Graphical Methods for Investigating the Size and Power of Hypothesis Tests

by

Russell Davidson

Department of Economics
Queen's University
Kingston, Ontario, Canada
K7L 3N6

GREQAM
Centre de la Vieille Charité
2 rue de la Charité
13002 Marseille, France

and

James G. MacKinnon

Department of Economics
Queen's University
Kingston, Ontario, Canada
K7L 3N6

Abstract

Simple techniques for the graphical display of simulation evidence concerning the size and power of hypothesis tests are developed and illustrated. Three types of figures — called P value plots, P value discrepancy plots, and size-power curves — are discussed. Some Monte Carlo experiments on the properties of alternative forms of the information matrix test for linear regression models and probit models are used to illustrate these figures. Tests based on the OPG regression generally perform much worse in terms of both size and power than efficient score tests.

This research was supported, in part, by the Social Sciences and Humanities Research Council of Canada. We are grateful to a referee and to participants in workshops at Cornell University, York University, Queen's University, the State University of New York at Albany, GREQAM, the Statistical Society of Canada, and the Canadian Econometric Study Group for helpful comments.

First Version, June, 1994
This Revision, March, 1997

1. Introduction

To obtain evidence on the finite-sample properties of hypothesis testing procedures, econometricians generally resort to simulation methods. As a result, many, if not most, of the papers that deal with specification testing and other forms of hypothesis tests include some Monte Carlo results. It is often a challenge to present these results in a form that is reasonably compact and easy to comprehend. This paper discusses some simple graphical methods that can be very useful for presenting simulation results on the size and the power of test statistics. The graphs convey much more information, in a more easily assimilated form, than tables can do.

Consider a Monte Carlo experiment in which N realizations of some test statistic τ are generated using a data generating process, or DGP, that is a special case of the null hypothesis. We may denote these simulated values by τ_j, $j = 1, \ldots, N$. Unless τ is extraordinarily expensive to compute (as it may be if bootstrapping is involved; see Section 3), N will generally be a large number, probably 5000 or more. In practice, of course, several different test statistics may be generated on each replication.

The conventional way to report the results of such an experiment is to tabulate the proportion of the time that τ_j exceeds one or more critical values, such as the 1%, 5%, and 10% values for the asymptotic distribution of τ. This approach has at least two serious disadvantages. First of all, the tables provide information about only a few points on the finite-sample distribution of τ; as we shall see in Section 4, this can be an important limitation. Secondly, the tables require some effort to interpret, and they generally do not make it easy to see how changes in the sample size, the number of degrees of freedom, and other factors affect test size. In this paper, we advocate graphical methods that provide more information, are easy to implement, and yield graphs that are easy to interpret.

The plan of the paper is as follows. In the next section, we discuss the graphs that we are proposing for experiments concerning test size. Then, in Section 3, we briefly discuss several forms of the information matrix (IM) test statistic, which was proposed originally by White (1982), for the normal linear regression model and the probit model. In Section 4, we present a number of Monte Carlo results on various forms of the IM test to illustrate the use of P value plots and P value discrepancy plots. In Section 5, we discuss and illustrate methods for smoothing P value discrepancy plots. Finally, in Section 6, we discuss and illustrate the use of size-power curves.

Some of the results on IM tests in Sections 4 and 6 are new and interesting. In particular, we find that bootstrapping IM tests in probit models dramatically improves their performance under the null. We also find that, for both linear regression models and probit models, the outer-product-of-the-gradient (OPG) form of the IM test generally has substantially lower power than other forms of the test, when all the forms are properly size-corrected.

2. *P* Value Plots and *P* Value Discrepancy Plots

All of the graphs we discuss are based on the empirical distribution function, or EDF, of the P values of the τ_j. The P value of τ_j is the probability of observing a value of τ as extreme as or more extreme than τ_j, according to some distribution $F(\tau)$. This distribution could be the asymptotic distribution of τ, or it could be a distribution derived by bootstrapping, or it could be an approximation to the (generally unknown) finite-sample distribution of τ. For notational simplicity, we shall assume that there is only one P value associated with τ_j, namely, $p_j \equiv p(\tau_j)$. Precisely how p_j is defined will vary. For example, if τ is asymptotically distributed as $\chi^2(r)$ and $F_{\chi^2}(x, r)$ denotes the c.d.f. of the $\chi^2(r)$ distribution evaluated at x, then $p_j = 1 - F_{\chi^2}(\tau_j, r)$.

The EDF of the p_j is simply an estimate of the c.d.f. of $p(\tau)$. At any point x_i in the $(0, 1)$ interval, it is defined by

$$\hat{F}(x_i) \equiv \frac{1}{N} \sum_{j=1}^{N} I(p_j \leq x_i), \tag{1}$$

where $I(p_j \leq x_i)$ is an indicator function that takes the value 1 if its argument is true and 0 otherwise. The EDF (1) is often evaluated at every data point. However, this is unnecessary. When N is large, as it often will be, storage space can be conserved by evaluating the EDF only at m points $x_i, i = 1, \ldots, m$, which should be chosen in advance so as to provide a reasonable snapshot of the $(0, 1)$ interval, or of that part of it which is of interest.

It is difficult to state categorically how large m should be and how the x_i should be chosen. A quite parsimonious way to choose the x_i is

$$x_i = .002, .004, \ldots, .01, .02, \ldots, .99, .992, \ldots, .998 \quad (m = 107). \tag{2}$$

Another choice, which may give slightly better results, is

$$x_i = .001, .002, \ldots, .010, .015, \ldots, .990, .991, \ldots, .999 \quad (m = 215). \tag{3}$$

For both (2) and (3), there are extra points near 0 and 1 in order to ensure that we do not miss any unusual behavior in the tails. As we shall see in Section 6, it may be necessary to add additional points in certain cases. Note that, because we evaluate the EDF only at a relatively small number of points, it is straightforward to employ variance reduction techniques. If control or antithetic variates are available, the techniques proposed by Davidson and MacKinnon (1992b) to estimate tail areas can simply be applied to each point on the EDF.

The simplest graph that we will discuss is a plot of $\hat{F}(x_i)$ against x_i. We shall refer to such a graph as a *P value plot*. If the distribution of τ used to compute the p_j is correct, each of the p_j should be distributed as uniform $(0, 1)$. Therefore, when $\hat{F}(x_i)$ is plotted against x_i, the resulting graph should be close to the 45° line. As we shall see in Section 4, P value plots allow us to distinguish at a glance among test statistics that systematically over-reject, test statistics that systematically under-reject, and test statistics that reject about the right

proportion of the time. However, because P value plots for all test statistics that behave approximately the way they should will look roughly like 45° lines, these plots are not very useful for distinguishing among such test statistics.

For dealing with test statistics that are well-behaved, it is much more revealing to graph $\hat{F}(x_i) - x_i$ against x_i. We shall refer to such a graph as a *P value discrepancy plot*. These plots have advantages and disadvantages. They convey a lot more information than P value plots for test statistics that are well behaved. However, some of this information is spurious, simply reflecting experimental randomness. In Section 5, we therefore discuss semi-parametric methods for smoothing them. Moreover, because there is no natural scale for the vertical axis, P value discrepancy plots can be harder to interpret than P value plots.

P value plots and P value discrepancy plots are very useful for dealing with test size, but not useful for dealing with test power. In Section 6, we will discuss graphical methods for comparing the power of competing tests using *size-power curves*. These curves can be constructed using two EDFs, one for an experiment in which the null hypothesis is true, and one for an experiment in which it is false.

It would be more conventional to graph the EDF of the τ_j instead of the EDF of their P values. If the former were graphed against a hypothesized distribution, the result would be what is often called a PP plot; see Wilk and Gnanadesikan (1968). Plotting P values makes it much easier to interpret the plots, since what they should look like will not depend on the null distribution of the test statistic. This fact also makes it easy to compare test statistics which have different distributions under the null, and to compare different procedures for making inferences from the same test statistics. Note that P value plots are really just PP plots of P values. Contrary to what we believed initially, we are not the first to employ them; see, for example, Fisher, Mammen, and Marron (1994). There are also some recent applications in econometrics, which generally cite the discussion paper version of this paper; see, for example, Hansen, Heaton, and Yaron (1996) and West and Wilcox (1996).

Another common type of plot is a QQ plot, in which the empirical quantiles of the τ_j are plotted against the actual quantiles of their hypothesized distribution. If the empirical distribution is close to the hypothesized one, the plot will be close to the 45° line. This approach, which has been used by Chesher and Spady (1991), among others, can yield useful information. If the plot is linear, then a suitable change of scale is all that is needed to make the test statistic perform as it should. This would suggest that a Bartlett-type correction may be appropriate; see Cribari-Neto and Cordeiro (1996) and Chesher and Smith (1997). On the other hand, if the plot is nonlinear, such a correction will not be adequate. Thus QQ plots can fruitfully be employed to see whether Bartlett-type corrections are worth investigating.

However, QQ plots also have disadvantages. One serious problem is that a QQ plot has no natural scale for the axes. Thus, if the hypothesized distribution changes, so will that scale. As we shall see in Section 4, this makes it difficult to plot on the same axes test statistics that have different distributions under the null. One way to get around this — see, for example, Mammen (1992) — is to

employ a QQ plot of P values. This will convey exactly the same information as a P value plot, except that the plot will be reflected about the 45° line. In our view, however, QQ plots of P values are harder to read than P value plots, because it is more natural to have nominal test size on the horizontal axis.

Of course, graphical methods by themselves are not always enough. When test performance depends on a number of factors, the two-dimensional nature of most graphs and all tables can be limiting. In such cases, it may be desirable to supplement the graphs with estimated response surfaces which relate size or power to sample size, parameter values, and so on; see, for example, Hendry (1984).

3. Alternative Forms of the Information Matrix Test

In order to illustrate and motivate P value and P value discrepancy plots, we use them to present the results of a study of the properties of alternative forms of the information matrix (IM) test. We compare tests based on the outer-product-of-the-gradient (OPG) regression, which was proposed as a way to compute IM tests by Chesher (1983) and Lancaster (1984), with two other forms of the IM test, including one proposed by us in Davidson and MacKinnon (1992a). Table 1 of that paper, which occupies a page and a half, is a particularly striking example of the disadvantages of presenting simulation results for test statistics in a non-graphical way. Even very recent papers about the IM test often fail to use graphical methods; one paper that, in our view, would benefit greatly from doing so is Cribari-Neto (1997).

We deal with two models, the univariate linear regression model with normal errors, and the probit model. In the former case, IM tests are pivotal, which means that their distributions do not depend on any nuisance parameters, and bootstrapping will therefore work perfectly. In the latter case, IM tests are not pivotal, so that bootstrapping will not work perfectly. However, because they are asymptotically pivotal, bootstrapping should yield inferences that are accurate to higher order, in the sample size, than the inferences provided by asymptotic theory; see Beran (1988), Horowitz (1994), and Davidson and MacKinnon (1996a).

