*	*0.393*	*0.554*	*0.664*	*0.766*
	1,000	*0.151*	*0.195*	*0.271*	*0.386*	*0.482*	*0.630*
	2,500	*0.099*	*0.109*	*0.172*	*0.217*	*0.271*	*0.345*
	5,000	*0.075*	*0.099*	*0.145*	*0.134*	*0.186*	*0.238*
	10,000	*0.065*	*0.073*	*0.085*	*0.081*	*0.134*	*0.131*

(1,000 Monte Carlo simulations using ***a*** *=0.05 asymptotic critical value. Values outside the range [.038,.066] are significant beyond the 95% level based on individual comparison with the .05 nominal rejection rate.)*

Table 2

Size of the Asymptotic IM Test - Probit Model

Monte Carlo results with full vector of indicators

Empirical rejection rates under the null hypothesis

		Number of Regressors					
Chesher-Lancaster		**2**	**3**	**4**	**5**	**6**	**7**
n	*50*	*0.552*	*0.746*	*0.698*	*0.756*	*0.753*	*0.620*
	100	*0.474*	*0.607*	*0.872*	*0.886*	*0.805*	*0.922*
	250	*0.249*	*0.513*	*0.597*	*0.710*	*0.944*	*0.976*
	500	*0.240*	*0.330*	*0.423*	*0.591*	*0.770*	*0.878*
	1,000	*0.145*	*0.193*	*0.345*	*0.463*	*0.581*	*0.766*
	2,500	*0.091*	*0.151*	*0.206*	*0.275*	*0.337*	*0.465*
	5,000	*0.079*	*0.116*	*0.143*	*0.162*	*0.209*	*0.299*
	10,000	*0.075*	*0.097*	*0.114*	*0.131*	*0.135*	*0.189*

		Number of Regressors					
White		**2**	**3**	**4**	**5**	**6**	**7**
n	*50*	*0.233*	*0.205*	*0.199*	*0.158*	*0.095*	*0.011*
	100	*0.218*	*0.336*	*0.323*	*0.419*	*0.359*	*0.470*
	250	*0.184*	*0.322*	*0.398*	*0.507*	*0.609*	*0.694*
	500	*0.181*	*0.275*	*0.340*	*0.470*	*0.591*	*0.704*
	1,000	*0.130*	*0.161*	*0.290*	*0.366*	*0.451*	*0.609*
	2,500	*0.081*	*0.132*	*0.179*	*0.243*	*0.286*	*0.400*
	5,000	*0.079*	*0.112*	*0.124*	*0.144*	*0.191*	*0.254*
	10,000	*0.071*	*0.095*	*0.110*	*0.124*	*0.126*	*0.169*

*(1,000 Monte Carlo simulations using **a** =0.05 asymptotic critical value. Values outside the range [.038,.066] are significant beyond the 95% level based on individual comparison with the .05 nominal rejection rate.)*

Table 3

Size of the Bootstrap IM Test - Linear Regression Model

Monte Carlo results with full vector of indicators

Empirical rejection rates under the null hypothesis

Non-Parametric Bootstrap

		Number of Regressors					
Chesher-Lancaster		**2**	**3**	**4**	**5**	**6**	**7**
n	*50*	*0.082*	*0.042*	*0.000*	*0.000*	*0.000*	*0.000*
	100	*0.108*	*0.054*	*0.033*	*0.011*	*0.004*	*0.001*
	250	*0.094*	*0.078*	*0.056*	*0.064*	*0.041*	*0.030*
	500	*0.074*	*0.082*	*0.058*	*0.052*	*0.056*	*0.042*
	1,000	*0.059*	*0.065*	*0.061*	*0.078*	*0.065*	*0.061*
	2,500	*0.058*	*0.043*	*0.060*	*0.060*	*0.053*	*0.048*
	5,000	*0.058*	*0.050*	*0.071*	*0.044*	*0.066*	*0.062*
	10,000	*0.044*	*0.046*	*0.055*	*0.045*	*0.058*	*0.046*

		Number of Regressors					
White		**2**	**3**	**4**	**5**	**6**	**7**
n	*50*	*0.150*	*0.115*	*0.026*	*0.001*	*0.000*	*0.000*
	100	*0.142*	*0.103*	*0.075*	*0.039*	*0.024*	*0.001*
	250	*0.115*	*0.095*	*0.086*	*0.092*	*0.073*	*0.052*
	500	*0.081*	*0.092*	*0.072*	*0.077*	*0.079*	*0.062*
	1,000	*0.068*	*0.075*	*0.070*	*0.085*	*0.087*	*0.068*
	2,500	*0.060*	*0.045*	*0.061*	*0.064*	*0.062*	*0.058*
	5,000	*0.059*	*0.052*	*0.073*	*0.044*	*0.069*	*0.066*
	10,000	*0.045*	*0.047*	*0.056*	*0.045*	*0.059*	*0.047*

Parametric Bootstrap

		Number of Regressors					
Chesher-Lancaster		**2**	**3**	**4**	**5**	**6**	**7**
n	*50*	*0.045*	*0.055*	*0.041*	*0.046*	*0.042*	*0.052*
	100	*0.050*	*0.045*	*0.040*	*0.046*	*0.037*	*0.056*
	250	*0.050*	*0.049*	*0.045*	*0.059*	*0.059*	*0.053*
	500	*0.046*	*0.058*	*0.049*	*0.045*	*0.047*	*0.050*
	1,000	*0.048*	*0.040*	*0.050*	*0.067*	*0.054*	*0.045*
	2,500	*0.040*	*0.040*	*0.049*	*0.052*	*0.046*	*0.046*
	5,000	*0.045*	*0.051*	*0.061*	*0.041*	*0.063*	*0.058*
	10,000	*0.040*	*0.052*	*0.057*	*0.047*	*0.064*	*0.047*

		Number of Regressors					
White		**2**	**3**	**4**	**5**	**6**	**7**
n	*50*	*0.042*	*0.057*	*0.036*	*0.052*	*0.050*	*0.055*
	100	*0.053*	*0.045*	*0.037*	*0.052*	*0.043*	*0.051*
	250	*0.050*	*0.048*	*0.049*	*0.064*	*0.052*	*0.054*
	500	*0.045*	*0.059*	*0.050*	*0.046*	*0.050*	*0.050*
	1,000	*0.048*	*0.042*	*0.050*	*0.064*	*0.055*	*0.047*
	2,500	*0.040*	*0.040*	*0.048*	*0.052*	*0.048*	*0.048*
	5,000	*0.045*	*0.050*	*0.061*	*0.041*	*0.063*	*0.059*
	10,000	*0.040*	*0.050*	*0.057*	*0.047*	*0.064*	*0.047*

(1,000 Monte Carlo simulations with 100 bootstrap iterations per simulation using α *=0.05 critical value.)*

Table 4

Size of the Bootstrap IM Test - Probit Model

Monte Carlo results with full vector of indicators

Empirical rejection rates under the null hypothesis

Non-Parametric Bootstrap

		Number of Regressors					
Chesher-Lancaster		**2**	**3**	**4**	**5**	**6**	**7**
n	*50*	*0.043*	*0.001*	*0.000*	*0.000*	*0.000*	*0.000*
	100	*0.186*	*0.092*	*0.006*	*0.001*	*0.000*	*0.000*
	250	*0.112*	*0.141*	*0.087*	*0.035*	*0.033*	*0.018*
	500	*0.123*	*0.106*	*0.081*	*0.073*	*0.048*	*0.058*
	1,000	*0.087*	*0.065*	*0.092*	*0.072*	*0.076*	*0.072*
	2,500	*0.055*	*0.064*	*0.071*	*0.069*	*0.061*	*0.060*
	5,000	*0.051*	*0.066*	*0.058*	*0.046*	*0.052*	*0.062*
	10,000	*0.055*	*0.064*	*0.068*	*0.058*	*0.046*	*0.057*

		Number of Regressors					
White		**2**	**3**	**4**	**5**	**6**	**7**
n	*50*	*0.103*	*0.005*	*0.000*	*0.000*	*0.000*	*0.000*
	100	*0.132*	*0.071*	*0.003*	*0.002*	*0.000*	*0.000*
	250	*0.125*	*0.090*	*0.064*	*0.028*	*0.008*	*0.002*
	500	*0.131*	*0.116*	*0.082*	*0.051*	*0.032*	*0.030*
	1,000	*0.090*	*0.063*	*0.094*	*0.062*	*0.062*	*0.053*
	2,500	*0.060*	*0.067*	*0.076*	*0.069*	*0.061*	*0.060*
	5,000	*0.053*	*0.067*	*0.058*	*0.046*	*0.054*	*0.060*
	10,000	*0.055*	*0.065*	*0.068*	*0.060*	*0.047*	*0.061*

Parametric Bootstrap

		Number of Regressors					
Chesher-Lancaster		**2**	**3**	**4**	**5**	**6**	**7**
n	*50*	*0.029*	*0.028*	*0.018*	*0.023*	*0.040*	*0.039*
	100	*0.048*	*0.045*	*0.028*	*0.016*	*0.006*	*0.008*
	250	*0.049*	*0.051*	*0.033*	*0.027*	*0.032*	*0.025*
	500	*0.039*	*0.059*	*0.035*	*0.047*	*0.031*	*0.038*
	1,000	*0.051*	*0.033*	*0.051*	*0.056*	*0.056*	*0.054*
	2,500	*0.040*	*0.043*	*0.045*	*0.054*	*0.047*	*0.055*
	5,000	*0.053*	*0.057*	*0.048*	*0.036*	*0.044*	*0.048*
	10,000	*0.052*	*0.062*	*0.059*	*0.056*	*0.055*	*0.054*

		Number of Regressors					
White		**2**	**3**	**4**	**5**	**6**	**7**
n	*50*	*0.059*	*0.061*	*0.052*	*0.071*	*0.088*	*0.094*
	100	*0.045*	*0.058*	*0.066*	*0.052*	*0.051*	*0.048*
	250	*0.052*	*0.049*	*0.042*	*0.044*	*0.051*	*0.051*
	500	*0.047*	*0.065*	*0.039*	*0.050*	*0.038*	*0.050*
	1,000	*0.048*	*0.036*	*0.059*	*0.052*	*0.065*	*0.058*
	2,500	*0.040*	*0.042*	*0.043*	*0.056*	*0.047*	*0.058*
	5,000	*0.053*	*0.056*	*0.052*	*0.035*	*0.045*	*0.051*
	10,000	*0.052*	*0.062*	*0.056*	*0.054*	*0.052*	*0.051*

(1,000 Monte Carlo simulations with 100 bootstrap iterations per simulation using α *=0.05 critical value.)*

Table 5

Size Adjusted Power of the IM Test Against Heteroskedasticity

Monte Carlo results with full vector of indicators

Empirical rejection rates under heteroskedastic alternative

Linear Regression

		Number of Regressors					
Chesher-Lancaster		**2**	**3**	**4**	**5**	**6**	**7**
n	*50*	*0.290*	*0.124*	*0.262*	*0.183*	*0.307*	*0.054*
	100	*0.657*	*0.496*	*0.582*	*0.409*	*0.424*	*0.196*
	250	*0.999*	*0.989*	*0.978*	*0.882*	*0.890*	*0.735*
	500	*1.000*	*1.000*	*1.000*	*1.000*	*0.999*	*0.999*
	1,000	*1.000*	*1.000*	*1.000*	*1.000*	*1.000*	*1.000*
	2,500	*1.000*	*1.000*	*1.000*	*1.000*	*1.000*	*1.000*

		Number of Regressors					
White		**2**	**3**	**4**	**5**	**6**	**7**
n	*50*	*0.147*	*0.088*	*0.157*	*0.099*	*0.198*	*0.023*
	100	*0.349*	*0.270*	*0.307*	*0.176*	*0.190*	*0.088*
	250	*0.969*	*0.912*	*0.819*	*0.627*	*0.673*	*0.451*
	500	*1.000*	*1.000*	*1.000*	*1.000*	*0.994*	*0.982*
	1,000	*1.000*	*1.000*	*1.000*	*1.000*	*1.000*	*1.000*
	2,500	*1.000*	*1.000*	*1.000*	*1.000*	*1.000*	*1.000*

Probit

		Number of Regressors					
Chesher-Lancaster		**2**	**3**	**4**	**5**	**6**	**7**
n	*50*	*0.075*	*0.124*	*0.170*	*0.181*	*0.110*	*0.093*
	100	*0.066*	*0.061*	*0.028*	*0.162*	*0.292*	*0.161*
	250	*0.084*	*0.054*	*0.105*	*0.131*	*0.083*	*0.107*
	500	*0.625*	*0.230*	*0.266*	*0.181*	*0.208*	*0.202*
	1,000	*0.935*	*0.941*	*0.866*	*0.527*	*0.473*	*0.298*
	2,500	*1.000*	*1.000*	*1.000*	*1.000*	*0.999*	*0.988*

		Number of Regressors					
White		**2**	**3**	**4**	**5**	**6**	**7**
n	*50*	*0.153*	*0.232*	*0.215*	*0.186*	*0.171*	*0.130*
	100	*0.267*	*0.235*	*0.226*	*0.331*	*0.359*	*0.303*
	250	*0.355*	*0.262*	*0.314*	*0.337*	*0.455*	*0.527*
	500	*0.849*	*0.473*	*0.454*	*0.404*	*0.549*	*0.541*
	1,000	*0.942*	*0.957*	*0.913*	*0.722*	*0.697*	*0.707*
	2,500	*1.000*	*1.000*	*1.000*	*1.000*	*1.000*	*0.993*

(1,000 Monte Carlo simulations using Monte Carlo simulated α =0.05 critical value.)

Table 6

Power of the IM Test Against Heteroskedasticity - Linear Regression Model

Monte Carlo results with full vector of indicators

Empirical rejection rates under heteroskedastic alternative

Asymptotic (unadjusted)

Number of Regressors

Chesher-Lancaster		2	3	4	5	6	7
n	50	0.980	0.981	0.986	0.991	0.972	0.046
	100	0.999	0.996	0.999	1.000	1.000	1.000
	250	1.000	1.000	1.000	1.000	1.000	1.000
	500	1.000	1.000	1.000	1.000	1.000	1.000
	1,000	1.000	1.000	1.000	1.000	1.000	1.000
	2,500	1.000	1.000	1.000	1.000	1.000	1.000

White		2	3	4	5	6	7
n	50	0.850	0.878	0.883	0.880	0.635	0.000
	100	0.981	0.928	0.976	0.982	0.987	0.988
	250	1.000	1.000	1.000	1.000	1.000	0.999
	500	1.000	1.000	1.000	1.000	1.000	1.000
	1,000	1.000	1.000	1.000	1.000	1.000	1.000
	2,500	1.000	1.000	1.000	1.000	1.000	1.000

Non-Parametric Bootstrap

Number of Regressors

Chesher-Lancaster		2	3	4	5	6	7
n	50	0.001	0.000	0.000	0.000	0.000	0.000
	100	0.009	0.000	0.000	0.000	0.000	0.000
	250	0.143	0.031	0.000	0.000	0.000	0.000
	500	0.589	0.355	0.061	0.027	0.005	0.000
	1,000	0.838	0.642	0.349	0.434	0.227	0.115
	2,500	0.971	0.977	0.848	0.859	0.888	0.853

White		2	3	4	5	6	7
n	50	0.003	0.003	0.000	0.000	0.000	0.000
	100	0.030	0.000	0.000	0.000	0.000	0.000
	250	0.210	0.040	0.001	0.001	0.000	0.000
	500	0.748	0.419	0.061	0.022	0.004	0.000
	1,000	0.911	0.765	0.452	0.511	0.213	0.082
	2,500	0.983	0.989	0.917	0.925	0.944	0.902

Parametric Bootstrap

Number of Regressors

Chesher-Lancaster		2	3	4	5	6	7
n	50	0.262	0.141	0.235	0.165	0.255	0.052
	100	0.629	0.458	0.499	0.378	0.362	0.206
	250	0.995	0.979	0.950	0.881	0.879	0.741
	500	1.000	1.000	1.000	1.000	0.999	0.999
	1,000	1.000	1.000	1.000	1.000	1.000	1.000
	2,500	1.000	1.000	1.000	1.000	1.000	1.000

White		2	3	4	5	6	7
n	50	0.139	0.090	0.137	0.102	0.187	0.031
	100	0.346	0.255	0.243	0.173	0.160	0.093
	250	0.943	0.871	0.775	0.639	0.671	0.465
	500	1.000	1.000	0.999	0.997	0.991	0.969
	1,000	1.000	1.000	1.000	1.000	1.000	1.000
	2,500	1.000	1.000	1.000	1.000	1.000	1.000

(1,000 Monte Carlo simulations with 100 bootstrap iterations per simulation using a=0.05 critical value.)

Table 7

Power of the IM Test Against Heteroskedasticity - Probit Model

Monte Carlo results with full vector of indicators

Empirical rejection rates under heteroskedastic alternative

Asymptotic (unadjusted)

		Number of Regressors					
Chesher-Lancaster		**2**	**3**	**4**	**5**	**6**	**7**
n	*50*	*0.501*	*0.685*	*0.810*	*0.850*	*0.833*	*0.632*
	100	*0.633*	*0.718*	*0.785*	*0.913*	*0.934*	*0.962*
	250	*0.880*	*0.832*	*0.858*	*0.886*	*0.976*	*0.993*
	500	*0.996*	*0.961*	*0.905*	*0.956*	*0.987*	*0.990*
	1,000	*0.999*	*0.999*	*0.999*	*0.989*	*0.998*	*0.999*
	2,500	*1.000*	*1.000*	*1.000*	*1.000*	*1.000*	*1.000*
White		**2**	**3**	**4**	**5**	**6**	**7**
n	*50*	*0.340*	*0.435*	*0.415*	*0.375*	*0.261*	*0.042*
	100	*0.555*	*0.609*	*0.603*	*0.730*	*0.713*	*0.735*
	250	*0.855*	*0.808*	*0.799*	*0.820*	*0.927*	*0.967*
	500	*0.996*	*0.950*	*0.890*	*0.941*	*0.977*	*0.980*
	1,000	*0.994*	*0.998*	*0.999*	*0.987*	*0.998*	*0.997*
	2,500	*1.000*	*1.000*	*1.000*	*1.000*	*1.000*	*1.000*

Non-Parametric Bootstrap

		Number of Regressors					
Chesher-Lancaster		**2**	**3**	**4**	**5**	**6**	**7**
n	*50*	*0.008*	*0.001*	*0.000*	*0.000*	*0.000*	*0.000*
	100	*0.020*	*0.006*	*0.000*	*0.000*	*0.000*	*0.000*
	250	*0.131*	*0.003*	*0.003*	*0.002*	*0.000*	*0.000*
	500	*0.517*	*0.154*	*0.030*	*0.009*	*0.002*	*0.000*
	1,000	*0.684*	*0.512*	*0.552*	*0.051*	*0.024*	*0.006*
	2,500	*0.999*	*0.856*	*0.960*	*0.955*	*0.954*	*0.410*
White		**2**	**3**	**4**	**5**	**6**	**7**
n	*50*	*0.021*	*0.004*	*0.001*	*0.000*	*0.000*	*0.000*
	100	*0.038*	*0.010*	*0.000*	*0.000*	*0.000*	*0.000*
	250	*0.187*	*0.005*	*0.004*	*0.005*	*0.000*	*0.000*
	500	*0.573*	*0.189*	*0.051*	*0.015*	*0.005*	*0.001*
	1,000	*0.678*	*0.523*	*0.600*	*0.076*	*0.040*	*0.009*
	2,500	*0.999*	*0.844*	*0.954*	*0.954*	*0.953*	*0.439*

Parametric Bootstrap

		Number of Regressors					
Chesher-Lancaster		**2**	**3**	**4**	**5**	**6**	**7**
n	*50*	*0.068*	*0.124*	*0.112*	*0.122*	*0.077*	*0.080*
	100	*0.094*	*0.101*	*0.055*	*0.116*	*0.127*	*0.112*
	250	*0.186*	*0.098*	*0.132*	*0.148*	*0.202*	*0.218*
	500	*0.618*	*0.421*	*0.292*	*0.269*	*0.280*	*0.336*
	1,000	*0.913*	*0.861*	*0.898*	*0.587*	*0.599*	*0.518*
	2,500	*1.000*	*1.000*	*1.000*	*1.000*	*1.000*	*0.988*
White		**2**	**3**	**4**	**5**	**6**	**7**
n	*50*	*0.164*	*0.258*	*0.219*	*0.233*	*0.204*	*0.179*
	100	*0.226*	*0.250*	*0.212*	*0.316*	*0.346*	*0.286*
	250	*0.354*	*0.246*	*0.279*	*0.311*	*0.437*	*0.490*
	500	*0.827*	*0.561*	*0.415*	*0.403*	*0.491*	*0.545*
	1,000	*0.912*	*0.894*	*0.934*	*0.723*	*0.751*	*0.739*
	2,500	*1.000*	*1.000*	*1.000*	*1.000*	*1.000*	*0.992*

(1,000 Monte Carlo simulations with 100 bootstrap iterations per simulation using α =0.05 critical value.)

Table 8

Relative Power of the IM Test - Linear Regression Model

Monte Carlo results with full vector of indicators

Difference between empirical rejection rates under heteroskedastic alternative and null hypotheses

Asymptotic (unadjusted) — **Number of Regressors**

Chesher-Lancaster		**2**	**3**	**4**	**5**	**6**	**7**
n	*50*	*0.332*	*0.165*	*0.110*	*0.043*	*0.084*	*0.003*
	100	*0.483*	*0.313*	*0.158*	*0.083*	*0.041*	*0.022*
	250	*0.666*	*0.519*	*0.382*	*0.209*	*0.113*	*0.045*
	500	*0.771*	*0.651*	*0.566*	*0.375*	*0.266*	*0.162*
	1,000	*0.845*	*0.791*	*0.703*	*0.585*	*0.485*	*0.321*
	2,500	*0.898*	*0.888*	*0.822*	*0.770*	*0.719*	*0.629*

White		**2**	**3**	**4**	**5**	**6**	**7**
n	*50*	*0.304*	*0.190*	*0.175*	*0.082*	*0.181*	*0.000*
	100	*0.533*	*0.349*	*0.264*	*0.162*	*0.107*	*0.067*
	250	*0.694*	*0.585*	*0.472*	*0.302*	*0.189*	*0.110*
	500	*0.786*	*0.675*	*0.607*	*0.446*	*0.336*	*0.234*
	1,000	*0.849*	*0.805*	*0.729*	*0.614*	*0.518*	*0.370*
	2,500	*0.901*	*0.891*	*0.828*	*0.783*	*0.729*	*0.655*

Non-Parametric Bootstrap — **Number of Regressors**

Chesher-Lancaster		**2**	**3**	**4**	**5**	**6**	**7**
n	*50*	*-0.081*	*-0.042*	*0.000*	*0.000*	*0.000*	*0.000*
	100	*-0.099*	*-0.054*	*-0.033*	*-0.011*	*-0.004*	*-0.001*
	250	*0.049*	*-0.047*	*-0.056*	*-0.064*	*-0.041*	*-0.030*
	500	*0.515*	*0.273*	*0.003*	*-0.025*	*-0.051*	*-0.042*
	1,000	*0.779*	*0.577*	*0.288*	*0.356*	*0.162*	*0.054*
	2,500	*0.913*	*0.934*	*0.788*	*0.799*	*0.835*	*0.805*

White		**2**	**3**	**4**	**5**	**6**	**7**
n	*50*	*-0.147*	*-0.112*	*-0.026*	*-0.001*	*0.000*	*0.000*
	100	*-0.112*	*-0.103*	*-0.075*	*-0.039*	*-0.024*	*-0.001*
	250	*0.095*	*-0.055*	*-0.085*	*-0.091*	*-0.073*	*-0.052*
	500	*0.667*	*0.327*	*-0.011*	*-0.055*	*-0.075*	*-0.062*
	1,000	*0.843*	*0.690*	*0.382*	*0.426*	*0.126*	*0.014*
	2,500	*0.923*	*0.944*	*0.856*	*0.861*	*0.882*	*0.844*

Parametric Bootstrap — **Number of Regressors**

Chesher-Lancaster		**2**	**3**	**4**	**5**	**6**	**7**
n	*50*	*0.217*	*0.086*	*0.194*	*0.119*	*0.213*	*0.000*
	100	*0.579*	*0.413*	*0.459*	*0.332*	*0.325*	*0.150*
	250	*0.945*	*0.930*	*0.905*	*0.822*	*0.820*	*0.688*
	500	*0.954*	*0.942*	*0.951*	*0.955*	*0.952*	*0.949*
	1,000	*0.952*	*0.960*	*0.950*	*0.933*	*0.946*	*0.955*
	2,500	*0.960*	*0.960*	*0.951*	*0.948*	*0.954*	*0.954*

White		**2**	**3**	**4**	**5**	**6**	**7**
n	*50*	*0.097*	*0.033*	*0.101*	*0.050*	*0.137*	*-0.024*
	100	*0.293*	*0.210*	*0.206*	*0.121*	*0.117*	*0.042*
	250	*0.893*	*0.823*	*0.726*	*0.575*	*0.619*	*0.411*
	500	*0.955*	*0.941*	*0.949*	*0.951*	*0.941*	*0.919*
	1,000	*0.952*	*0.958*	*0.950*	*0.936*	*0.945*	*0.953*
	2,500	*0.960*	*0.960*	*0.952*	*0.948*	*0.952*	*0.952*

(1,000 Monte Carlo simulations with 100 bootstrap iterations per simulation using α *=0.05 critical value.)*

Table 9

Relative Power of the IM Test - Probit Model

Monte Carlo results with full vector of indicators

Difference between empirical rejection rates under heteroskedastic alternative and null hypotheses

Asymptotic (unadjusted)

		Number of Regressors					
Chesher-Lancaster		**2**	**3**	**4**	**5**	**6**	**7**
n	*50*	*-0.051*	*-0.061*	*0.112*	*0.094*	*0.080*	*0.012*
	100	*0.159*	*0.111*	*-0.087*	*0.027*	*0.129*	*0.040*
	250	*0.631*	*0.319*	*0.261*	*0.176*	*0.032*	*0.017*
	500	*0.756*	*0.631*	*0.482*	*0.365*	*0.217*	*0.112*
	1,000	*0.854*	*0.806*	*0.654*	*0.526*	*0.417*	*0.233*
	2,500	*0.909*	*0.849*	*0.794*	*0.725*	*0.663*	*0.535*
White		**2**	**3**	**4**	**5**	**6**	**7**
n	*50*	*0.107*	*0.230*	*0.216*	*0.217*	*0.166*	*0.031*
	100	*0.337*	*0.273*	*0.280*	*0.311*	*0.354*	*0.265*
	250	*0.671*	*0.486*	*0.401*	*0.313*	*0.318*	*0.273*
	500	*0.815*	*0.675*	*0.550*	*0.471*	*0.386*	*0.276*
	1,000	*0.864*	*0.837*	*0.709*	*0.621*	*0.547*	*0.388*
	2,500	*0.919*	*0.868*	*0.821*	*0.757*	*0.714*	*0.600*

Non-Parametric Bootstrap

		Number of Regressors					
Chesher-Lancaster		**2**	**3**	**4**	**5**	**6**	**7**
n	*50*	*-0.035*	*0.000*	*0.000*	*0.000*	*0.000*	*0.000*
	100	*-0.166*	*-0.086*	*-0.006*	*-0.001*	*0.000*	*0.000*
	250	*0.019*	*-0.138*	*-0.084*	*-0.033*	*-0.033*	*-0.018*
	500	*0.394*	*0.048*	*-0.051*	*-0.064*	*-0.046*	*-0.058*
	1,000	*0.597*	*0.447*	*0.460*	*-0.021*	*-0.052*	*-0.066*
	2,500	*0.944*	*0.792*	*0.889*	*0.886*	*0.893*	*0.350*
White		**2**	**3**	**4**	**5**	**6**	**7**
n	*50*	*-0.082*	*-0.001*	*0.001*	*0.000*	*0.000*	*0.000*
	100	*-0.094*	*-0.061*	*-0.003*	*-0.002*	*0.000*	*0.000*
	250	*0.062*	*-0.085*	*-0.060*	*-0.023*	*-0.008*	*-0.002*
	500	*0.442*	*0.073*	*-0.031*	*-0.036*	*-0.027*	*-0.029*
	1,000	*0.588*	*0.460*	*0.506*	*0.014*	*-0.022*	*-0.044*
	2,500	*0.939*	*0.777*	*0.878*	*0.885*	*0.892*	*0.379*

Parametric Bootstrap

		Number of Regressors					
Chesher-Lancaster		**2**	**3**	**4**	**5**	**6**	**7**
n	*50*	*0.039*	*0.096*	*0.094*	*0.099*	*0.037*	*0.041*
	100	*0.046*	*0.056*	*0.027*	*0.100*	*0.121*	*0.104*
	250	*0.137*	*0.047*	*0.099*	*0.121*	*0.170*	*0.193*
	500	*0.579*	*0.362*	*0.257*	*0.222*	*0.249*	*0.298*
	1,000	*0.862*	*0.828*	*0.847*	*0.531*	*0.543*	*0.464*
	2,500	*0.960*	*0.957*	*0.955*	*0.946*	*0.953*	*0.933*
White		**2**	**3**	**4**	**5**	**6**	**7**
n	*50*	*0.105*	*0.197*	*0.167*	*0.162*	*0.116*	*0.085*
	100	*0.181*	*0.192*	*0.146*	*0.264*	*0.295*	*0.238*
	250	*0.302*	*0.197*	*0.237*	*0.267*	*0.386*	*0.439*
	500	*0.780*	*0.496*	*0.376*	*0.353*	*0.453*	*0.495*
	1,000	*0.864*	*0.858*	*0.875*	*0.671*	*0.686*	*0.681*
	2,500	*0.960*	*0.958*	*0.957*	*0.944*	*0.953*	*0.934*

(1,000 Monte Carlo simulations with 100 bootstrap iterations per simulation using α =0.05 critical value.)

Table 10

Size of the White Test Under Non-Normality - Linear Regression Model

Monte Carlo results with partial vector of indicators

Empirical rejection rates under non-normal alternative

Asymptotic

		Number of Regressors					
Chesher-Lancaster		**2**	**3**	**4**	**5**	**6**	**7**
n	*50*	*0.107*	*0.124*	*0.127*	*0.157*	*0.111*	*0.089*
	100	*0.091*	*0.098*	*0.087*	*0.144*	*0.139*	*0.128*
	250	*0.061*	*0.071*	*0.076*	*0.089*	*0.082*	*0.113*
	500	*0.057*	*0.051*	*0.064*	*0.059*	*0.089*	*0.089*
	1,000	*0.058*	*0.051*	*0.056*	*0.077*	*0.067*	*0.075*
	2,500	*0.063*	*0.063*	*0.062*	*0.057*	*0.051*	*0.052*
White		**2**	**3**	**4**	**5**	**6**	**7**
n	*50*	*0.024*	*0.002*	*0.000*	*0.000*	*0.000*	*0.000*
	100	*0.064*	*0.016*	*0.003*	*0.001*	*0.000*	*0.000*
	250	*0.046*	*0.044*	*0.032*	*0.021*	*0.007*	*0.003*
	500	*0.049*	*0.043*	*0.036*	*0.025*	*0.025*	*0.023*
	1,000	*0.054*	*0.046*	*0.045*	*0.057*	*0.033*	*0.033*
	2,500	*0.063*	*0.061*	*0.055*	*0.048*	*0.040*	*0.035*

Non-Parametric Bootstrap

		Number of Regressors					
Chesher-Lancaster		**2**	**3**	**4**	**5**	**6**	**7**
n	*50*	*0.010*	*0.001*	*0.000*	*0.000*	*0.000*	*0.000*
	100	*0.041*	*0.007*	*0.001*	*0.000*	*0.000*	*0.000*
	250	*0.044*	*0.026*	*0.016*	*0.012*	*0.002*	*0.000*
	500	*0.047*	*0.035*	*0.037*	*0.020*	*0.017*	*0.014*
	1,000	*0.049*	*0.045*	*0.037*	*0.046*	*0.028*	*0.024*
	2,500	*0.060*	*0.059*	*0.057*	*0.046*	*0.041*	*0.032*
White		**2**	**3**	**4**	**5**	**6**	**7**
n	*50*	*0.028*	*0.002*	*0.000*	*0.000*	*0.000*	*0.000*
	100	*0.065*	*0.011*	*0.002*	*0.000*	*0.000*	*0.000*
	250	*0.053*	*0.033*	*0.025*	*0.017*	*0.003*	*0.000*
	500	*0.054*	*0.046*	*0.040*	*0.024*	*0.021*	*0.016*
	1,000	*0.049*	*0.046*	*0.040*	*0.048*	*0.035*	*0.024*
	2,500	*0.061*	*0.065*	*0.060*	*0.053*	*0.045*	*0.042*