Three of the variants of the IM test that we shall deal with pertain to the normal linear regression model

$$y_t = \delta_1 + \sum_{i=2}^{k} \delta_i X_{ti} + u_t, \quad u_t \sim \text{NID}(0, \sigma^2), \quad t = 1, \ldots, n. \tag{4}$$

The regressors X_{ti} are normal random variables, independent across observations, and equicorrelated with correlation coefficient one-half. Because all versions of the IM test for this model are pivotal, the values of σ and the δ_i are chosen arbitrarily.

The OPG variant of the IM test statistic is obtained by regressing an $n-$vector of 1s on $\tilde{u}_t X_{ti}$, for $i = 1, \ldots, k$, and on $\frac{1}{2}(k^2 + 3k)$ test regressors.

These test regressors are functions of $\tilde{e}_t \equiv \tilde{u}_t/\tilde{\sigma}$ and the X_{ti}. Here $\tilde{u}_t$ is the t^{th} OLS residual and $\tilde{\sigma}^2 = (1/n)\sum_{t=1}^{n} \tilde{u}_t^2$. There are $k(k+1)/2 - 1$ test regressors of the form $(\tilde{e}_t^2 - 1)X_{ti}X_{tj}$, which test for heteroskedasticity, k regressors of the form $(\tilde{e}_t^3 - 3\tilde{e}_t)X_{ti}$, which test for skewness interacting with the regressors, and one regressor of the form $\tilde{e}_t^4 - 5\tilde{e}_t^2 + 2$, which tests for kurtosis. The test statistic is n minus the sum of squared residuals from the regression, and it is asymptotically distributed as $\chi^2\big(\frac{1}{2}(k^2 + 3k)\big)$.

The DLR form of the IM test is a bit more complicated. It involves a double-length artificial regression with $2n$ "observations," and the test statistic is $2n$ minus the sum of squared residuals from this regression. The number of regressors is the same as for the OPG test, and so is the asymptotic distribution of the test statistic. See Davidson and MacKinnon (1984, 1992a).

A third form of the IM test, which is less widely available than the other two, is the efficient score, or ES, form. The ES form of the Lagrange Multiplier test is often considered to have optimal or nearly optimal properties, because the only random quantities in the estimate of the information matrix are the restricted maximum likelihood parameter estimates. In this case, the ES form of the IM test is actually the sum of three test statistics:

$$\tfrac{1}{2}\boldsymbol{h}_2'\boldsymbol{Z}\big(\boldsymbol{Z}'\boldsymbol{Z}\big)^{-1}\boldsymbol{Z}'\boldsymbol{h}_2 + \tfrac{1}{6}\boldsymbol{h}_3'\boldsymbol{X}\big(\boldsymbol{X}'\boldsymbol{X}\big)^{-1}\boldsymbol{X}'\boldsymbol{h}_3 + \frac{1}{24n}\sum_{t=1}^{n}\big(\tilde{e}_t^4 - 3\big)^2, \tag{5}$$

where $\boldsymbol{X}$ has typical element X_{ti}, $\boldsymbol{Z}$ has typical element $X_{ti}X_{tj}$, $\boldsymbol{h}_2$ has typical element $\tilde{e}_t^2 - 1$, and $\boldsymbol{h}_3$ has typical element $\tilde{e}_t^3$. These three statistics test for heteroskedasticity, skewness, and kurtosis, respectively; see Hall (1987).

The other two variants of the IM test that we shall deal with pertain to the probit model. Consider any binary response model of the form

$$E(y_t \,|\, \Omega_t) = D(\boldsymbol{X}_t\boldsymbol{\beta}), \quad t = 1, \ldots, n, \tag{6}$$

where y_t is a 0-1 dependent variable, Ω_t denotes an information set, $D(\cdot)$ is a twice continuously differentiable, increasing function that maps from the real line to the 0-1 interval, $\boldsymbol{X}_t$ is a $k \times 1$ row vector of observations on variables that belong to Ω_t, and $\boldsymbol{\beta}$ is a k-vector of unknown parameters. The contribution to the loglikelihood by observation t is

$$\ell_t = y_t \log D(\boldsymbol{X}_t\boldsymbol{\beta}) + (1 - y_t)\log\big(1 - D(\boldsymbol{X}_t\boldsymbol{\beta})\big), \tag{7}$$

and the derivative with respect to β_i is

$$\frac{\partial \ell_t}{\partial \beta_i} = \frac{X_{ti}d_t(y_t - D_t)}{D_t(1 - D_t)}, \tag{8}$$

where $d_t \equiv d(\boldsymbol{X}_t\boldsymbol{\beta})$ is the derivative of $D_t \equiv D(\boldsymbol{X}_t\boldsymbol{\beta})$. After some algebra, the derivative of (8) with respect to β_j can be seen to be

$$\frac{X_{ti}X_{tj}}{D_t^2(1 - D_t)^2}\big((y_t - D_t)d_t'D_t(1 - D_t) - d_t^2(y_t - D_t)^2\big), \tag{9}$$

where $d'_t \equiv d'(\boldsymbol{X}_t\boldsymbol{\beta})$ is the derivative of d_t. Thus the (i,j) direction of the IM test is given by

$$\frac{\partial^2 \ell_t}{\partial\beta_i\partial\beta_j} + \frac{\partial\ell_t}{\partial\beta_i}\frac{\partial\ell_t}{\partial\beta_j} = \frac{X_{ti}X_{tj}d_t}{D_t(1-D_t)}\frac{d'_t}{d_t}(y_t - D_t). \tag{10}$$

The OPG regression simply involves regressing a vector of 1s on k regressors with typical element (8) and on $k(k+1)/2-1$ test regressors with typical element (10), where all regressors are evaluated at the ML estimates $\tilde{\boldsymbol{\beta}}$. The test statistic is n minus the sum of squared residuals from this regression. The number of test regressors is normally one less than $k(k+1)/2$ because, if there is a constant term, the test regressor for which i and j both index the constant will be perfectly collinear with the regressor for the constant from (8).

It is also quite easy to derive an efficient score variant of the IM test for binary response models; see Orme (1988). Consider the artificial regression

$$\tilde{w}_t(y_t - \tilde{D}_t) = \tilde{w}_t\tilde{d}_t\boldsymbol{X}_t\boldsymbol{b} + \tilde{w}_t\tilde{d}_t(\tilde{d}'_t/\tilde{d}_t)\boldsymbol{W}_t\boldsymbol{c} + \text{residual}, \tag{11}$$

where $\tilde{D}_t \equiv D_t(\boldsymbol{X}_t\tilde{\boldsymbol{\beta}})$, $\tilde{d}_t \equiv d_t(\boldsymbol{X}_t\tilde{\boldsymbol{\beta}})$, $\tilde{d}'_t \equiv d'(\boldsymbol{X}_t\tilde{\boldsymbol{\beta}})$, $\tilde{w}_t \equiv \left(\tilde{D}_t(1-\tilde{D}_t)\right)^{-1/2}$, and $\boldsymbol{W}_t$ is a vector with typical element $X_{ti}X_{tj}$. By standard results for artificial regressions — see Davidson and MacKinnon (1993, Chapter 15) — the explained sum of squares from regression (11) is a test statistic for $\boldsymbol{c} = \boldsymbol{0}$, and from (10) it is clear that this test statistic is testing in the directions of the IM test. That this test statistic is the efficient score variant is evident from the fact that none of the regressors in (11) depend directly on y_t; their only dependence on the data is through the ML estimates $\tilde{\boldsymbol{\beta}}$.

In the case of the probit model, (11) simplifies slightly to

$$\tilde{w}_t(y_t - \tilde{\Phi}_t) = \tilde{w}_t\tilde{\phi}_t\boldsymbol{X}_t\boldsymbol{b} + \tilde{w}_t\tilde{\phi}_t(-\boldsymbol{X}_t\tilde{\boldsymbol{\beta}})\boldsymbol{W}_t\boldsymbol{c} + \text{residual}. \tag{12}$$

This artificial regression is precisely the one for testing the hypothesis that $\boldsymbol{\gamma} = \boldsymbol{0}$ in the model

$$E(y_t \,|\, \Omega_t) = \Phi\left(\frac{\boldsymbol{X}_t\boldsymbol{\beta}}{\exp(\boldsymbol{W}_t\boldsymbol{\gamma})}\right). \tag{13}$$

This model involves a form of heteroskedasticity, and it reduces to the ordinary probit model when $\boldsymbol{\gamma} = \boldsymbol{0}$. Thus we see that, for the probit model, the IM test is equivalent to a test for a certain type of heteroskedasticity.

In our experiments, we deal with a probit model that has between 2 and 5 regressors, one of them a constant term and the rest independent $N(0,1)$ random variates. For this model, $D(\boldsymbol{X}_t\boldsymbol{\beta})$ is the cumulative standard normal distribution function $\Phi(\boldsymbol{X}_t\boldsymbol{\beta})$, $d(\boldsymbol{X}_t\boldsymbol{\beta})$ is the standard normal density $\phi(\boldsymbol{X}_t\boldsymbol{\beta})$, and $d'_t/d_t = -\boldsymbol{X}_t\boldsymbol{\beta}$.

It is well-known that many forms of the IM statistic, most notoriously the OPG form, have very poor finite-sample properties under the null. See, among others, Taylor (1987), Chesher and Spady (1991), Davidson and MacKinnon (1992a), and Horowitz (1994). One way to improve these finite-sample properties

is to use the parametric bootstrap. This idea was investigated by Horowitz (1994), who found that it worked very well for IM tests in probit and tobit models. Horowitz used the bootstrap to compute critical values. An alternative approach, which we prefer, is to use it to compute P values. After a test statistic, say $\hat{\tau}$, has been obtained, B sets of simulated data are generated from a DGP that satisfies the null hypothesis, using the parameter estimates from the actual sample, and B bootstrap test statistics are computed. Suppose that B^* of these test statistics are greater than $\hat{\tau}$, or greater in absolute value if it is a two-tailed test. Then the estimated P value for $\hat{\tau}$ is B^*/B.

In the case of the linear regression model (4), all forms of the IM test statistic are pivotal. This implies that, as $B \to \infty$, the EDF of the bootstrapped P values must tend to the 45° line, and that the size-power curve will be unaffected by bootstrapping; see Beran (1988). Therefore, in this case, we do not need to do a Monte Carlo experiment to see how the parametric bootstrap works; except for experimental error, it will work perfectly. The error terms for the bootstrap samples must be obtained from a normal distribution rather than by resampling from the residuals, since the residuals will not be normally distributed. Thus the result that the bootstrap will work perfectly for IM tests in linear regression models is only for the parametric bootstrap.