Parametric Bootstrap

		Number of Regressors					
Chesher-Lancaster		**2**	**3**	**4**	**5**	**6**	**7**
n	*50*	*0.019*	*0.013*	*0.014*	*0.021*	*0.032*	*0.043*
	100	*0.017*	*0.012*	*0.006*	*0.011*	*0.009*	*0.004*
	250	*0.020*	*0.015*	*0.012*	*0.009*	*0.003*	*0.002*
	500	*0.028*	*0.012*	*0.011*	*0.006*	*0.002*	*0.002*
	1,000	*0.037*	*0.017*	*0.013*	*0.017*	*0.006*	*0.003*
	2,500	*0.046*	*0.048*	*0.033*	*0.012*	*0.014*	*0.005*
White		**2**	**3**	**4**	**5**	**6**	**7**
n	*50*	*0.013*	*0.006*	*0.014*	*0.015*	*0.017*	*0.033*
	100	*0.016*	*0.002*	*0.002*	*0.003*	*0.003*	*0.001*
	250	*0.017*	*0.009*	*0.011*	*0.001*	*0.001*	*0.001*
	500	*0.028*	*0.009*	*0.007*	*0.003*	*0.000*	*0.000*
	1,000	*0.035*	*0.016*	*0.010*	*0.016*	*0.002*	*0.003*
	2,500	*0.046*	*0.046*	*0.032*	*0.010*	*0.014*	*0.002*

(1,000 Monte Carlo simulations with 100 bootstrap iterations per simulation using ***a*** *=0.05 critical value.)*

Figure 1

R-squared as a function of *k*

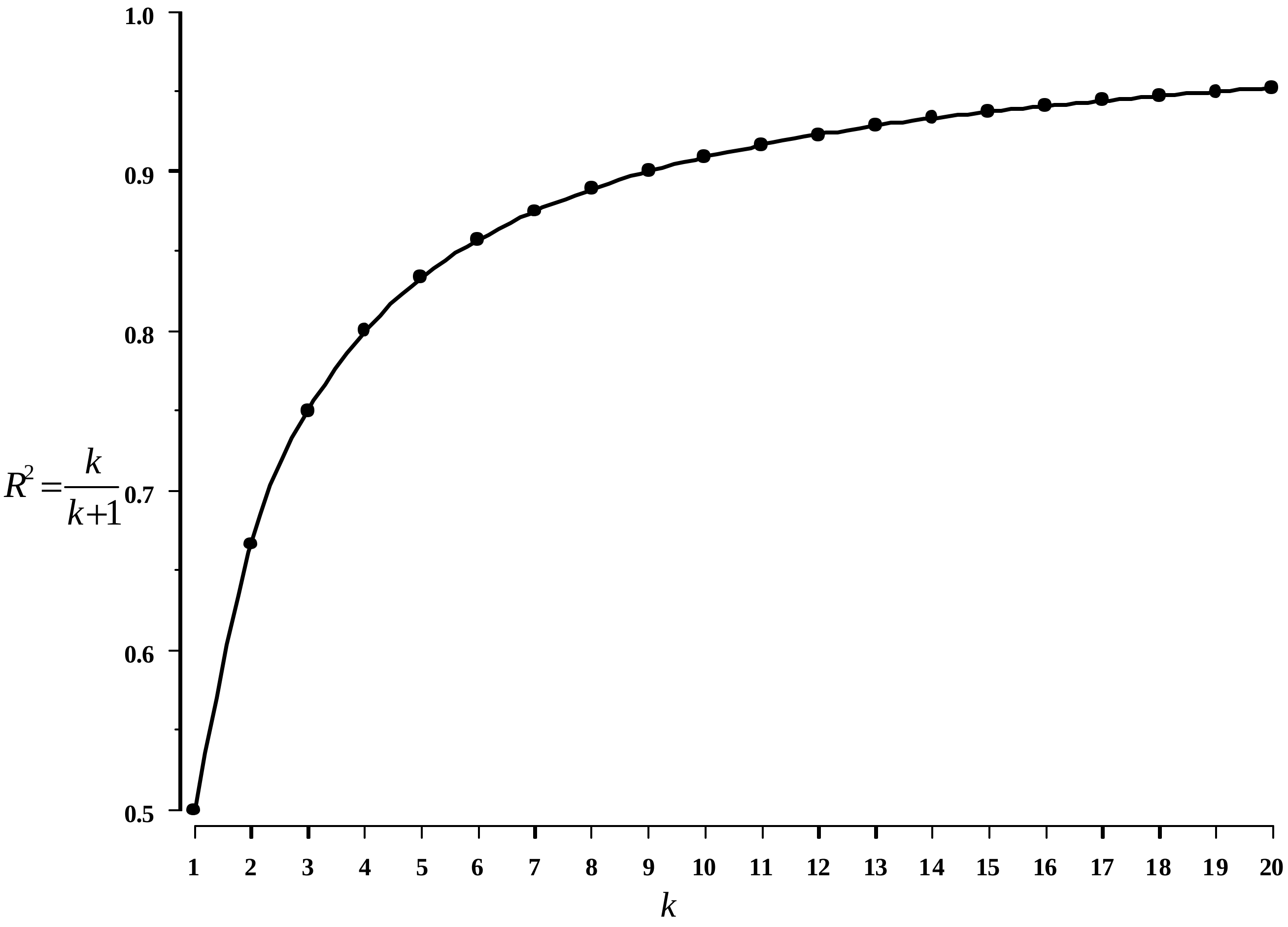
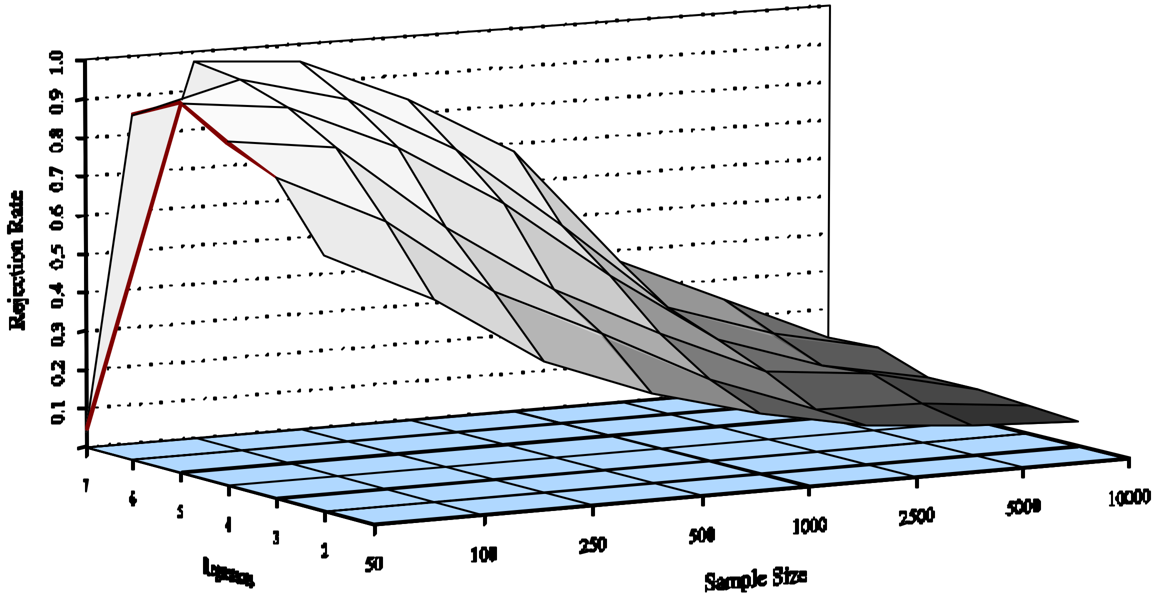


Figure 2

Surface Plots of Monte Carlo Rejection Rates for Full IM Test
Linear Regression Model

Chesher-Lancaster Version



White Version

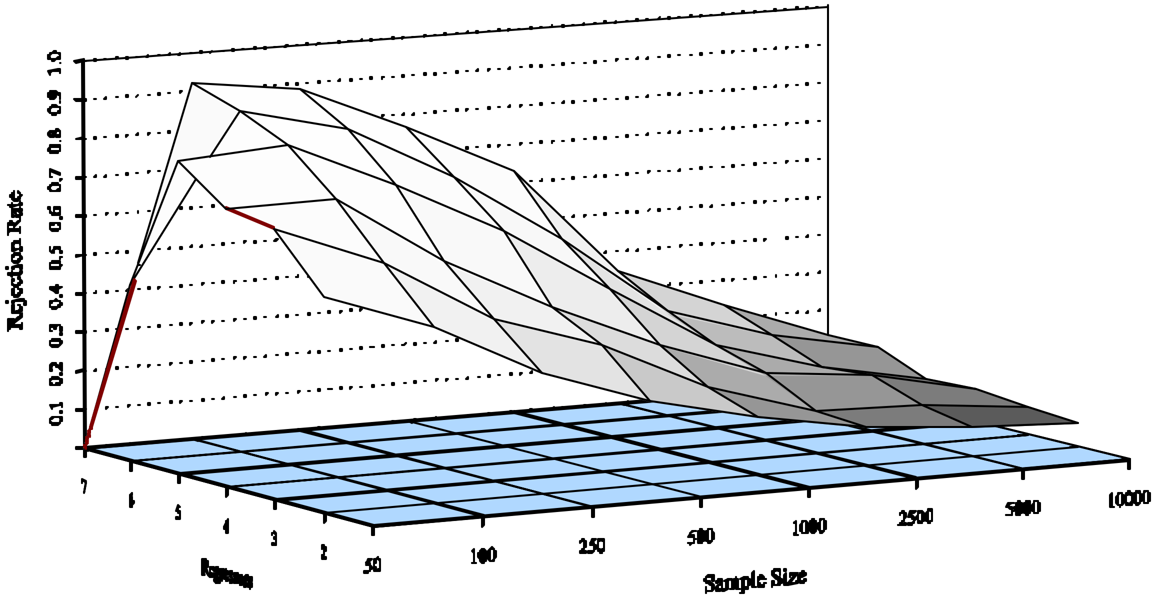
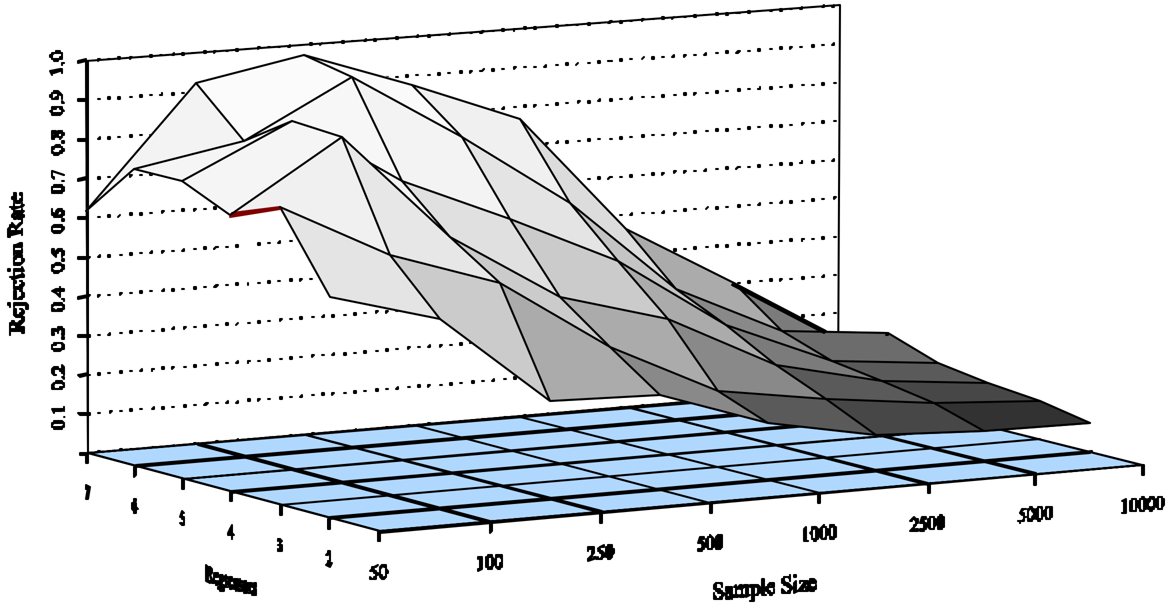


Figure 3

Surface Plots of Monte Carlo Rejection Rates Full IM Test Probit Model

Chesher-Lancaster Version



White Version

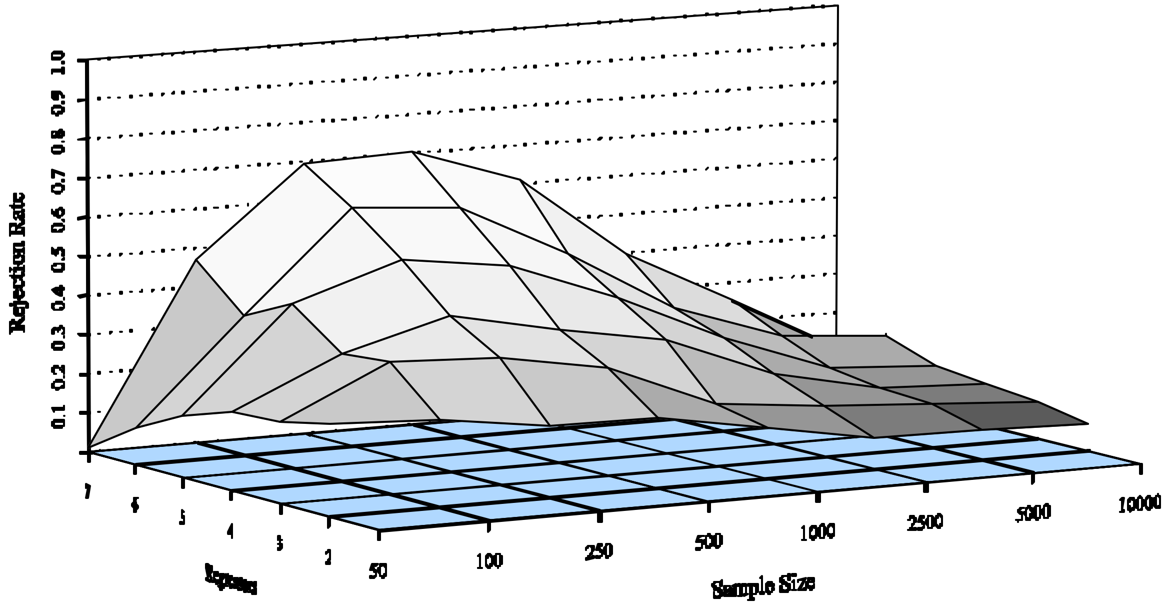


Figure 4

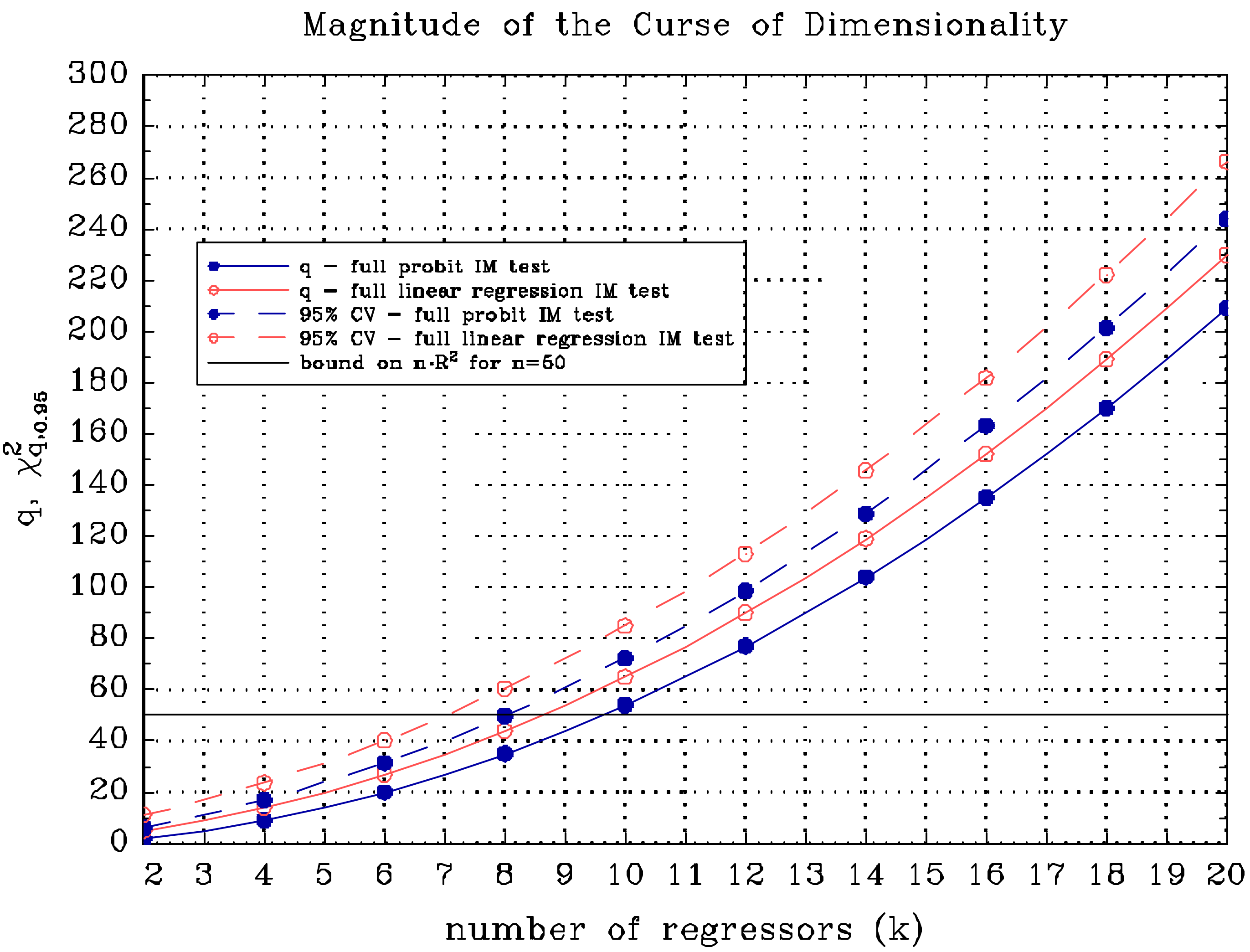
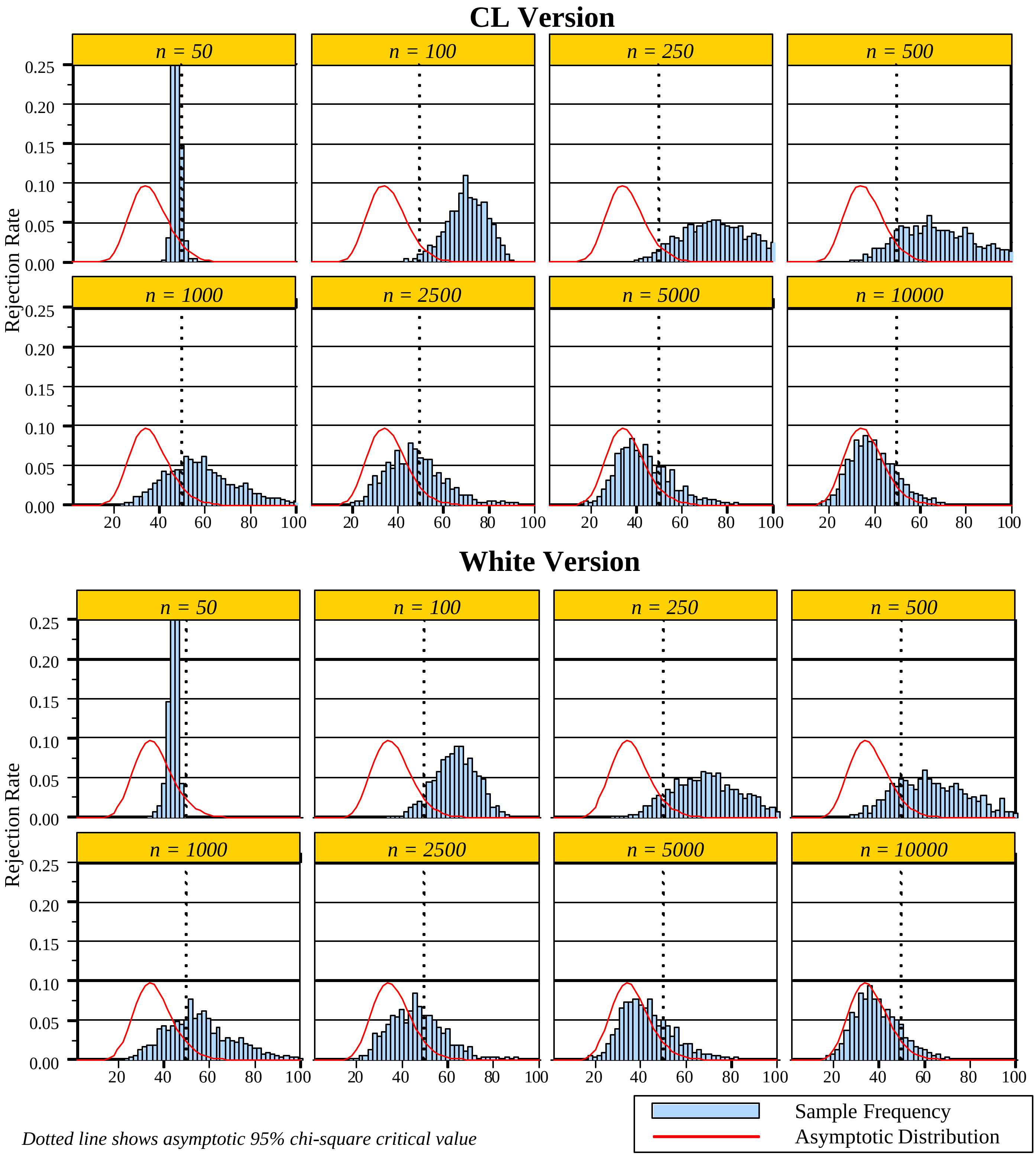

Magnitude of the Curse of Dimensionality
q – full probit IM test
q – full linear regression IM test
95% CV – full probit IM test
95% CV – full linear regression IM test
bound on $n \cdot R^2$ for n=50
$q,\ \chi^2_{q,0.95}$
number of regressors (k)
0
20
40
60
80
100
120
140
160
180
200
220
240
260
280
300
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20

Figure 5

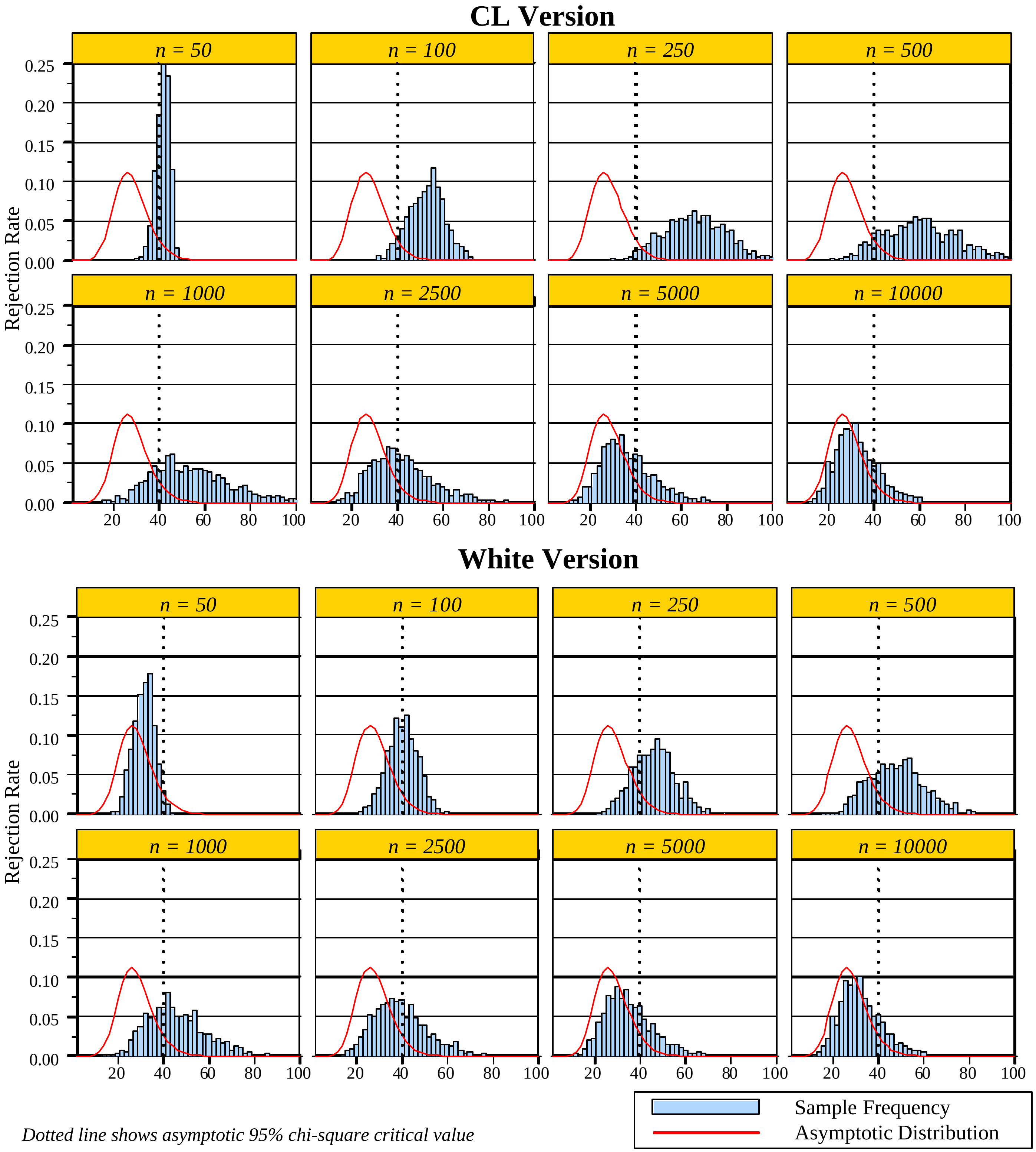
Full IM Test - Linear Regression With k=7