Since IM tests for the probit model are only asymptotically pivotal, it is of interest to see how bootstrapping affects their size and power. The results, which are presented in Sections 4, 5, and 6, turn out to be interesting.

4. IM Tests under the Null

Figure 1 shows P value plots for the OPG, DLR, and ES variants of the IM test for the regression case with $n = 100$ and $k = 2$. These are based on an experiment with 100,000 replications. The x_i were chosen as in (3), so that $m = 215$. The figure makes it dramatically clear that the OPG form of the IM test works very badly under the null. In this case, it rejects just under half the time at the nominal 5% level. In contrast, the DLR form seems to work quite well, and the ES form over-rejects in the left tail and under-rejects elsewhere.

Figure 1 illustrates some of the advantages and disadvantages of P value plots. On the one hand, they make it very easy to distinguish tests that work well, such as the DLR form, from tests that work badly, such as the OPG form. Moreover, because they show how a test performs for all nominal sizes, P value plots are particularly useful for tests that both over- and under-reject, such as the ES form. On the other hand, P value plots do not make it easy to see patterns in the behavior of tests that work well. For example, one has to look quite closely at the figure to see that DLR systematically under-rejects for small test sizes.

Another potential disadvantage of P value plots is that they can take up a lot of space. Since, in most cases, we are primarily interested in reasonably small test sizes, it may make sense to truncate the plot at some value of x less than unity. Figure 2 shows two sets of P value plots, both truncated at $x = 0.25$. These plots provide a great deal of information about how the sample size n

and the number of regressors k affect the performance of the OPG form of the IM test. From Figure 2a, we see that their performance improves as n increases but is still unsatisfactory for $n = 1600$. From Figure 2b, we see that their performance deteriorates dramatically as k (and hence the number of degrees of freedom) increases. These figures tell us all we really need to know about how the OPG IM test behaves under the null.

For the DLR variant of the IM test, P value plots are not very informative because the tests perform so well. Figure 3 therefore provides P value discrepancy plots for this test for the same cases as Figure 2, but truncated at $x = 0.40$. It is natural to ask whether P value discrepancies, such as those in Figure 3, can be explained by experimental randomness. The Kolmogorov-Smirnov (KS) test is often used to test whether an EDF is compatible with some specified distribution function. For EDFs of P values, the KS test statistic is simply

$$\max_{j=1,\ldots,N} \left| \hat{F}(p_j) - p_j \right|. \tag{14}$$

This is *almost* equal to the largest absolute value of the P value discrepancy plot,

$$\max_{i=1,\ldots,m} \left| \hat{F}(x_i) - x_i \right|. \tag{15}$$

Note, however, that expression (14) will almost always be slightly bigger than expression (15) because the maximum is being taken over a larger number of points. If we ignore this very slight difference, there is an extremely easy way to see if observed P value discrepancies could be the result of experimental randomness. We simply need to draw horizontal lines representing the critical values for the KS test on our figure. The lines labelled $\text{KS}_{.05}$ in Figure 3 represent the .05 critical values.

Figure 3a makes it clear that the DLR form of the IM test systematically under-rejects for small test sizes when $k = 2$. However, as we see from Figure 3b, things change as k increases. For $n = 200$, the DLR form over-rejects for all test sizes whenever $k \geq 5$. For large values of k and small values of n, the DLR form is sufficiently badly behaved that a P value plot might be more appropriate than a P value discrepancy plot.

Figure 4 shows P value discrepancy plots for the ES form of the IM test. This figure is comparable to Figure 3, but the vertical scale is somewhat different. It is very clear that the ES form over-rejects when the nominal test size is small but under-rejects when the nominal size is large enough. These over- and under-rejections become more pronounced as n falls and k rises, although the effects of reducing n and increasing k are by no means the same. The nominal test size at which the test switches from over-rejection to under-rejection becomes larger as k increases.

We now turn to the results for the probit model. Figure 5 shows that the OPG variant of the IM test rejects far too often under the null. This over-rejection becomes worse as the number of regressors k increases and as the sample size n declines. It also depends on the parameters of the DGP. The figure reports results from two sets of experiments, one with $\beta_0 = 0$ and the

other β_i equal to 1, and the second with all the β_i equal to 1. The performance of the OPG test is always somewhat worse in the latter case. From the figure, we see that the OPG test should never be used without bootstrapping and that, because the test is definitely nonpivotal, bootstrapping cannot be expected to work perfectly.

Figure 6 shows P value discrepancy plots for the ES form of the IM test for probit models. P value plots would also be reasonably informative in this case, but they would be hard to decipher in the left-hand tail, which is the region of greatest interest. To keep the figure readable, only results for the $\beta_0 = 0$ case are shown. The comparable figure for $\beta_0 = 1$, which is not shown, looks similar but not identical to this one. Like Figure 5, this figure is based on 100,000 replications. The striking feature of Figure 6 is that the ES test rejects too often for small test sizes and not nearly often enough for larger sizes. When $k = 3$, tests at the .05 level perform quite well, especially for $n = 50$ and $n = 100$. But the test over-rejects quite severely at the .01 and .001 levels, and its performance at the .05 level actually deteriorates as n increases. Thus reporting results only for the .05 level could be very misleading.

By this point, it should be clear why P value plots and P value discrepancy plots are often much more informative than tables of rejection frequencies at conventional significance levels like .05. First of all, there is nothing special about the .05 level. Some investigators may prefer to use the .01 level or even the .001 level, while pretests are often performed at levels of .25 or even higher. Moreover, if investigators are simply looking at P values rather than testing at some specified level, they will generally place much more weight on a P value of .001 than on a P value of .049. This can be a dangerous thing to do.

Another reason for being concerned with the entire distribution of a test statistic, or at least the left-hand part of it, is that, in our experience, a test statistic that is seriously unreliable at any level for some combination of parameter values and sample size, is likely to be seriously unreliable at the .05 level for some other combination of parameter values and sample size. Consider Figures 4 and 6. Thus serious P value discrepancies at any level should be a cause for concern.

As we discussed above, an alternative graphical technique, which has long been used in the statistics literature, is the QQ plot. Figure 7 shows QQ plots for the same experiments as Figure 4b, that is, for the regression model ES IM test with 200 observations and several values of k, the number of regressors. Since these experiments had 100,000 replications, the plots are based only on the 215 quantiles given in (3). It is instructive to compare Figure 7 with Figure 4b. Although the QQ plots in Figure 7 certainly make it clear that the ES IM test does not perform perfectly, this figure provides much less useful information, and takes up much more space, than does Figure 4b. It is extremely difficult, on the basis of Figure 7, to see how the performance of the ES IM test changes with k, something that is immediately obvious from the P value discrepancy plots of Figure 4b.

Of course, because they use one dimension for nominal test size, P value plots and P value discrepancy plots cannot use that dimension to represent

something else, such as the value of some parameter or the sample size. In consequence, there are bound to be cases in which other types of plots are equally or more informative. There may well be occasions when, like residual plots for regressions, P value plots provide valuable information to investigators but do not convey it succinctly enough to merit publication.

5. Smoothing P Value Discrepancy Plots

Even though Figure 3 is based on 100,000 replications, it is somewhat jagged as a consequence of experimental randomness. When P value discrepancy plots are based on more modest numbers of replications, as will generally be the case when test statistics are expensive to compute, perhaps because they involve bootstrapping, these plots can be quite jagged. Therefore, it is natural to think about how to obtain smoother plots that may be easier to interpret.

One way to smooth a P value discrepancy plot is to regress the discrepancies on smooth functions of x_i, such as polynomials or trigonometric functions. If we let v_i denote $\hat{F}(x_i) - x_i$ and $f_l(x_i)$ denote the l^{th} function of x_i, the first of which may be a constant term, such a regression can be written as

$$v_i = \sum_{l=1}^{L} \gamma_l f_l(x_i) + u_i, \qquad i = 1, \ldots, m. \tag{16}$$

One difficulty is that the u_i are neither homoskedastic nor serially uncorrelated. However, it turns out to be relatively easy to derive a feasible GLS procedure, and this is done in the Appendix.

It is not clear how to choose L and the functions $f_l(x_i)$ that appear in (16). One obvious approach is to use powers of x_i as regressors. Another is to use the functions $\sin(l\pi x_i)$ for $l = 1, 2, 3 \ldots$, and no constant term. The advantage of the latter approach is that $\sin(0) = \sin(l\pi) = 0$, so that the approximation, like z_i itself, is constrained to equal zero at $x_i = 0$ and $x_i = 1$. However, this may not always be an advantage. If a test over-rejects severely, $F_i - x_i$ may be large even for x_i near zero, and it may be hard for a function that equals zero at $x_i = 0$ to fit well with a reasonable number of terms. For a given set of regressors, the choice of L can be made in various ways. We have chosen it to maximize the Akaike Information Criterion, that is, the value of the loglikelihood function minus L; see the Appendix.

Figure 8 illustrates the smoothing procedure we have just described. It shows P value discrepancy functions for the bootstrapped OPG IM test for the regression model with $n = 100$ and $k = 2$, based on 199 bootstrap samples. Because IM tests in linear regression models are pivotal, this test should work perfectly, except for experimental error. The AIC chose the trigonometric model with two regressors, $\sin(\pi x_i)$ and $\sin(2\pi x_i)$, and the resulting smoothed P value discrepancy function is shown in the figure. The figure also shows confidence bands, which were obtained by adding to and subtracting from the $\tilde{F}_i$ twice the square roots of the diagonal elements of the covariance matrix of the fitted values. These confidence bands are very narrow near 0 and 1, because of the constraint

that the trigonometric function must always pass through those points, and they always include the true value of zero.

When using P value discrepancy plots to study the performance of bootstrap tests, it is extremely important to choose B, the number of bootstrap samples, with care. B should be chosen so that $x(B+1)$ is an integer for every x at which the discrepancy will be calculated. See Davidson and MacKinnon (1996b) for an intuitive explanation and references. Since all the curves shown in Figure 8 are plotted at the points $0.005, 0.010, 0.015, \ldots, 0.995$, the smallest number of bootstrap samples that it is appropriate to use is 199.

Using the wrong number of bootstrap samples can produce quite unsatisfactory results. This is illustrated in Figure 9, which is based on the same set of 10,000 replications as Figure 8. The P value discrepancy plots based on $B = 200$ and $B = 100$ are shown along with the one based on $B = 199$. Using just one too many bootstrap samples substantially changes the results: The discrepancies are much more positive than they should be in the left-hand part of the figure. If the inequality in the definition (1) of the EDF had been made strict, the discrepancies would instead have been too negative in the right-hand part of the figure. In either case, the shape of the discrepancy plot will be wrong. Using $B = 100$, as some authors have done in studies of bootstrap testing, produces even worse results. Note that for $B = 100$ the figure uses only the points $0.01, 0.02, \ldots, 0.99$. If the full set of points were used, the plot would exhibit an extreme sawtooth pattern.