Dotted line shows asymptotic 95% chi-square critical value

Figure 6

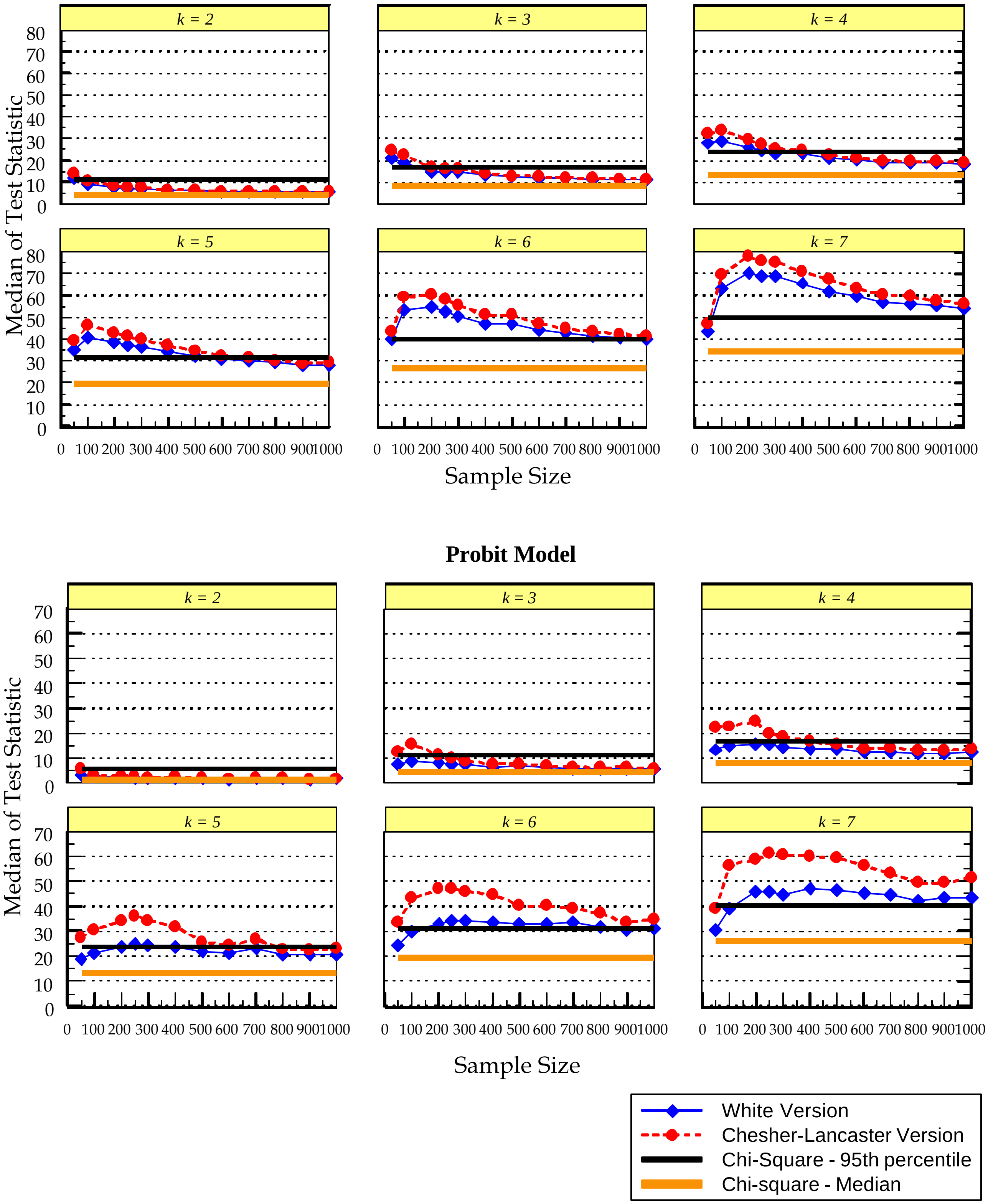
Full IM Test - Probit Model With k=7



Dotted line shows asymptotic 95% chi-square critical value

Figure 7

Median From Monte Carlo Distribution of the Full IM Test Statistic

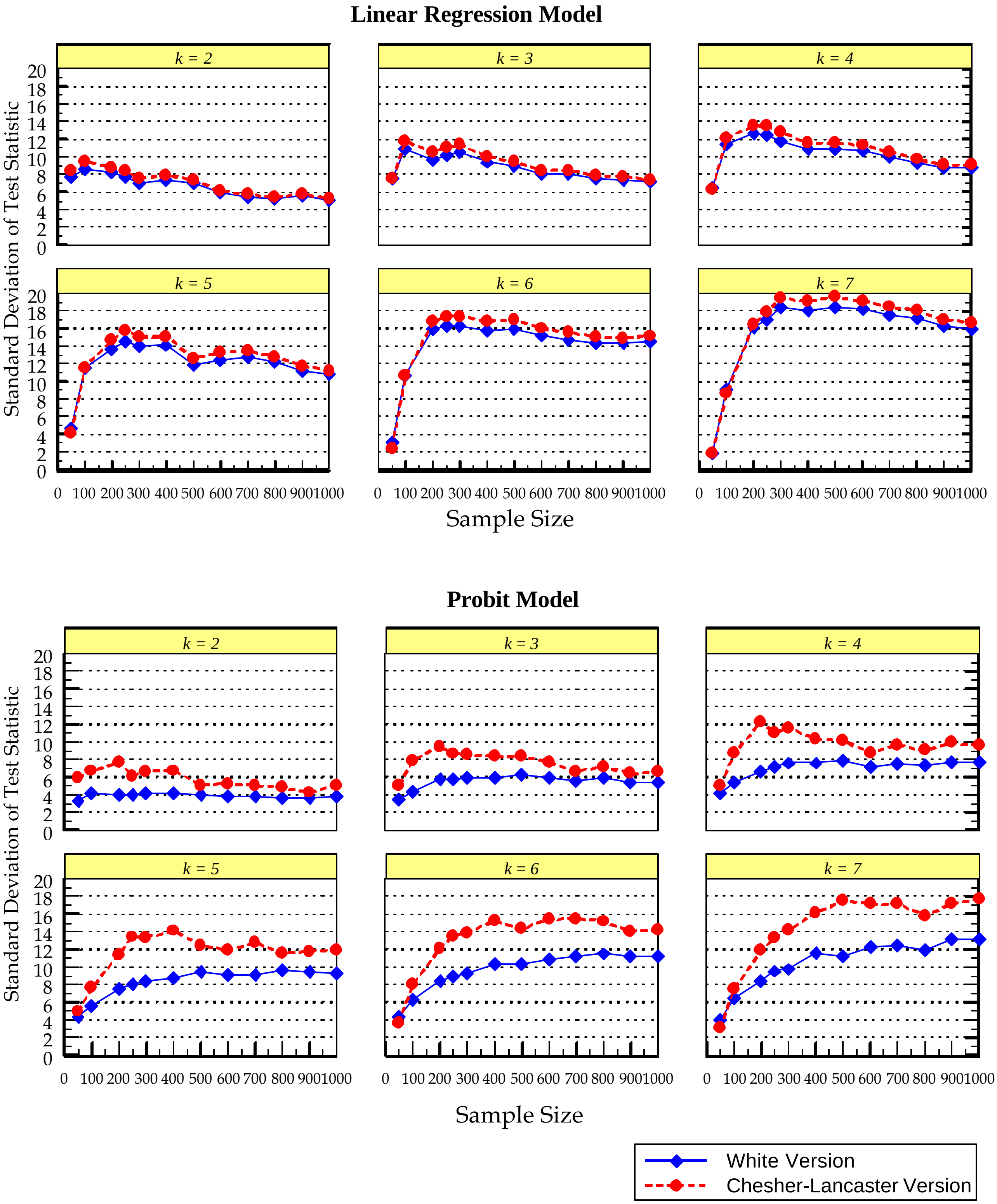


Linear Regression Model

Probit Model

Figure 8

Standard Deviation From Monte Carlo Distribution of the Full IM Test Statistic



Linear Regression Model

Probit Model

Figure 9

Full IM Test - Non-Parametric Bootstrap Rejection Rates Under Null Hypothesis

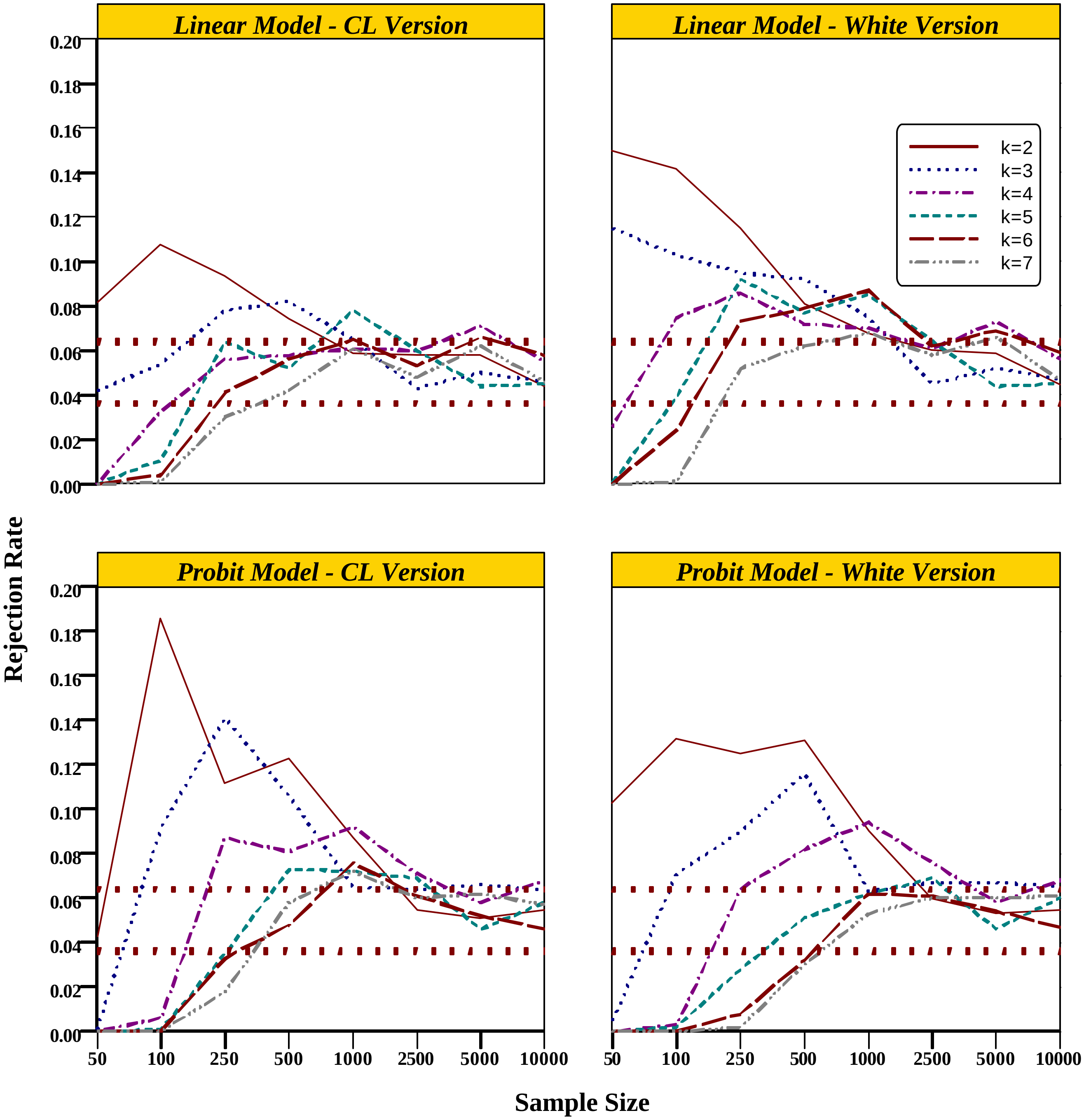


Figure 10

QQ Plot Comparing Bootstrap Distributions with Sampling Distribution

Full Chesher-Lancaster Version of Linear Regression IM Test

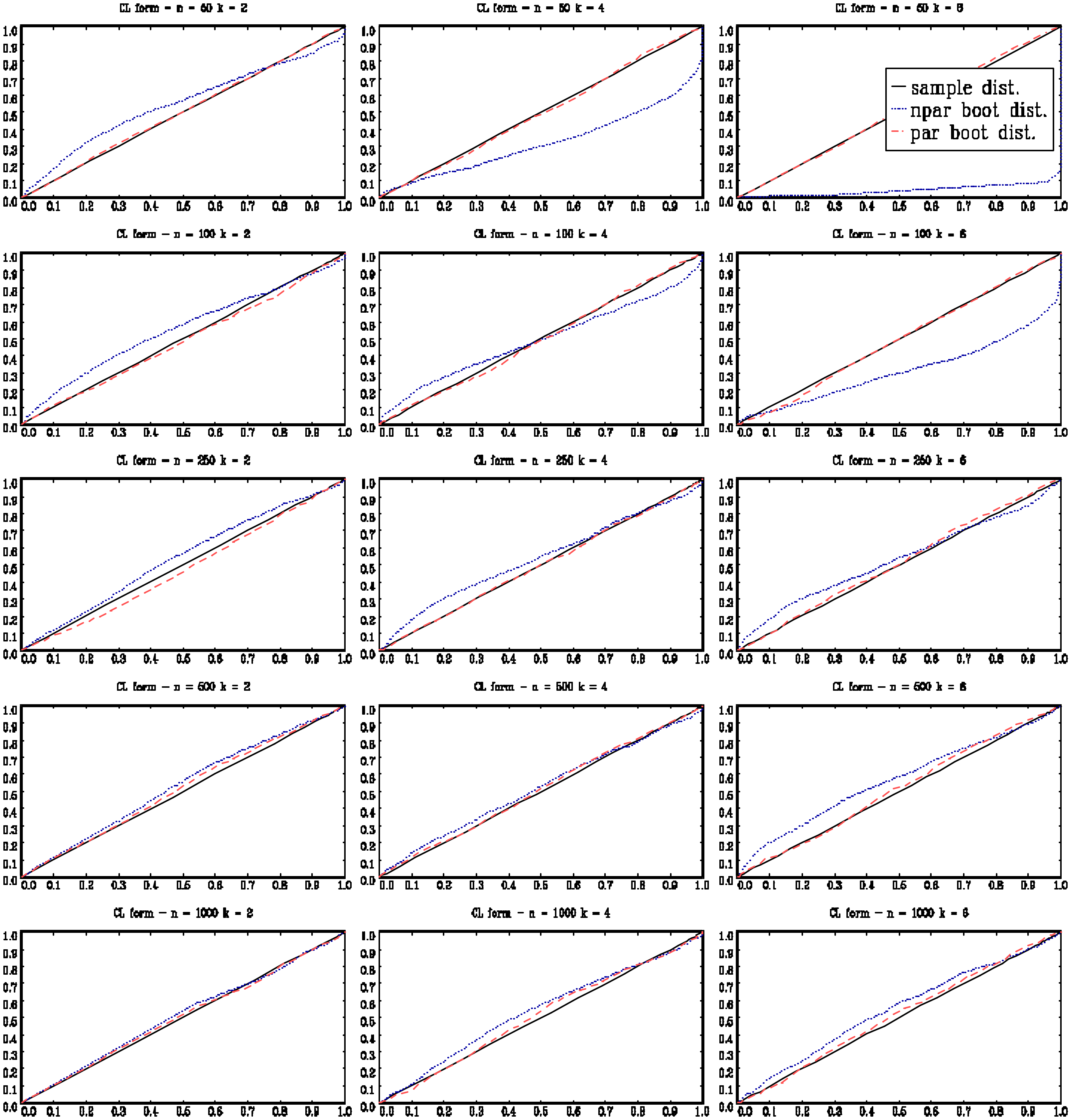


Figure 11

QQ Plot Comparing Bootstrap Distributions with Sampling Distribution