Smoothing provides a way to test the null hypothesis that the observed P value discrepancies are entirely due to experimental error. A test that should usually be much more powerful than a KS test is a likelihood ratio test for $\boldsymbol{\gamma} = \mathbf{0}$. Such a test may easily be computed. As an example, for the bootstrapped OPG test with $B = 200$, both the correctly computed KS test, based on (14), and the approximate one, based on (15), are equal to 0.0117. These tests are not significant at the .05 level ($P = .129$). In contrast, the LR test for the smoothing regression with polynomial regressors is 34.00 with 2 degrees of freedom, which is significant at any imaginable level. Note that the LR test statistics for the bootstrapped OPG test with $B = 199$ have P values of .256 for the trigonometric specification and .785 for the polynomial one. Thus the evidence suggests that the bootstrapped OPG test performs exactly the way theory says it should, provided that B is chosen correctly.

We now turn our attention to the bootstrapped IM tests for the probit model. These results are based on 10,000 replications with 399 bootstrap samples per replication. Figure 10 shows smoothed P value discrepancy plots for both forms of the IM test, for $k = 3$, several sample sizes, and the two sets of parameter values used in Figure 5. Several things are evident from this figure. Bootstrapping works remarkably well, but it generally does not work perfectly. The discrepancies between nominal and actual size are reasonably small, much smaller than the discrepancies associated with using asymptotic P values. However, we can reject the hypothesis that these discrepancies are zero in every case but one (the ES form for $n = 100$ and $\beta_0 = 0$). These discrepancies can be of either sign, and they are usually smaller for the larger sample sizes, as we would

expect. Parameter values clearly matter a lot. Because of this, it is unreasonable to expect that bootstrapped IM tests for the probit model will always work as well as they seem to in this example.

6. Size-Power Curves

It is often desirable to compare the power of alternative test statistics, but this can be difficult to do if all the tests do not have the correct size. Suppose we perform a Monte Carlo experiment in which the data are generated by a process belonging to the alternative hypothesis. The test statistics of interest are calculated for each replication, and corresponding P values are obtained. If the EDFs of these P values are plotted, the result will not be very useful, since we will be plotting power against *nominal* test size. Unfortunately, this is what is often implicitly done when test power is reported in a table.

In order to plot power against true size, we need to perform two experiments, preferably using the same sequence of random numbers. In the first experiment, the DGP satisfies the null hypothesis, and in the second it does not. Let the points on the two approximate EDFs be denoted $\hat{F}(x)$ and $\hat{F}^*(x)$, respectively. These are to be evaluated at a prechosen set of points $x_i, i = 1, \ldots, m$. As before, $F(x)$ is the probability of getting a nominal P value less than x under the null. Similarly, $F^*(x)$ is the probability of getting a nominal P value less than x under the alternative. Tracing the locus of points $\big(F(x), F^*(x)\big)$ inside the unit square as x varies from 0 to 1 thus generates a size-power curve on a correct size-adjusted basis. Plotting the points $\big(\hat{F}(x_i), \hat{F}^*(x_i)\big)$, including the points $(0, 0)$ and $(1, 1)$, does exactly the same thing, except for experimental error. By using the same set of random numbers in both experiments, we can reduce experimental error, since the correlation between $\hat{F}(x_i) - F(x_i)$ and $\hat{F}^*(x_i) - F^*(x_i)$ will normally be quite high. The idea of plotting power against true size to obtain a size-power curve is not at all new. When we wrote the first version of this paper, we believed that this very simple method of doing so was new, but in fact it was used by Wilk and Gnanadesikan (1968).

Unless the null is a simple one, there will be an infinite number of DGPs that satisfy it. When the test statistic is pivotal, it does not matter which one we use to generate $\hat{F}(x)$. However, when it is not pivotal, the choice of which DGP to use can matter greatly. In Davidson and MacKinnon (1996b), we argue that a reasonable choice is the *pseudo-true null*, which is the DGP that satisfies the null hypothesis and is as close as possible, in the sense of the Kullback-Leibler Information Criterion, to the DGP used to generate $\hat{F}^*(x)$; see also Horowitz (1994, 1995). In the experiments for the probit model, to be discussed below, we used the pseudo-true null.

Figure 11 shows size-power curves for the OPG, DLR, and ES forms of the IM test for regression models with $k = 2$ for $n = 100$ and $n = 200$. The error terms in the non-null DGP were generated as a mixture of normals with different variances, and they therefore displayed kurtosis. Several results are immediate from this figure. The ES form has the greatest power for a given size of test, followed by the DLR form. The OPG form has far less power than the other

two, and for $n = 100$ it actually has power less than its size for true sizes that are small enough to be interesting. Numerous other experiments, for which results are not shown, produced the same ordering of the three tests. The inferiority of the OPG form was always very marked.

These experimental results make it clear that, if one is going to use an IM test for a linear regression model, the best one to use is the bootstrapped ES form. It must have essentially the correct size (because the test is pivotal), and it will have better power than any of the other tests. This conclusion contradicts that of Davidson and MacKinnon (1992a), who failed to allow for the possibility of bootstrapping and ignored the issue of power. Of course, it might well be better yet to test for heteroskedasticity, skewness, and kurtosis separately. The component pieces of the ES form (5), with P values determined by bootstrapping, could be used for this purpose.

There is one practical problem with drawing size-power curves by plotting $\hat{F}^*(x_i)$ against $\hat{F}(x_i)$. For tests that under- or over-reject severely under the null, there may be a region of the size-power curve that is left out by a choice of values of x_i such as those given in (2) or (3). For instance, suppose that a test over-rejects severely for small sizes, as the OPG IM test does. Then, even if x_i is very small, there may be many replications under the null for which the realized P value is still smaller. As an example, for the OPG test in the regression model with $n = 100$ and $k = 5$, $\hat{F}(.001) = .562$ and $\hat{F}^*(.001) = .397$. Therefore, if the size-power curve were plotted with $x_1 = .001$, there would be a long straight segment extending from $(0, 0)$ to $(.562, .397)$. Such a straight segment would bear clear witness to the gross over-rejection of which the OPG IM test is guilty, but it would also bear witness to a severe lack of detail in the depiction of how the test behaves.

It could well be argued that tests which behave very badly under the null are not of much interest, so that this is not a serious problem. In any case, the problem is not difficult to solve. We simply have to make sure that the x_i include enough very small numbers. Experience suggests that adding the following 18 points to (3) will produce reasonably good results even in extreme cases:

$$x_i = .1 \times 10^{-8}, .2 \times 10^{-8}, .5 \times 10^{-8}, \ldots, .1 \times 10^{-3}, .2 \times 10^{-3}, .5 \times 10^{-3}.$$

Of course, this assumes that small P values are computed accurately, which may not always be the case.

In the remainder of this section, we discuss size-power curves for IM tests in probit models. These turn out to be quite interesting. The data are generated by the model

$$E(y_t \,|\, \Omega_t) = \Phi\left(\frac{\beta_0 + X_{t1} + X_{t2}}{\exp(.5X_{t1} + .5X_{t2})}\right), \tag{17}$$

where $\beta_0 = 0$ or $\beta_0 = 1$. This model is a special case of (13), and it is similar to one of the alternatives used by Horowitz (1994).

We performed six experiments, for $n = 50$, $n = 100$, and $n = 200$, for each of $\beta_0 = 0$ and $\beta_0 = 1$. We also ran a few additional experiments with $k = 2$ instead

of $k = 3$ to verify that the number of regressors did not affect the principal conclusions. These experiments used 10,000 replications with $B = 799$; the value of B was relatively large in order to ensure that power loss from bootstrapping would be minimal. It is not possible to graph the results for all six experiments in one figure, but Figure 12 does show the principal results. The ES test with asymptotic P values performs very much better than the asymptotic OPG test when $\beta_0 = 1$. However, when $\beta_0 = 0$, it performs a bit better when $n = 100$ and a bit less well when $n = 200$.

Bootstrapping often has a substantial effect on the performance of the OPG test, but it generally has little effect on the performance of the ES test. The power of both tests is sometimes increased and sometimes reduced. Both these results are consistent with the theoretical results of Davidson and MacKinnon (1996b), since, as we saw in Figure 10, the bootstrapped OPG test is much farther from being pivotal than the bootstrapped ES test. Because the asymptotic tests, especially the OPG form, are emphatically not pivotal, the size-power curves for those tests in Figure 12 might have been very different if we had used a different null DGP to generate $\hat{F}(x)$. It is safer, but still not entirely safe, to compare the size-power curves for the bootstrap tests.

From these results and those discussed earlier, we conclude that the ES form of the IM test with bootstrap P values is the procedure of choice for probit models, just as it is for regression models. It never has substantially less power than the bootstrap OPG form, and it sometimes has substantially more. The bootstrap P values are much more accurate than the asymptotic ones, especially when they are very small: Compare Figures 5 and 6 with Figure 10. However, Figure 10 makes it clear that bootstrapping will not produce entirely accurate test sizes for small samples.

7. Conclusion

Monte Carlo experiments are a valuable tool for obtaining information about the properties of specification testing procedures in finite samples. However, the rich detail in the results they provide can be difficult to apprehend if presented in the usual tabular form. In this paper, we have discussed several graphical techniques that can make the principal results of an experiment immediately obvious. All of these techniques rely on the construction of an estimated c.d.f. (EDF) of the nominal P values associated with some test statistic. From these, we can easily obtain a variety of diagrams, namely, P value plots, P value discrepancy plots (which may optionally be smoothed), and size-power curves.

We have illustrated these techniques by presenting the results of a number of experiments concerning alternative forms of the information matrix test. We believe that these results, which are entirely presented in graphical form, are of considerable interest and provide more information in a more easily assimilated fashion than a tabular presentation could possibly have done.

Appendix

This appendix discusses GLS estimation of equation (16). If the regression function in (16) were chosen correctly, the error term u_i would be equal to $\hat{F}(x_i) - F(x_i)$. It can be shown that, for any two points x and x' in the $(0,1)$ interval,

$$\begin{aligned} \mathrm{Var}(\hat{F}(x)) &= N^{-1}F(1-F), \text{ and} \\ \mathrm{Cov}(\hat{F}(x), \hat{F}(x')) &= N^{-1}\big(\min(F,F') - FF'\big). \end{aligned} \tag{A.1}$$

Here $F \equiv F(x)$ and $F' \equiv F(x')$. The first line of (A.1) is just a restatement of the well-known result about the variance of the mean of N Bernoulli trials. Expression (A.1) makes it clear that $\boldsymbol{\Omega}$, the $m \times m$ covariance matrix of the u_i in (16), exhibits a moderate amount of heteroskedasticity and a great deal of serial correlation. Both the standard deviation of u_i and the correlation between u_i and u_{i-1} are greatest when $F_i = 0.5$ and decline as F_i approaches 0 or 1.