Full White Version of Linear Regression IM Test

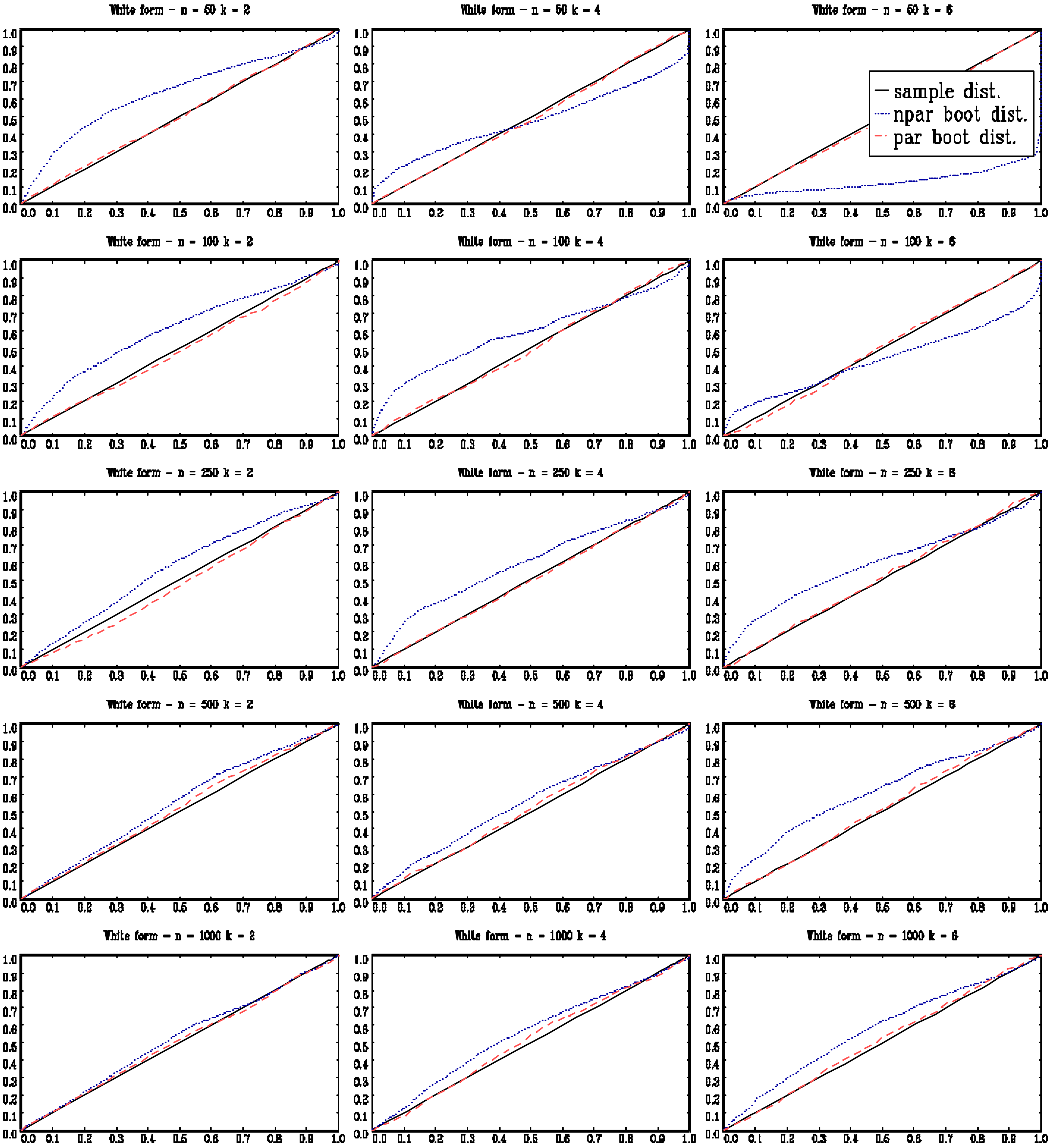


Figure 12

QQ Plot Comparing Bootstrap Distributions with Sampling Distribution

Full Chesher-LancasterVersion of Probit IM Test

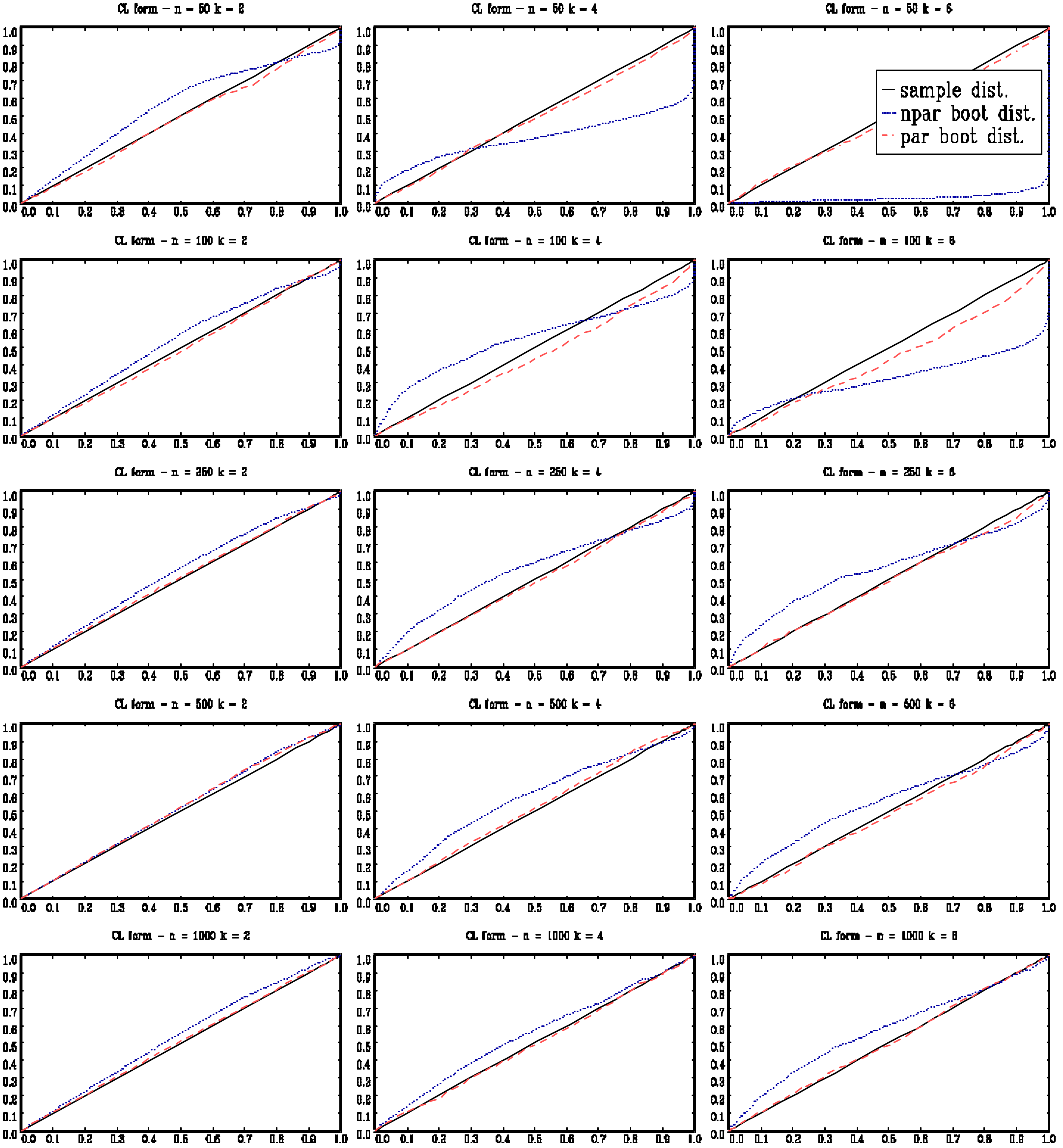


Figure 13

QQ Plot Comparing Bootstrap Distributions with Sampling Distribution

Full White Version of Probit IM Test

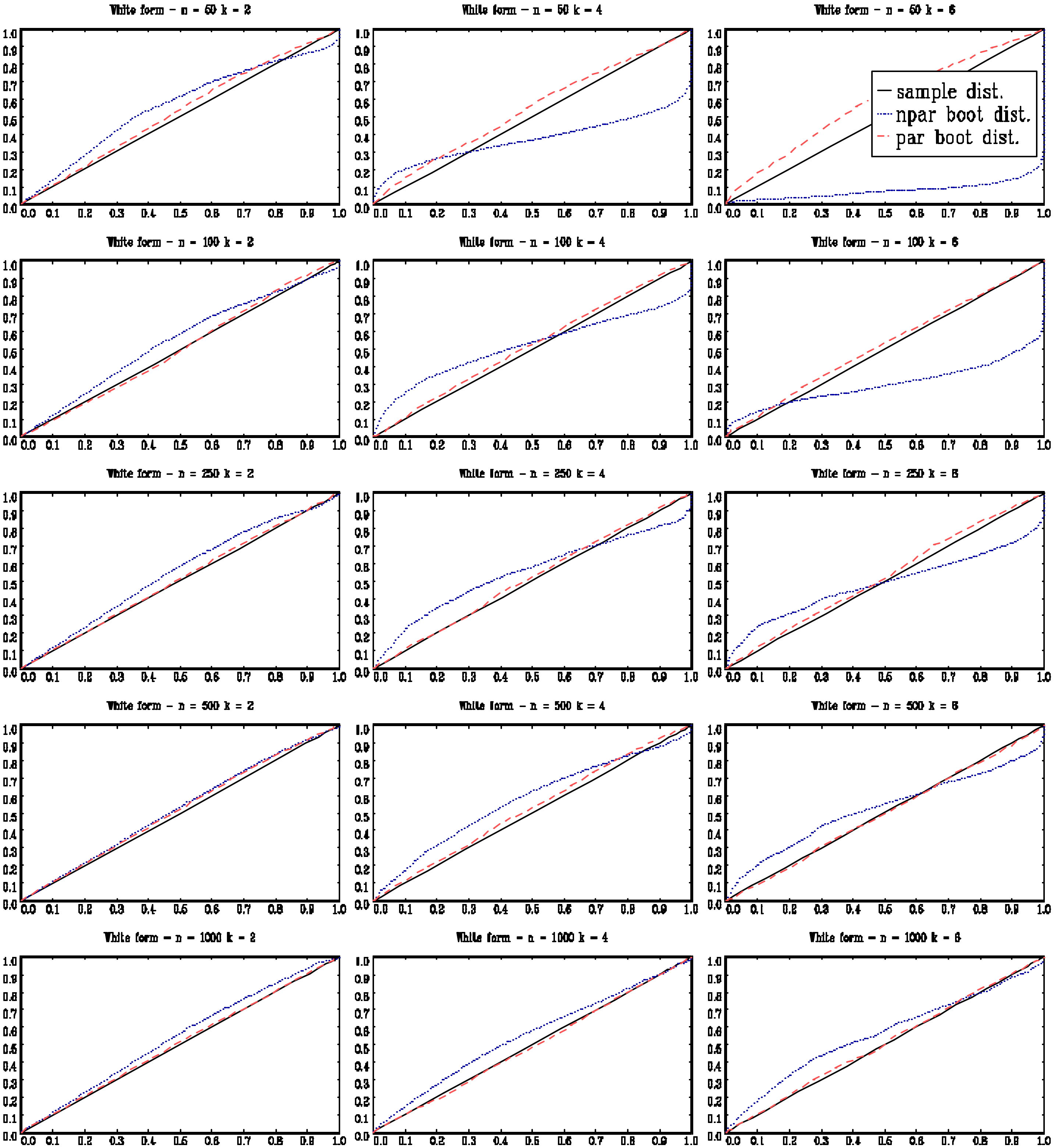


Figure 14

Plot Comparing Histogram of Bootstrap Distributions with Sampling Distribution

Full Chesher-Lancaster Version of Linear Regression IM Test

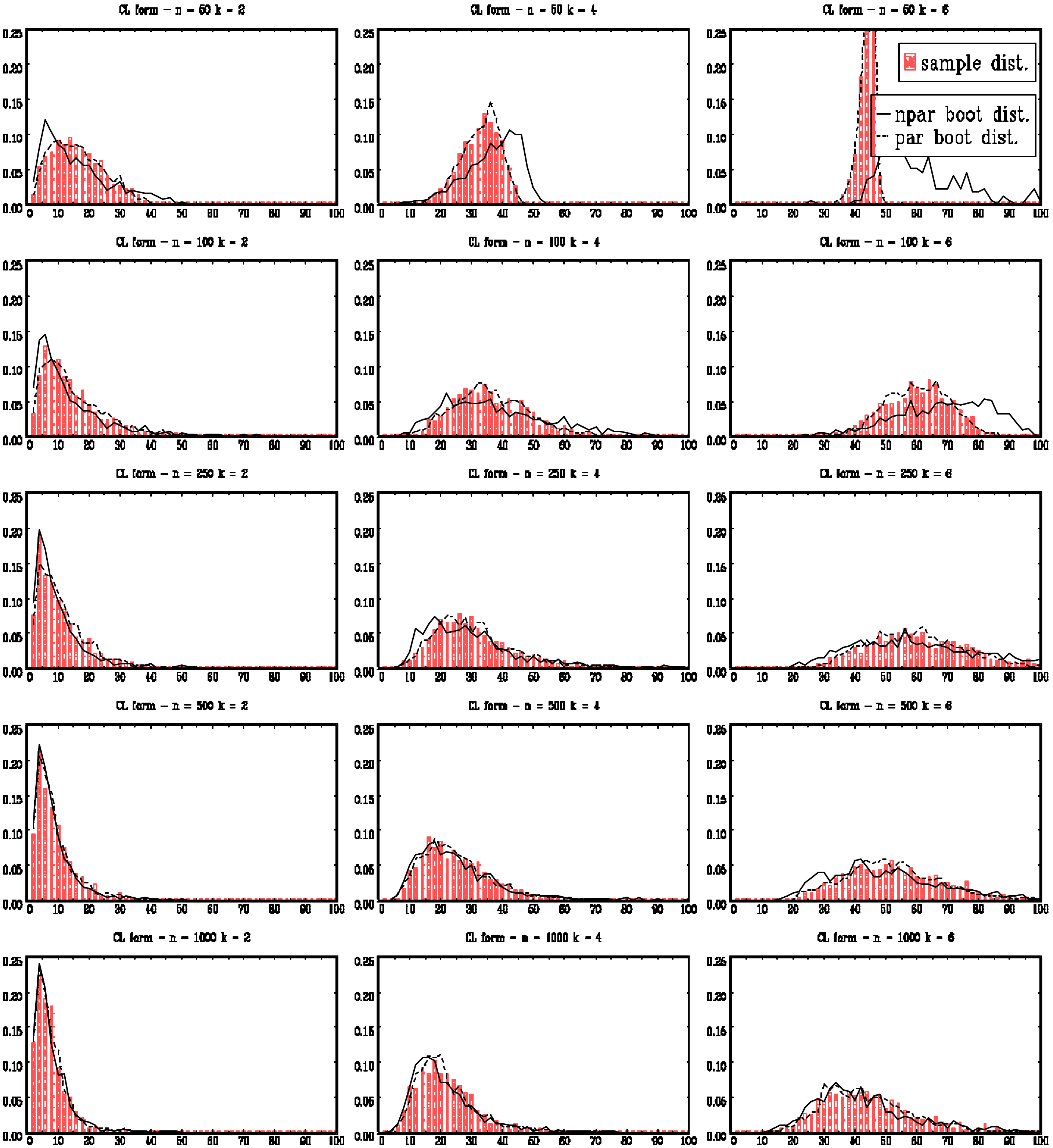


Figure 15

Plot Comparing Histogram of Bootstrap Distributions with Sampling Distribution

Full White Version of Linear Regression IM Test

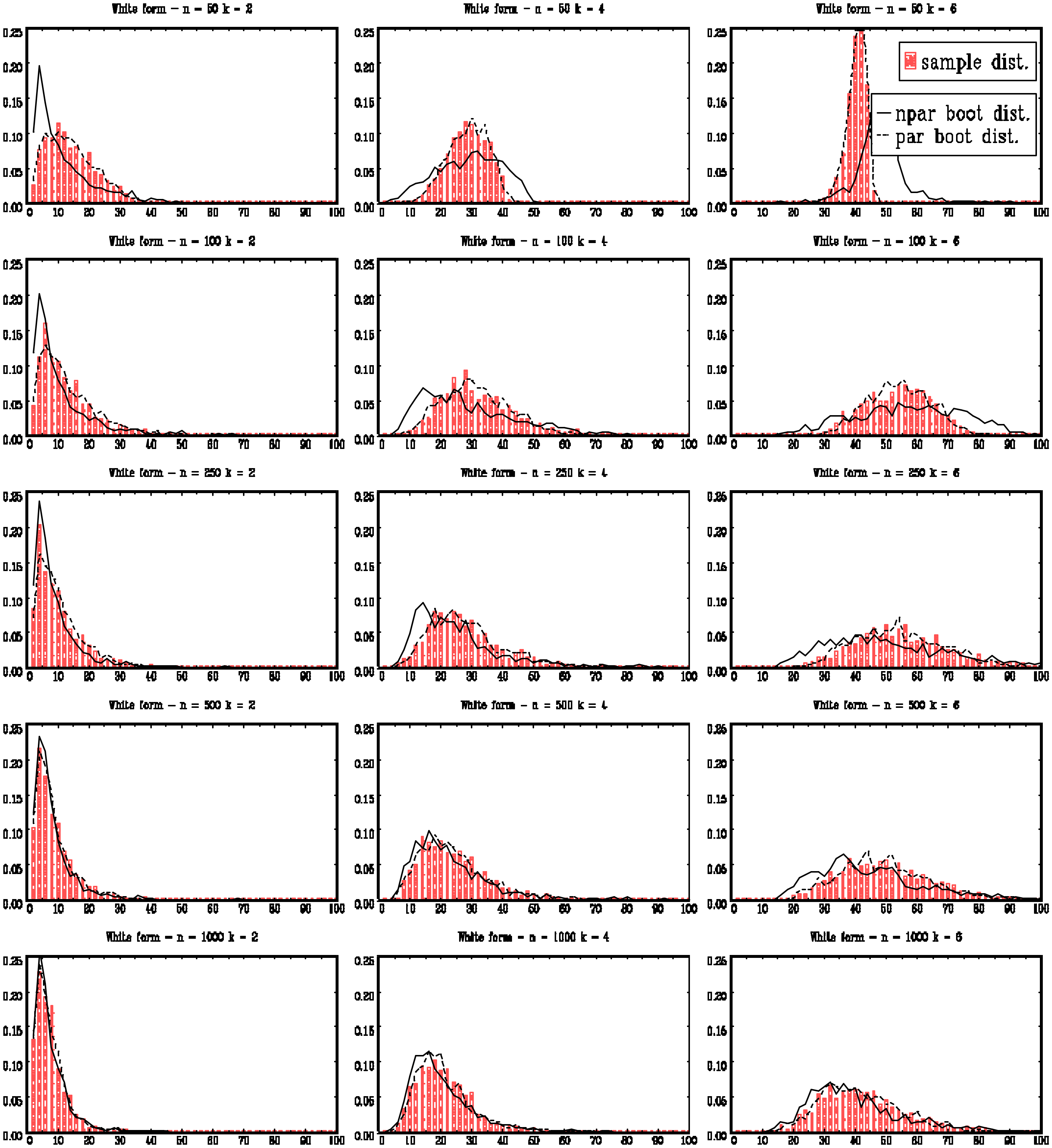


Figure 16