Equation (16) can be rewritten using matrix notation as

$$\boldsymbol{v} = \boldsymbol{Z}\boldsymbol{\gamma} + \boldsymbol{u}, \quad E(\boldsymbol{u}\boldsymbol{u}') = \boldsymbol{\Omega}, \tag{A.2}$$

where $\boldsymbol{Z}$ is an $n \times L$ matrix, the columns of which are the regressors in (16). The GLS estimator of $\boldsymbol{\gamma}$ is

$$\tilde{\boldsymbol{\gamma}} = \big(\boldsymbol{Z}'\boldsymbol{\Omega}^{-1}\boldsymbol{Z}\big)^{-1}\boldsymbol{Z}'\boldsymbol{\Omega}^{-1}\boldsymbol{v}. \tag{A.3}$$

The smoothed discrepancies are the fitted values $\boldsymbol{Z}\tilde{\boldsymbol{\gamma}}$ from (A.2), and the covariance matrix of these fitted values is

$$\boldsymbol{Z}\big(\boldsymbol{Z}'\boldsymbol{\Omega}^{-1}\boldsymbol{Z}\big)^{-1}\boldsymbol{Z}'. \tag{A.4}$$

Confidence bands can be constructed by using the square roots of the diagonal elements of (A.4).

It is easily verified that the inverse of $\boldsymbol{\Omega}$ has non-zero entries only on the principal diagonal and the two adjacent diagonals. Specifically,

$$\begin{aligned} \boldsymbol{\Omega}^{-1}_{ii} &= N(F_{i+1} - F_i)^{-1} + N(F_i - F_{i-1})^{-1}, \\ \boldsymbol{\Omega}^{-1}_{i,i+1} &= -N(F_{i+1} - F_i)^{-1}, \\ \boldsymbol{\Omega}^{-1}_{i,i-1} &= -N(F_i - F_{i-1})^{-1}, \end{aligned} \tag{A.5}$$

for $i = 1, \ldots, m$, and $\boldsymbol{\Omega}^{-1}_{i,j} = 0$ for $|i-j| > 1$. In these formulae, $F_0 = 0$ and $F_{m+1} = 1$. In contrast to many familiar examples of GLS estimation, it is neither easy nor necessary to triangularize $\boldsymbol{\Omega}^{-1}$ in this case. Because $\boldsymbol{\Omega}^{-1}$ has non-zero elements only along three diagonals, it is not difficult to compute $\boldsymbol{Z}'\boldsymbol{\Omega}^{-1}\boldsymbol{Z}$ and $\boldsymbol{Z}'\boldsymbol{\Omega}^{-1}\boldsymbol{v}$ directly. For example, the lj^{th} element of $\boldsymbol{Z}'\boldsymbol{\Omega}^{-1}\boldsymbol{Z}$ is

$$\sum_{i=1}^{m} Z_{il}Z_{ij}\boldsymbol{\Omega}^{-1}_{ii} + \sum_{i=2}^{m} Z_{i-1,l}Z_{ij}\boldsymbol{\Omega}^{-1}_{i,i-1} + \sum_{i=1}^{m-1} Z_{i+1,l}Z_{ij}\boldsymbol{\Omega}^{-1}_{i,i+1}, \tag{A.6}$$

where the needed elements of $\boldsymbol{\Omega}^{-1}$ were defined in (A.5). Thus, by using (A.6), it is straightforward to compute the GLS estimates (A.3).

True GLS estimation is not feasible here. If the test statistic being studied is well-behaved, however, $F(x_i)$ will be close to x_i for all x_i, and it will be reasonable to use x_i instead of the unknown F_i in (A.5). This will yield approximate GLS estimates. If the test statistic is not so well-behaved, it is natural to use a two-stage procedure. In the first stage, the approximate GLS estimates are obtained. In the second stage, the unknown F_i in (A.5) are replaced by $\acute{F}_i \equiv x_i + \acute{z}_i$, where $\acute{z}_i$ denotes the fitted values from the approximate GLS procedure. Note that the estimates of both F_i and $F_i - F_{i-1}$ must be positive for the GLS procedure to work. Since these two conditions may not always be satisfied by the $\acute{F}_i$, it may be necessary to modify them slightly before computing the feasible GLS estimates and the final estimates $\tilde{F}_i$.

The loglikelihood function associated with GLS estimation of (A.2) is

$$-\frac{m}{2}\log(2\pi) + \frac{1}{2}|\boldsymbol{\Omega}^{-1}| - \frac{1}{2}(\boldsymbol{v} - \boldsymbol{Z}\boldsymbol{\gamma})'\boldsymbol{\Omega}^{-1}(\boldsymbol{v} - \boldsymbol{Z}\boldsymbol{\gamma}). \qquad (A.7)$$

The determinant of $\boldsymbol{\Omega}^{-1}$ which appears in (A.7) is

$$N^m\left(\prod_{i=0}^{m} a_i\right)\sum_{j=0}^{m} 1/a_j, \qquad (A.8)$$

where $a_i \equiv (F_{i+1} - F_i)^{-1}$, for $i = 0, \ldots, m$, and, as before, $F_0 = 0$ and $F_{m+1} = 1$.

It is important to make sure that (16) fits satisfactorily, as it may not if $\boldsymbol{Z}$ has been chosen poorly. One simple approach is to calculate the GLS equivalent of the regression standard error:

$$s \equiv \left(\frac{1}{n-L}\tilde{\boldsymbol{u}}'\boldsymbol{\Omega}^{-1}\tilde{\boldsymbol{u}}\right)^{1/2},$$

where $\tilde{\boldsymbol{u}}$ is the vector of (feasible) GLS residuals from estimation of (16). If $\boldsymbol{Z}$ has been specified correctly, s should be approximately equal to unity.