Plot Comparing Histogram of Bootstrap Distributions with Sampling Distribution

Full Chesher-Lancaster Version of Probit IM Test

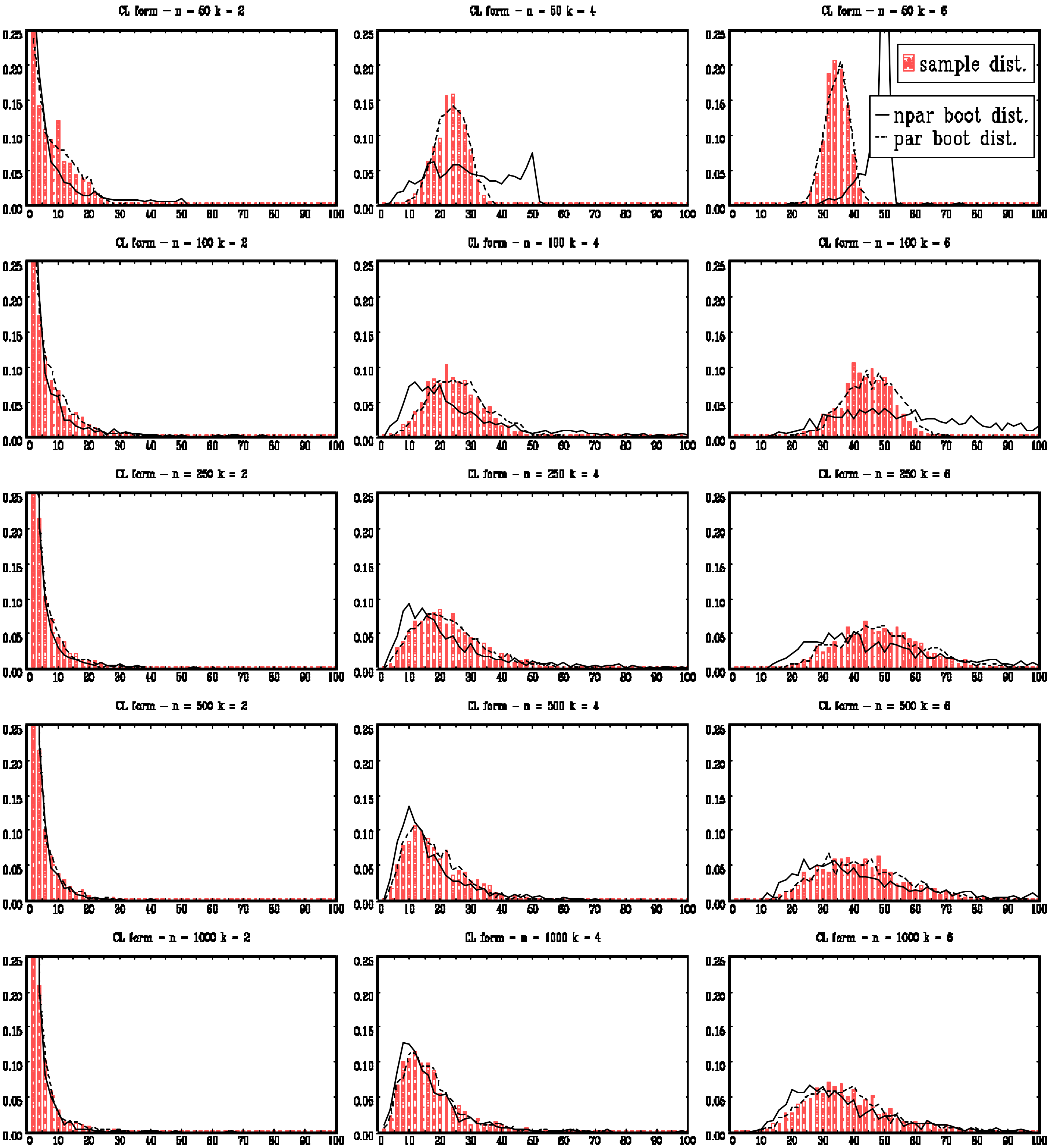


Figure 17

Plot Comparing Histogram of Bootstrap Distributions with Sampling Distribution

Full Chesher-Lancaster Version of Probit IM Test

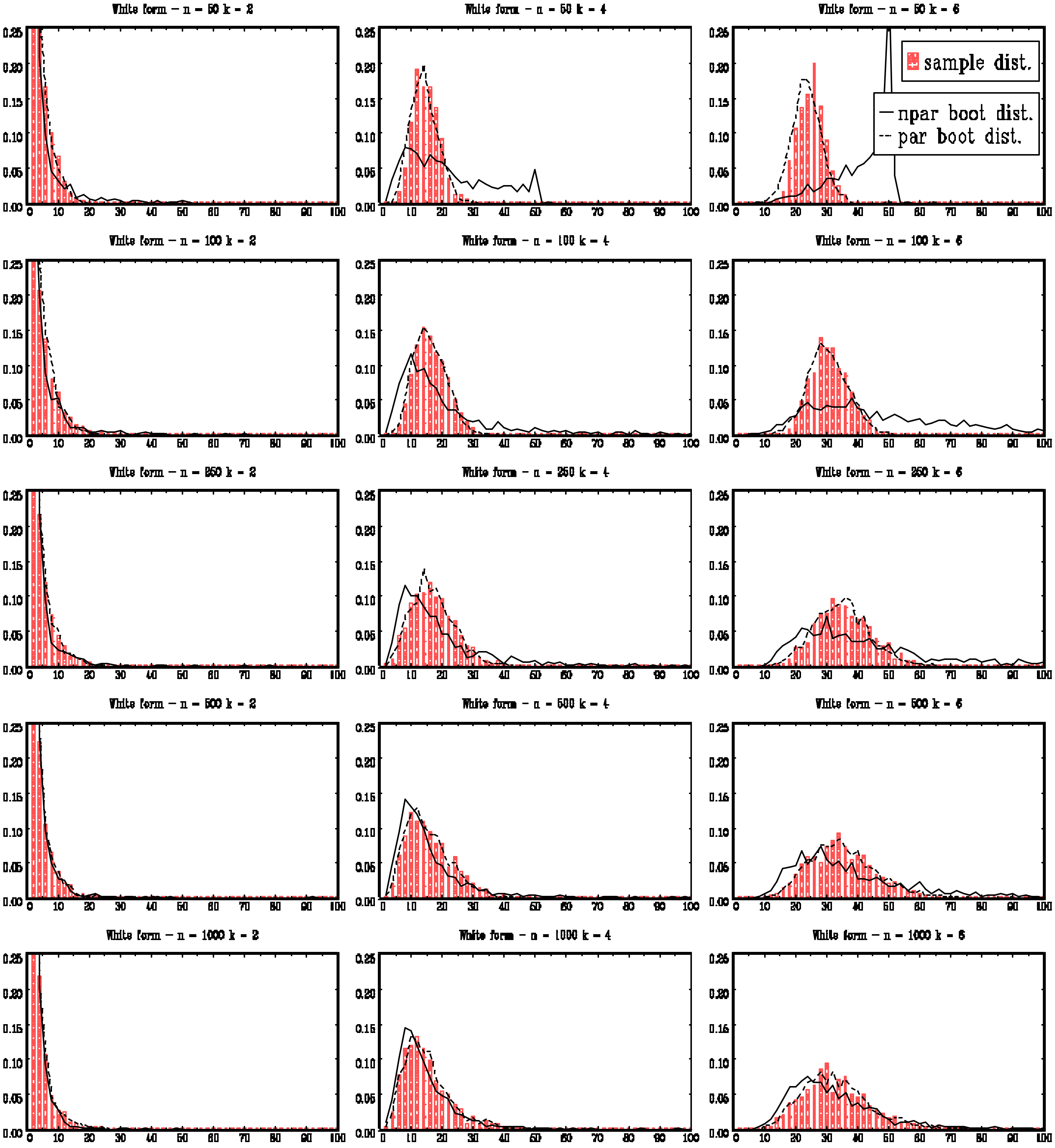
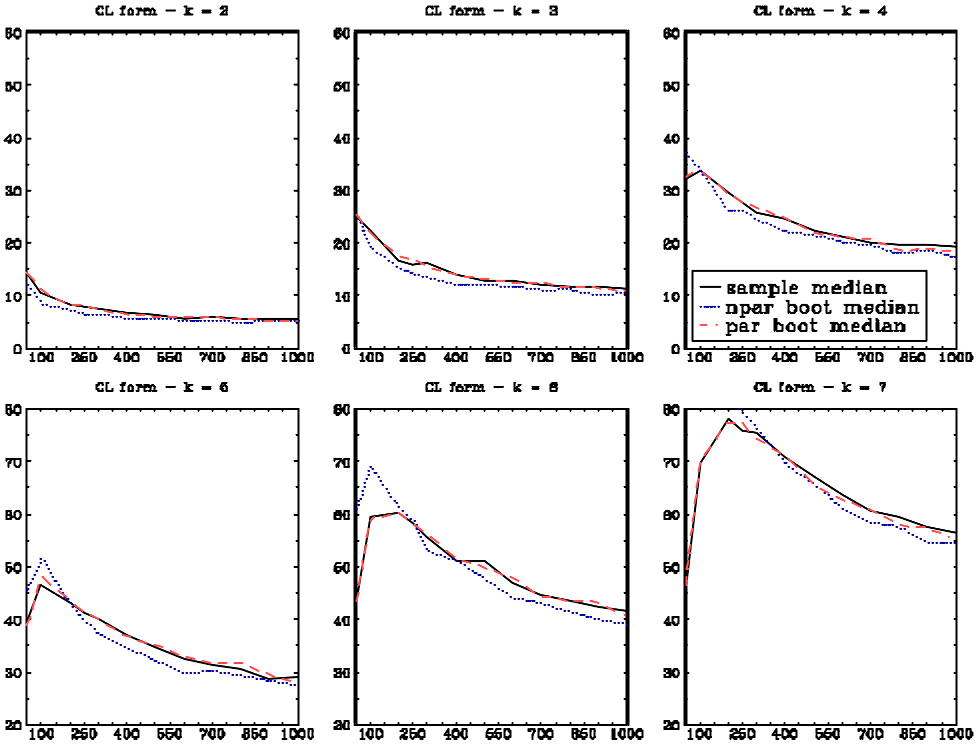


Figure 18

Median Realizations from Bootstrap distributions

Full Chesher-Lancaster Version of Linear Regression IM Test



Full White Version of Linear Regression IM Test

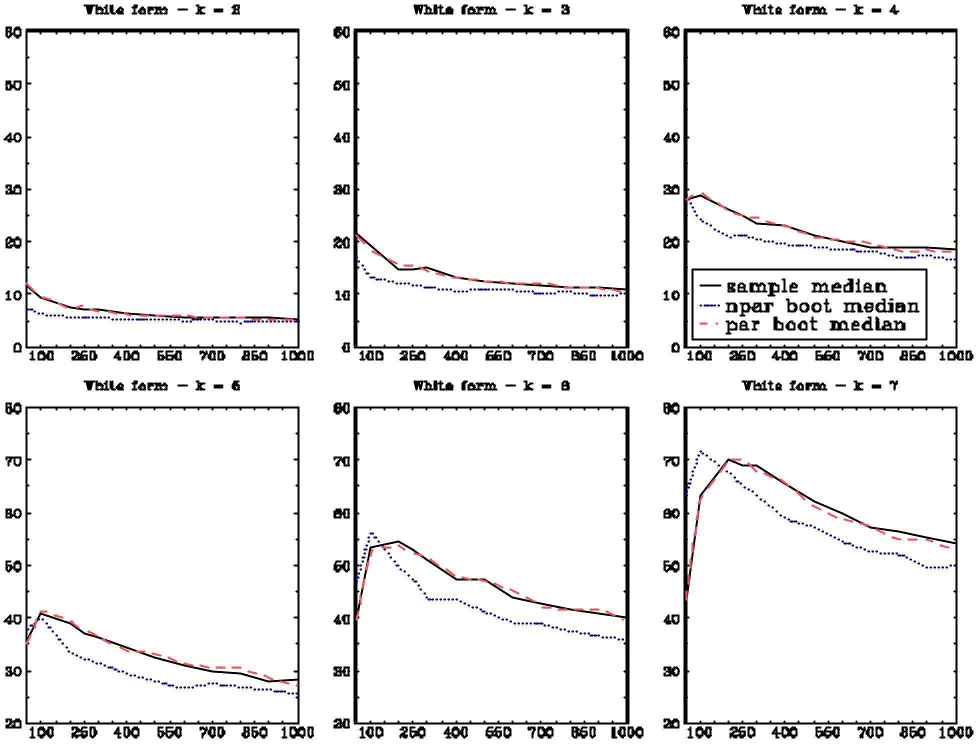
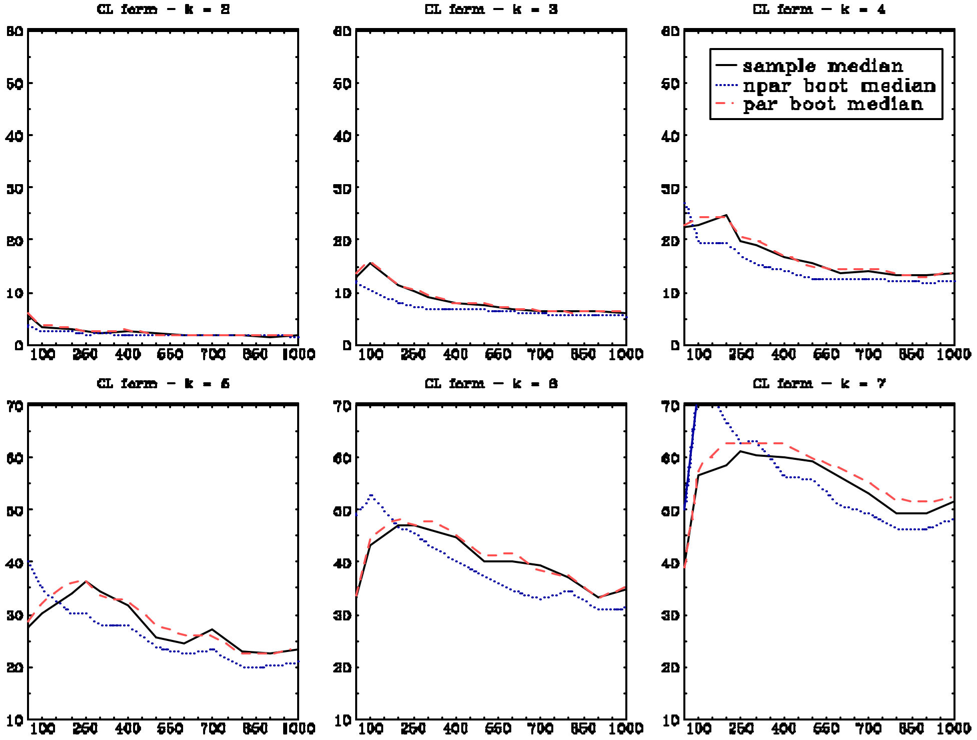


Figure 19

Median Realizations from Bootstrap distributions

Full Chesher-Lancaster Version of Probit IM Test



Full White Version of Probit IM Test

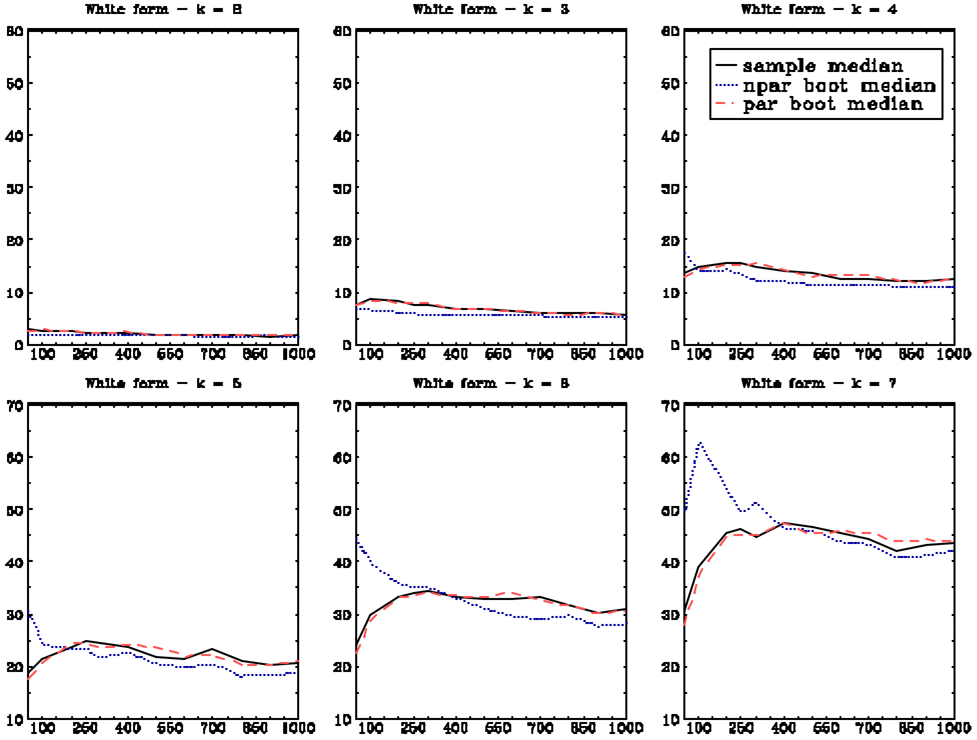
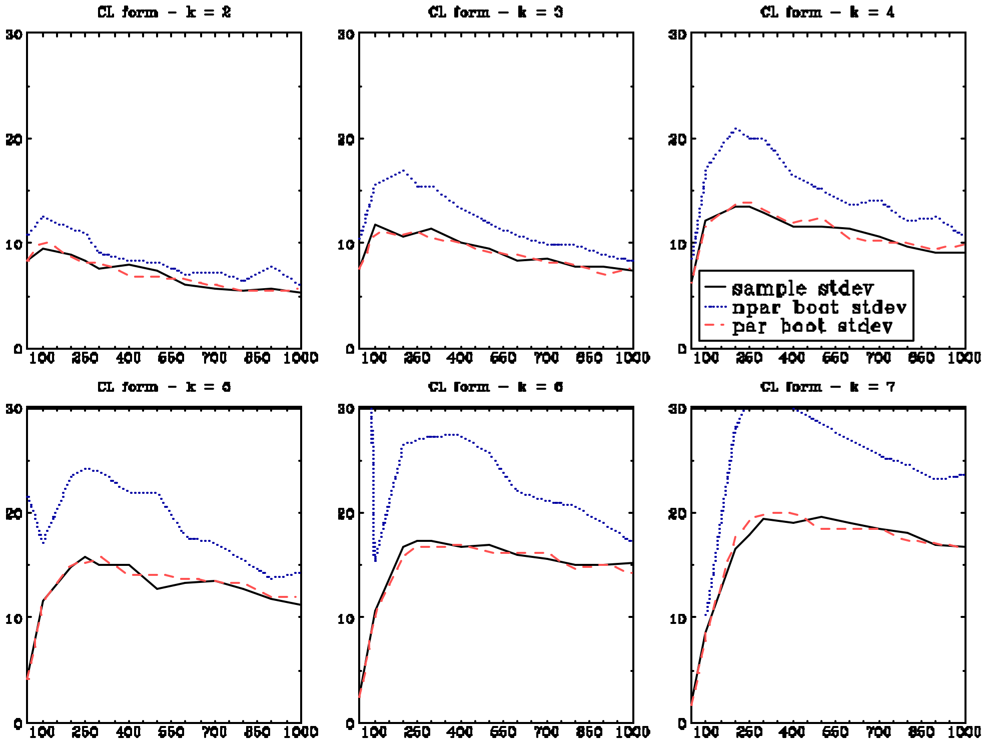