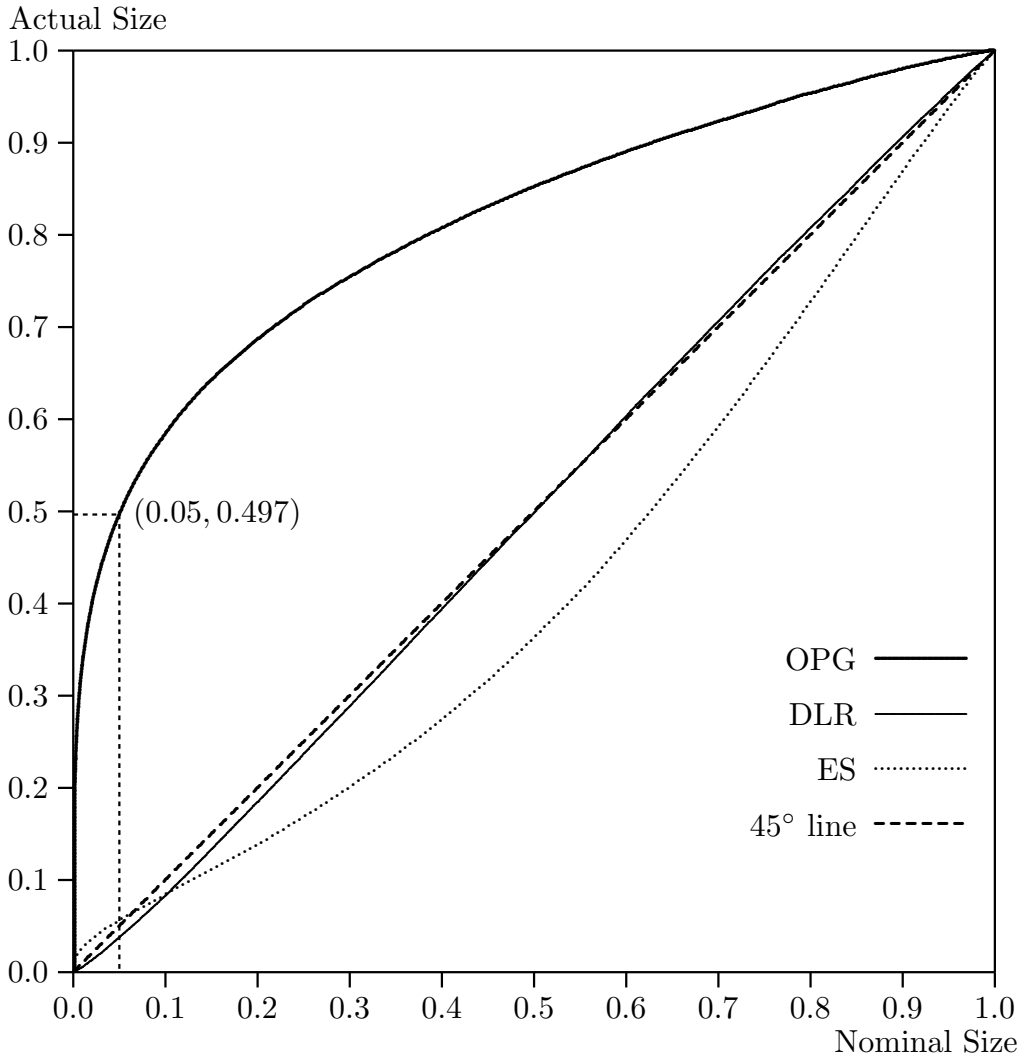


Figure 1. *P* value plots for IM tests, regression model, $n = 100$, $k = 2$

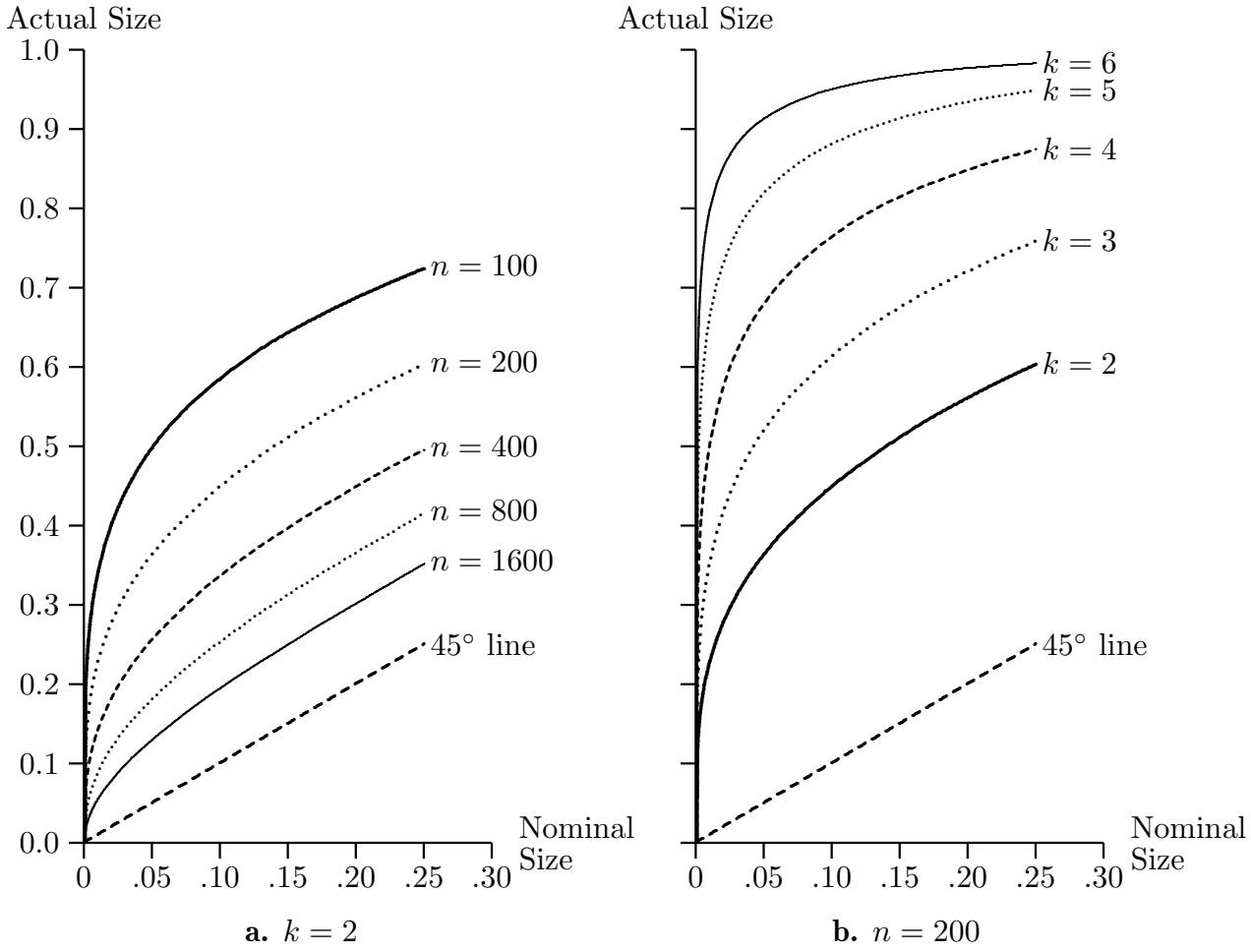
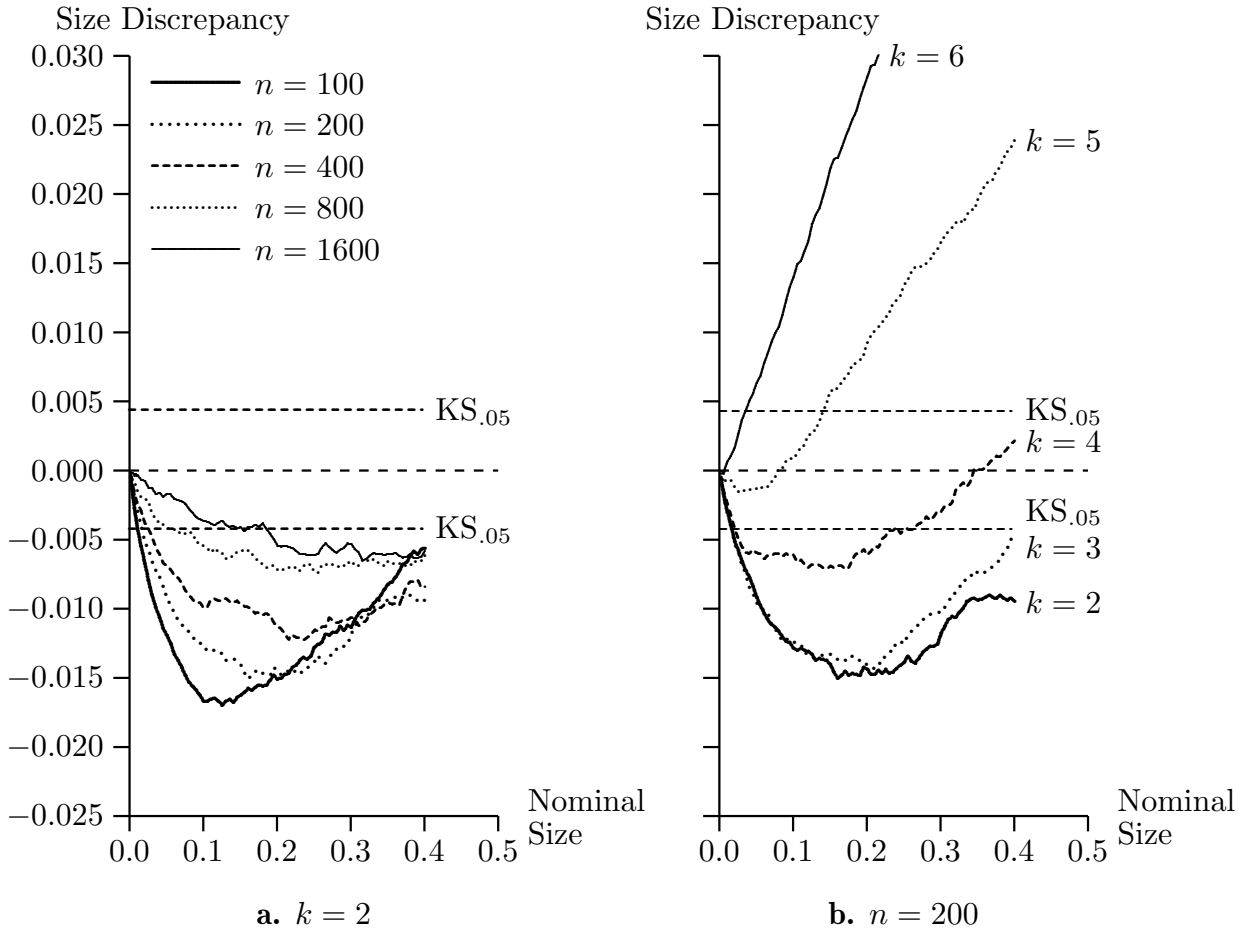


Figure 2. *P* value plots for OPG IM tests, regression model



a. $k = 2$ **b.** $n = 200$

Figure 3. *P* value discrepancy plots for DLR IM tests, regression model

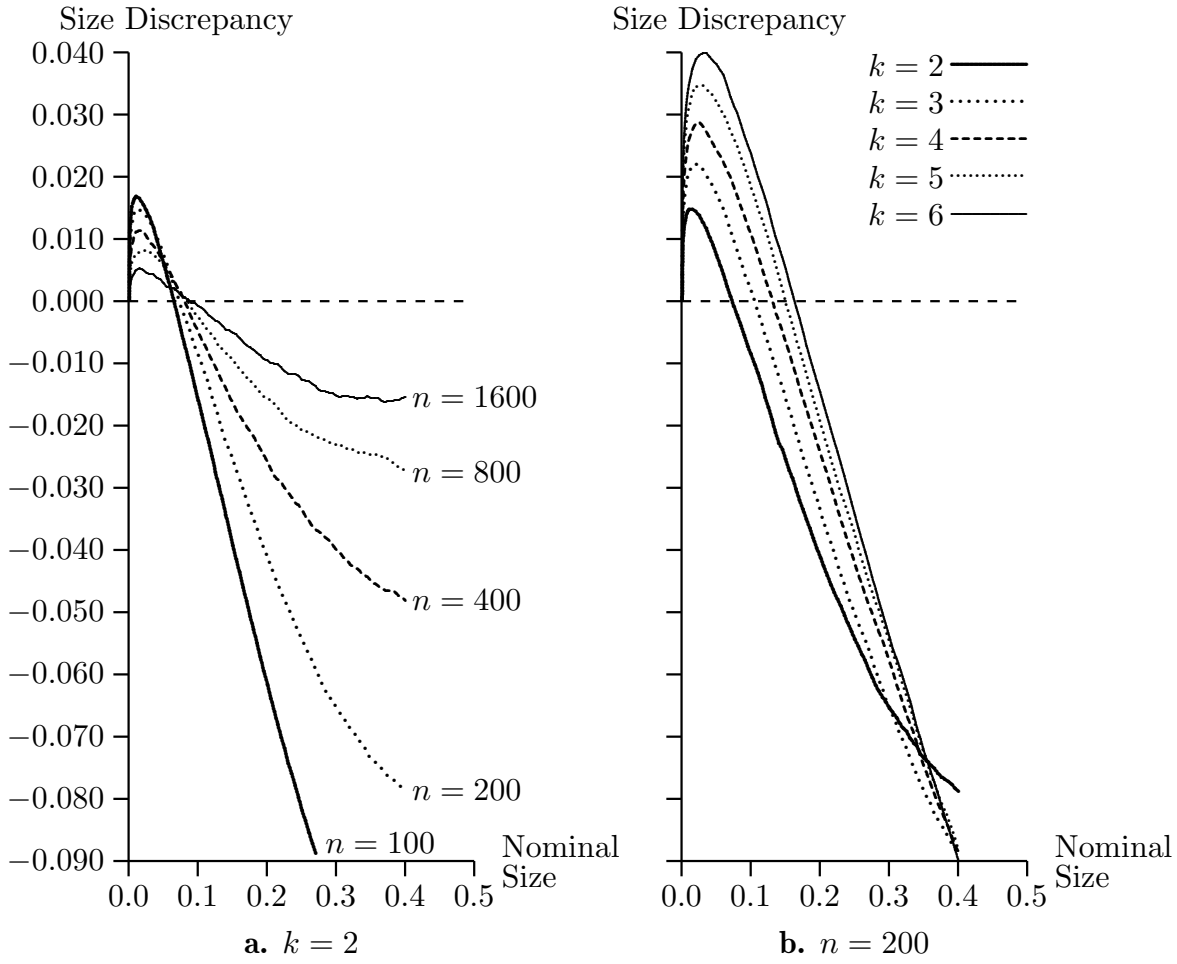
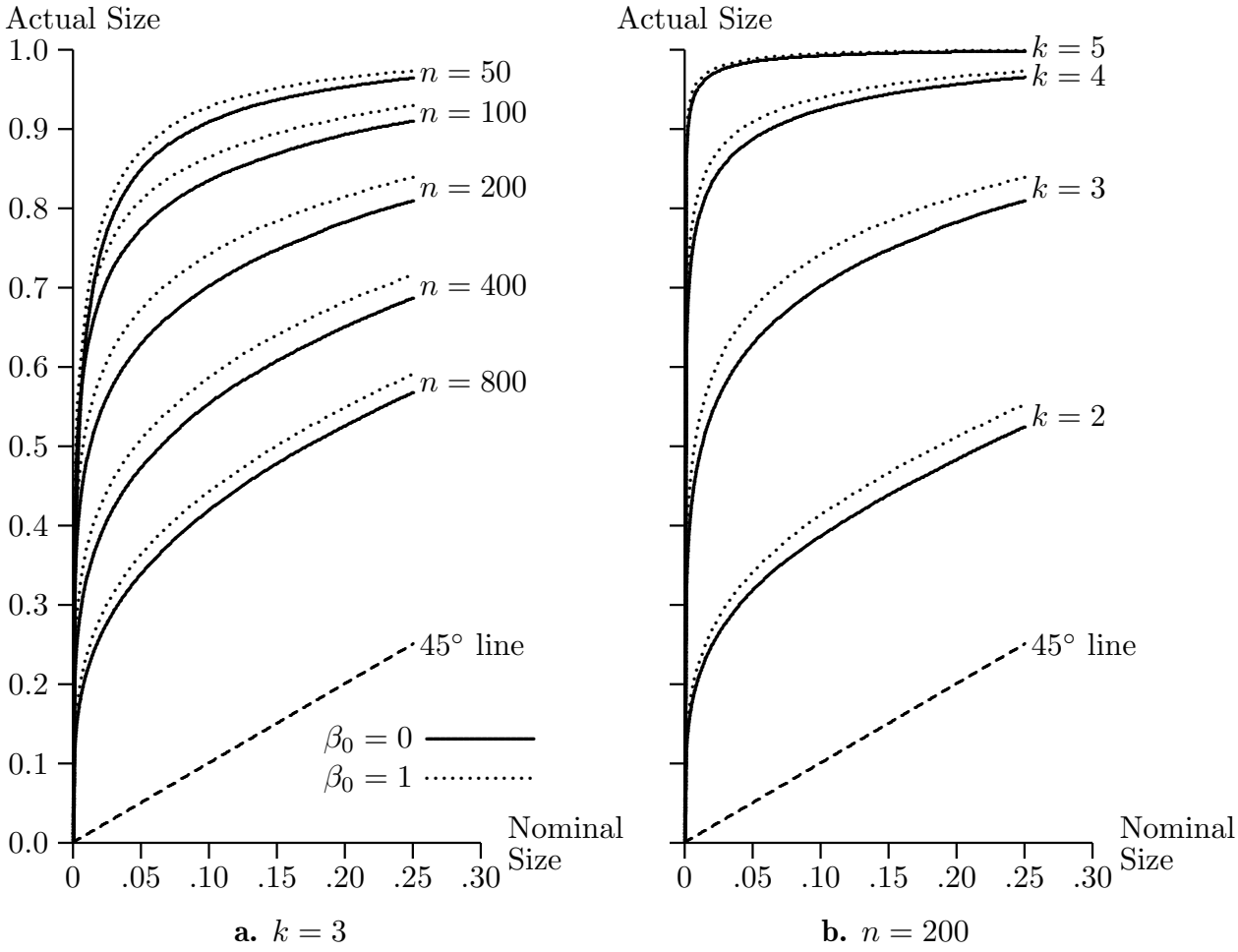