Figure 20

Standard Deviation of Realizations from Bootstrap distributions

Full Chesher-Lancaster Version of Linear Regression IM Test



Full White Version of Linear Regression IM Test

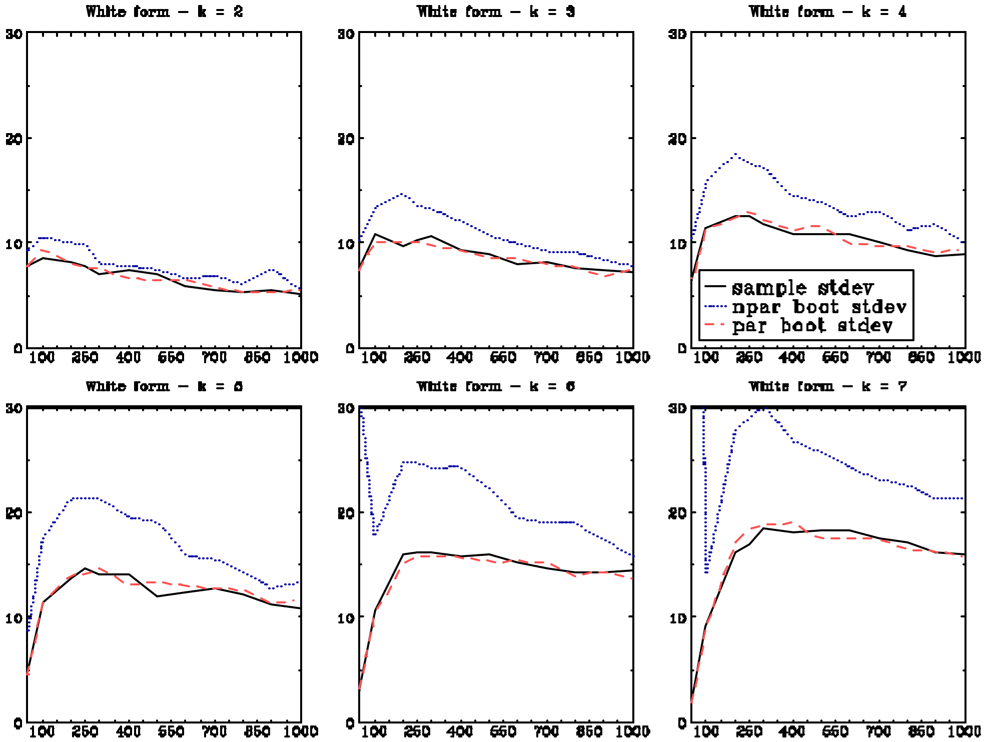
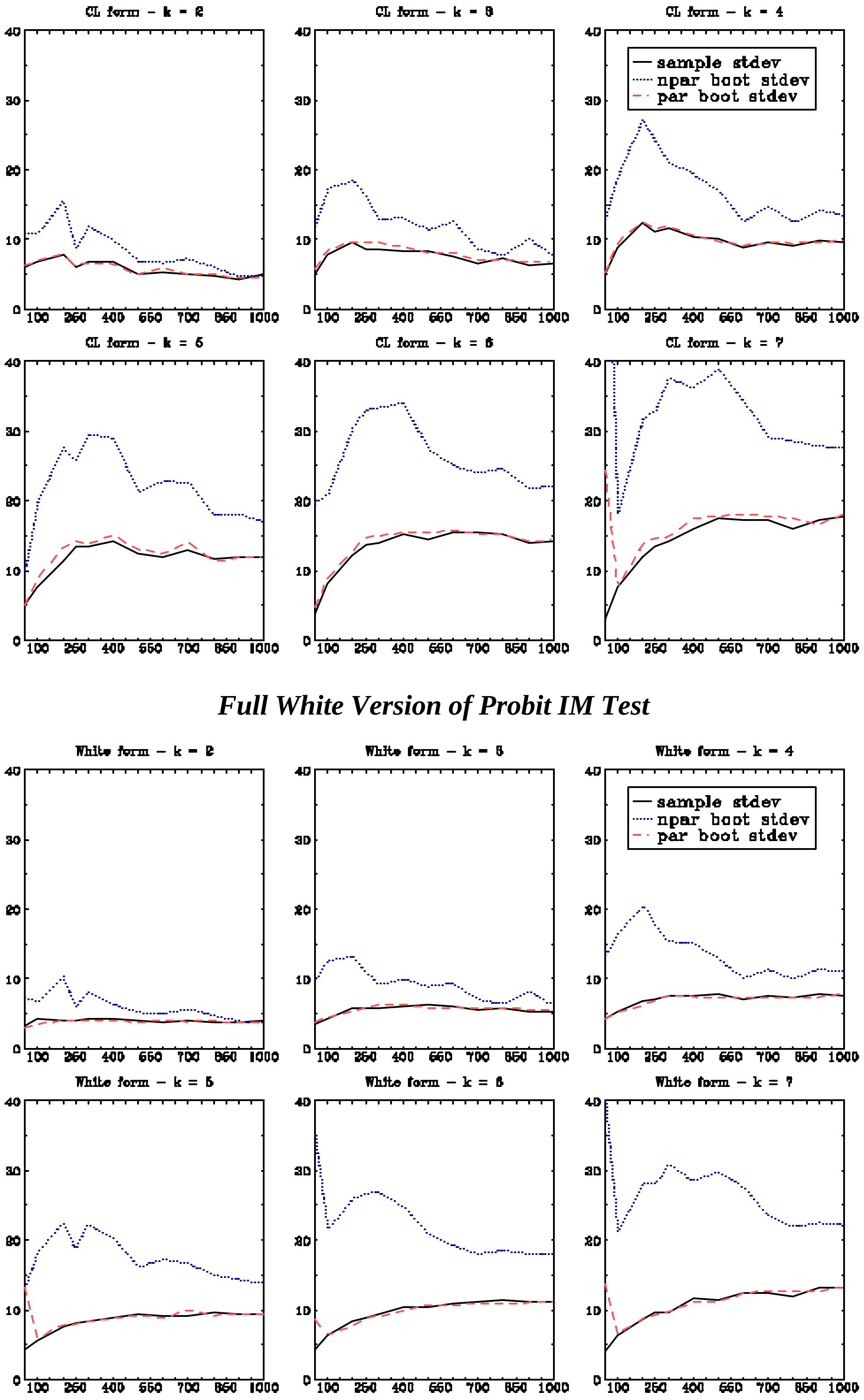


Figure 21

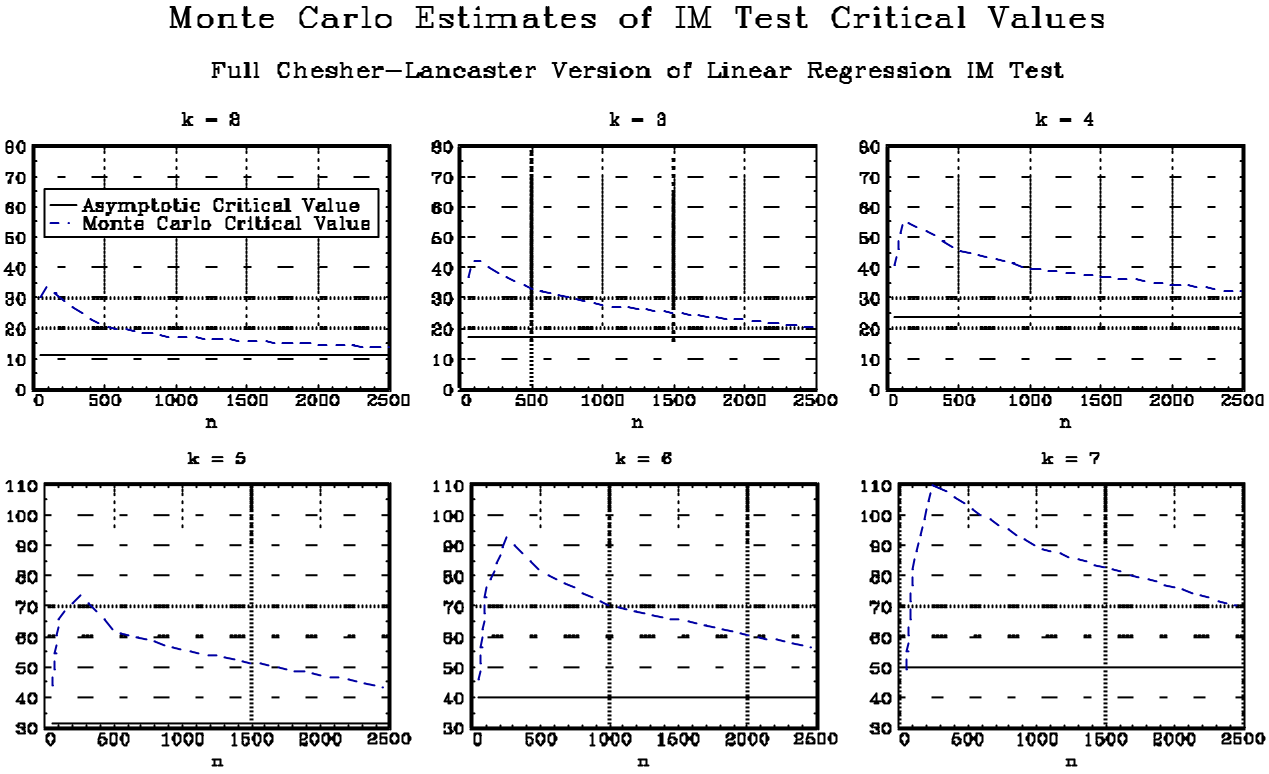
Standard Deviation of Realizations from Bootstrap distributions

Full Chesher-Lancaster Version of Probit IM Test



Full White Version of Probit IM Test

Figure 22


Monte Carlo Estimates of IM Test Critical Values
Full Chesher–Lancaster Version of Linear Regression IM Test
k = 2
k = 3
k = 4
k = 5
k = 6
k = 7
Asymptotic Critical Value
Monte Carlo Critical Value
0
10
20
30
40
50
60
70
80
90
100
110
0
500
1000
1500
2000
2500
n

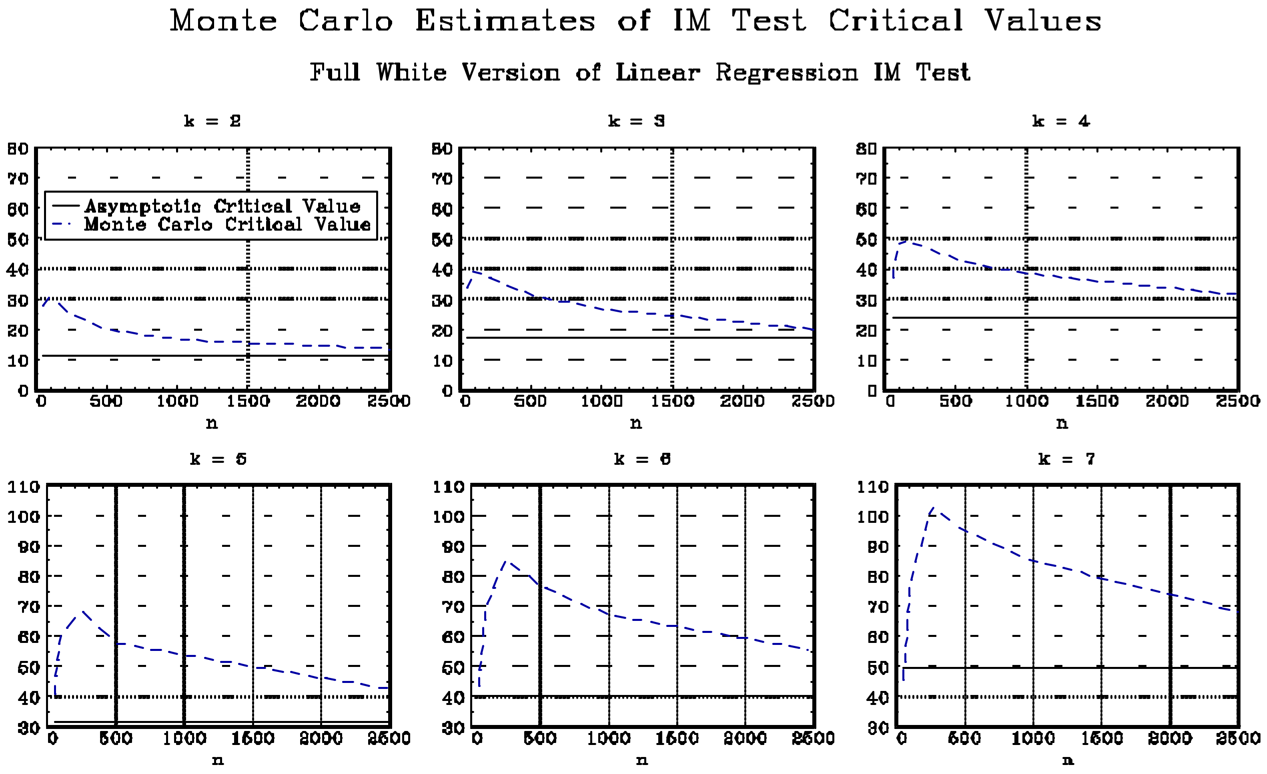
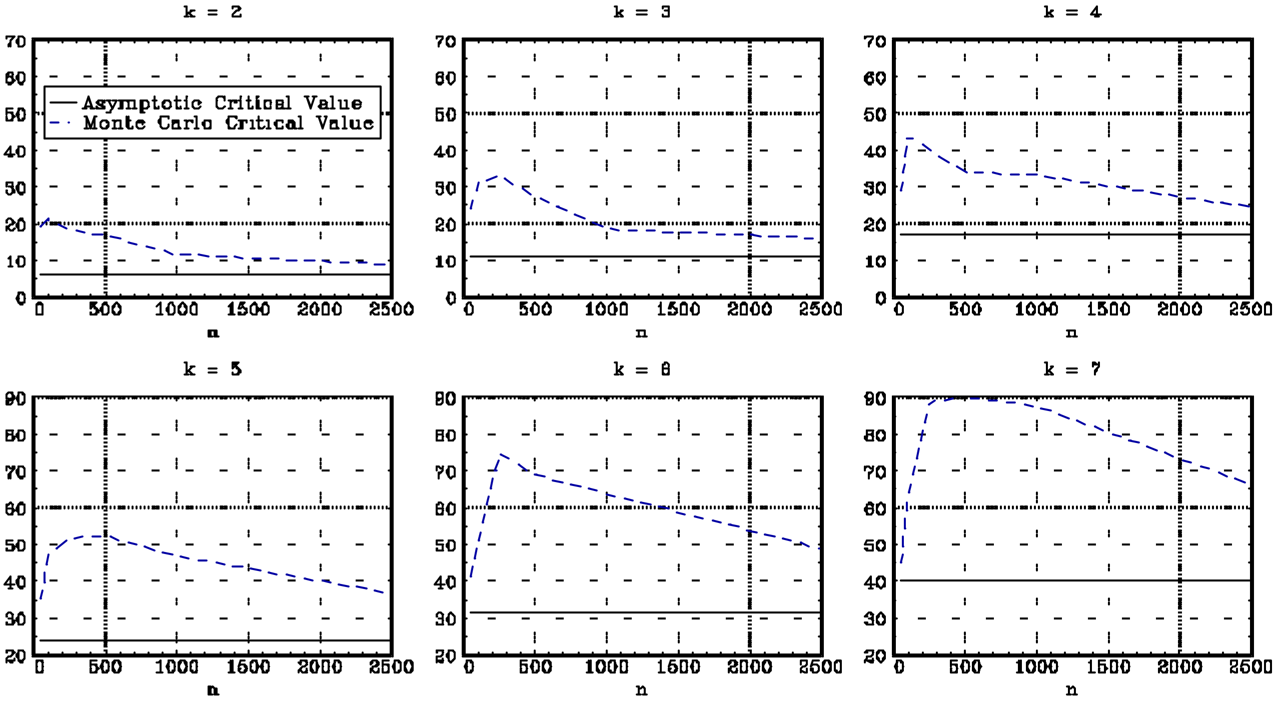

Monte Carlo Estimates of IM Test Critical Values
Full White Version of Linear Regression IM Test
k = 2
k = 3
k = 4
k = 5
k = 6
k = 7
Asymptotic Critical Value
Monte Carlo Critical Value
0
10
20
30
40
50
60
70
80
90
100
110
0
500
1000
1500
2000
2500
n

Figure 23

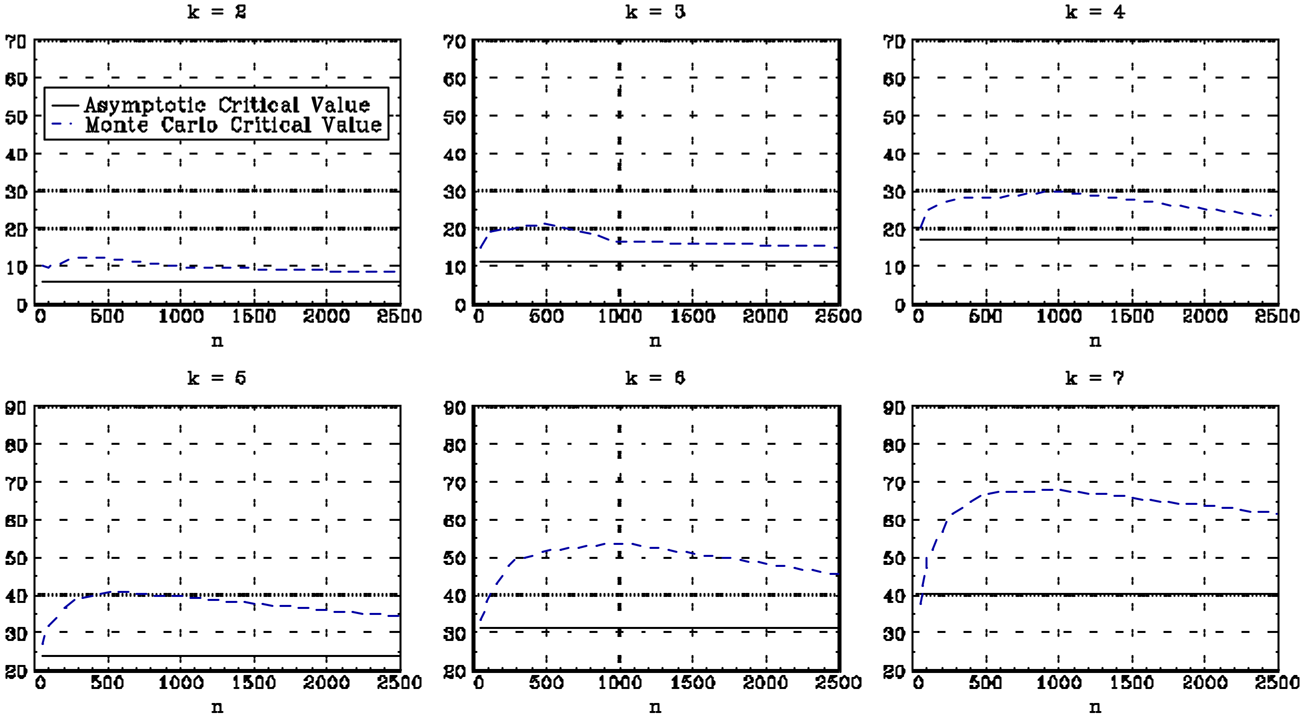
Monte Carlo Estimates of IM Test Critical Values

Full Chesher-Lancaster Version of Probit IM Test



Monte Carlo Estimates of IM Test Critical Values

Full White Version of Probit IM Test


k = 2
k = 3
k = 4
k = 5
k = 6
k = 7
Asymptotic Critical Value
Monte Carlo Critical Value
n