Figure 4. *P* value discrepancy plots for ES IM tests, regression model



a. $k = 3$ **b.** $n = 200$

Figure 5. *P* value plots for OPG IM tests, probit model

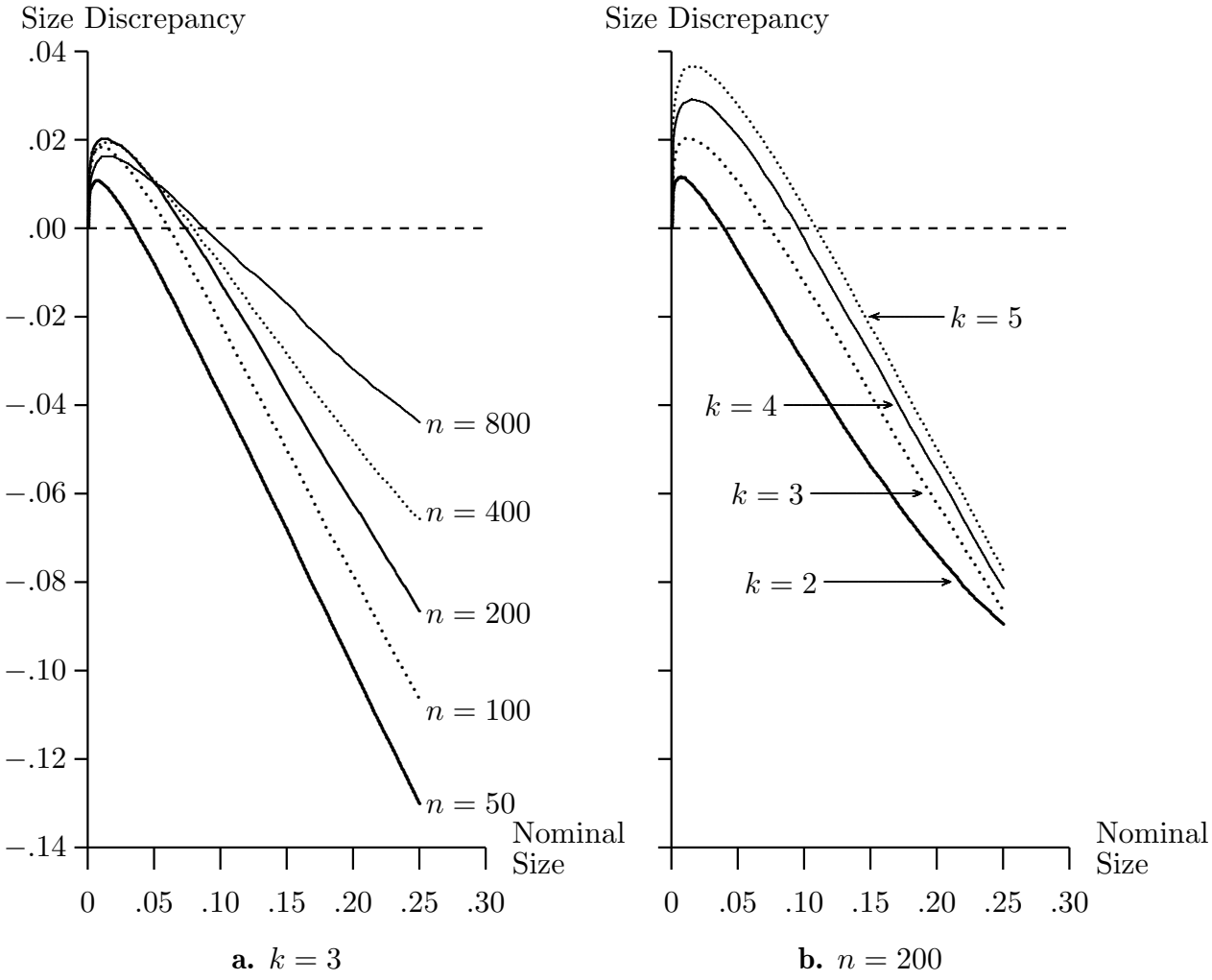


Figure 6. *P* value discrepancy plots for ES IM tests, probit model

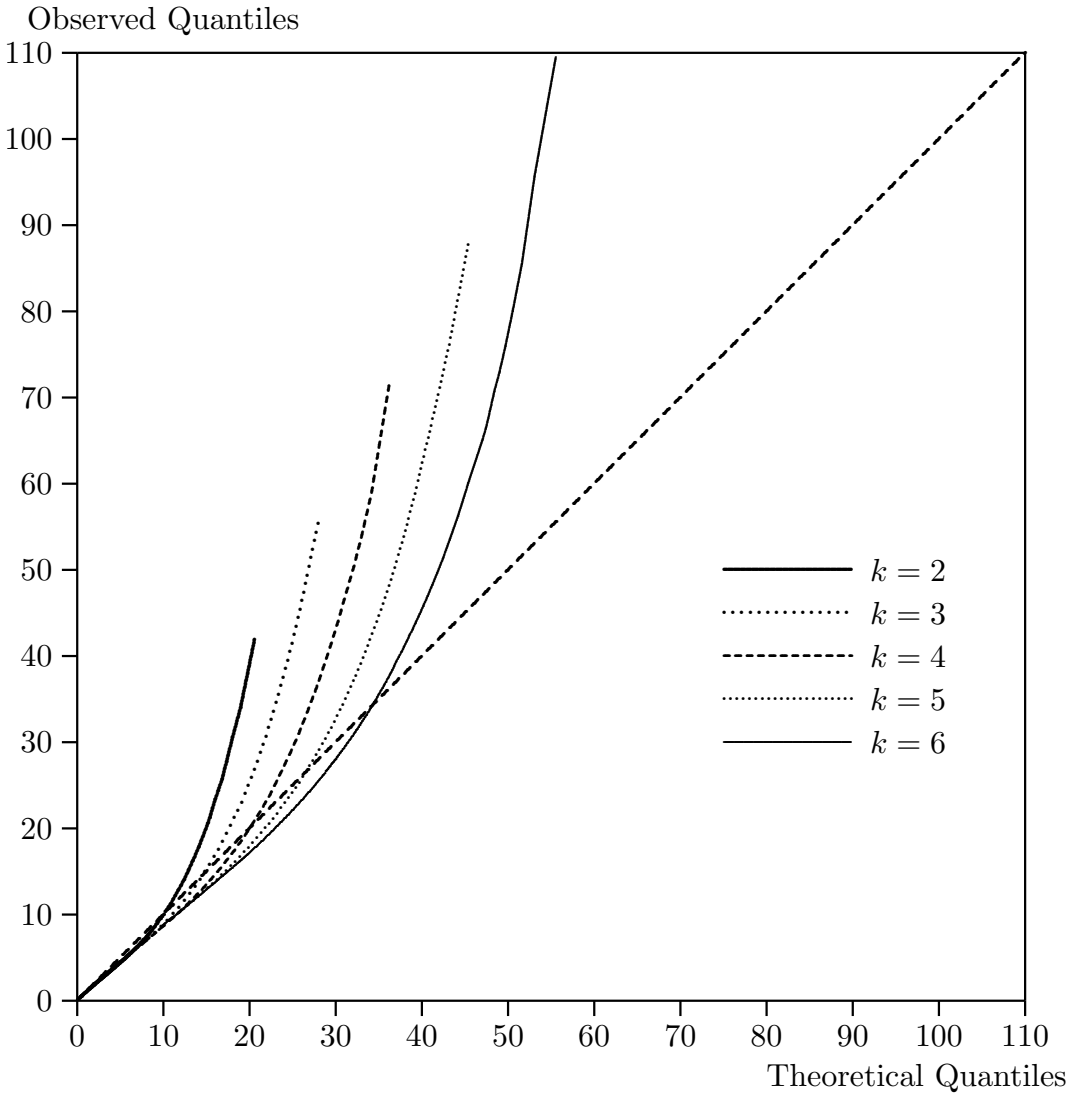


Figure 7. QQ plots for ES IM tests, regression model, $n = 200$

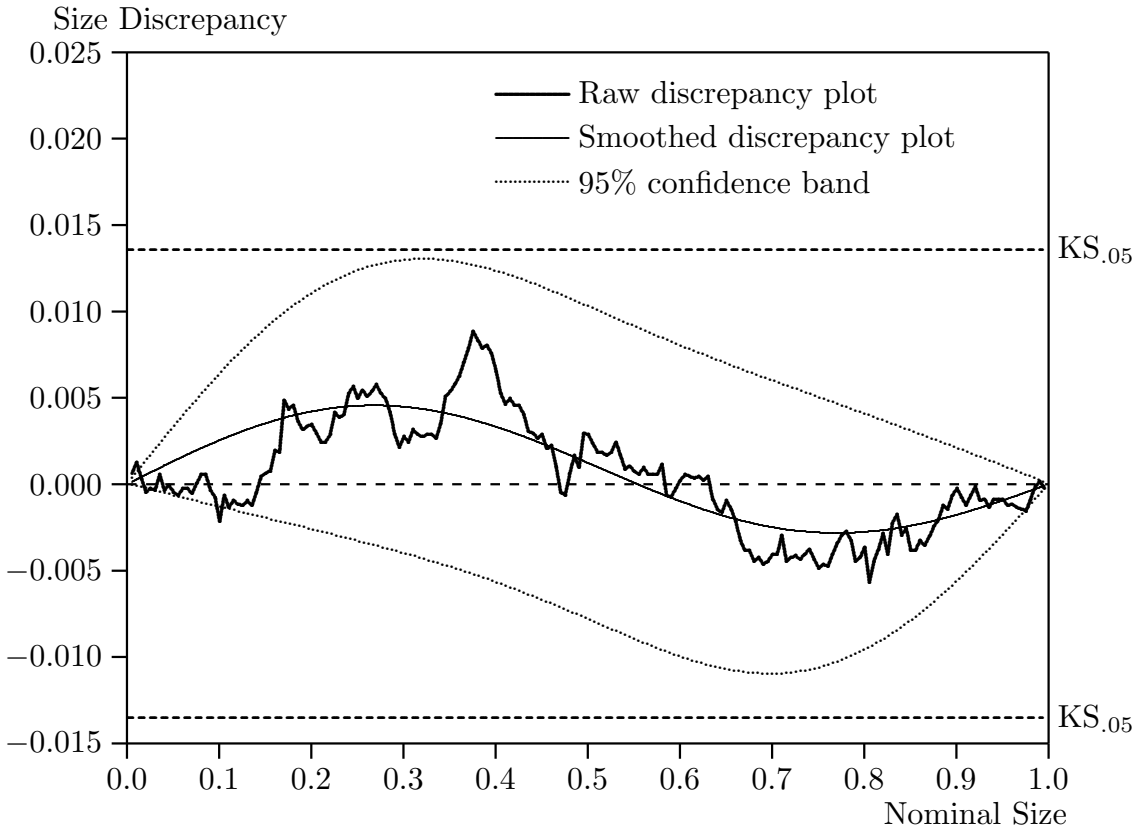


Figure 8. *P* value discrepancy plots for bootstrap OPG IM test, regression model, $n = 100$, $k = 2$

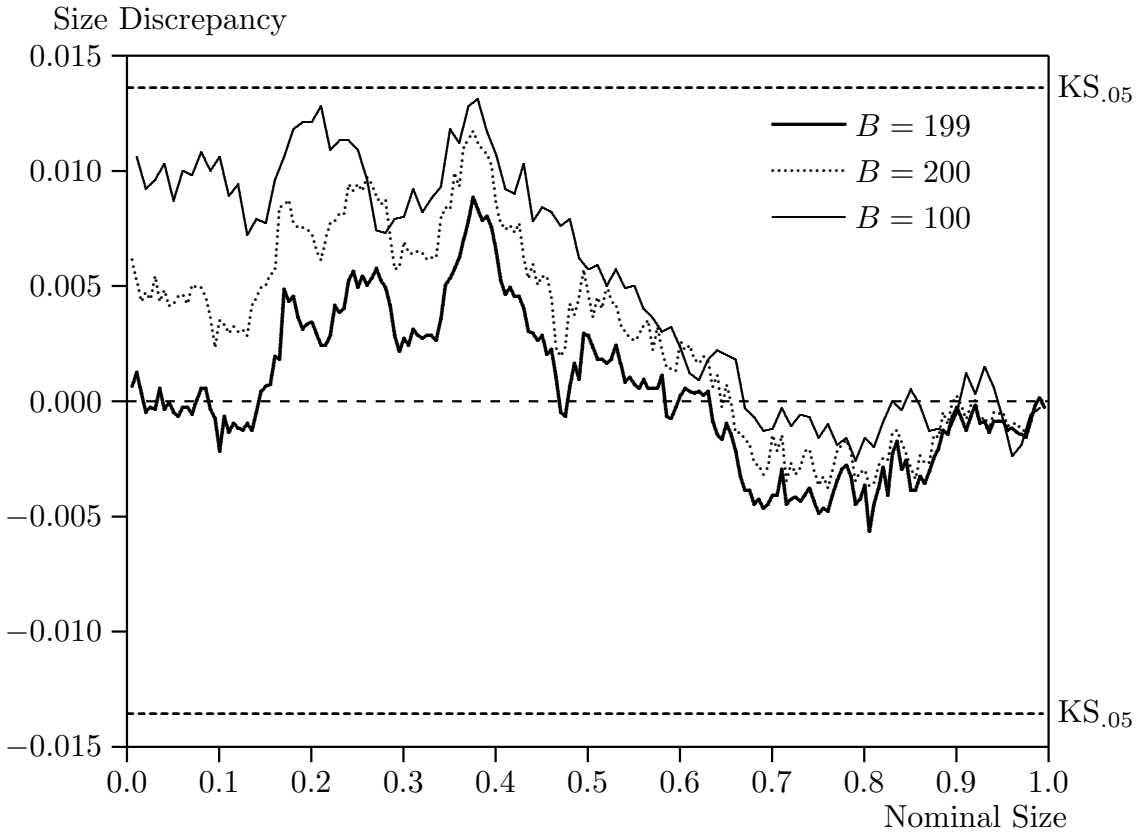


Figure 9. *P* value discrepancy plots for bootstrap OPG IM test, regression model, $n = 100$, $k = 2$

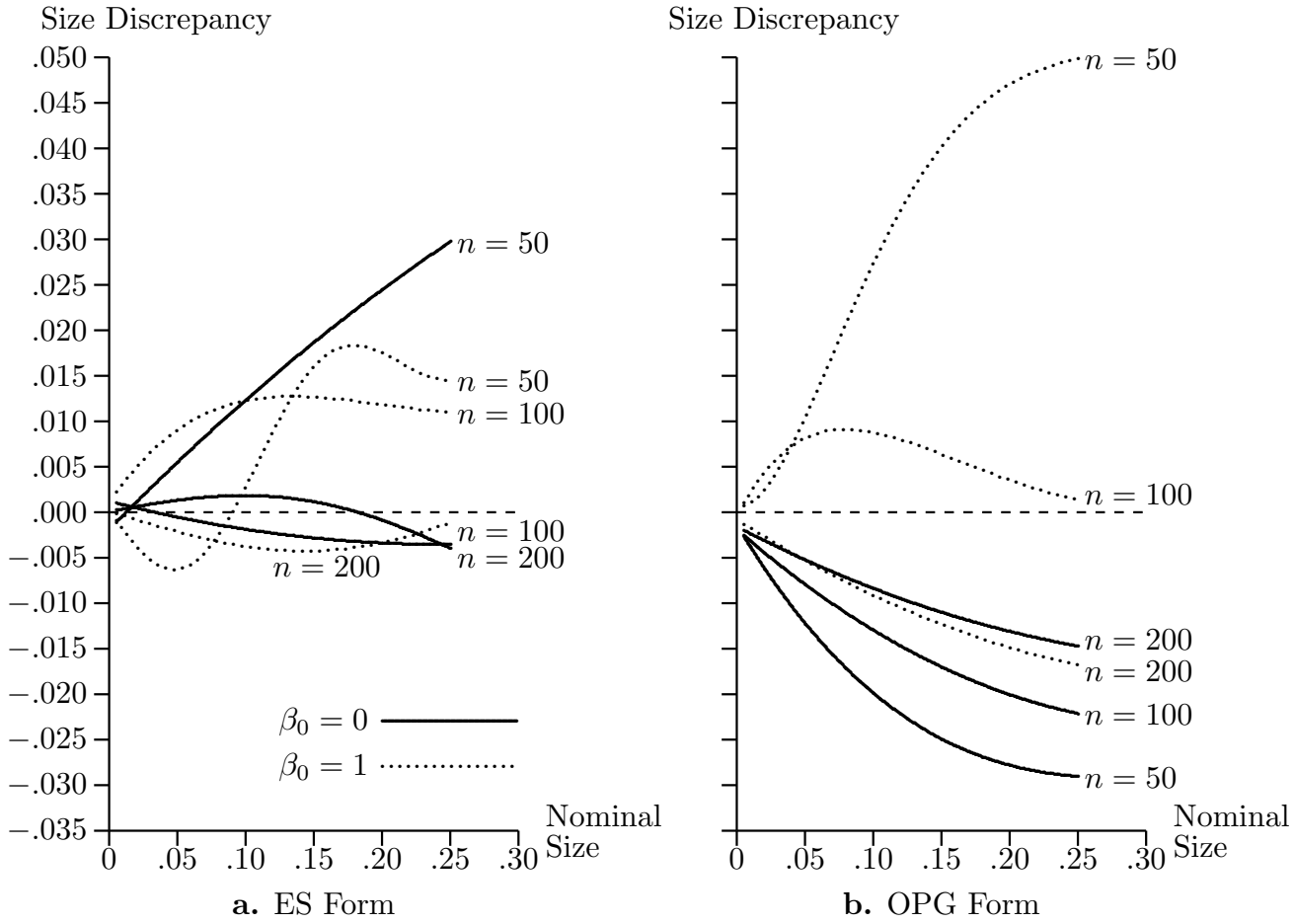


Figure 10. Smoothed *P* value discrepancy plots, bootstrap tests, probit, $k = 3$

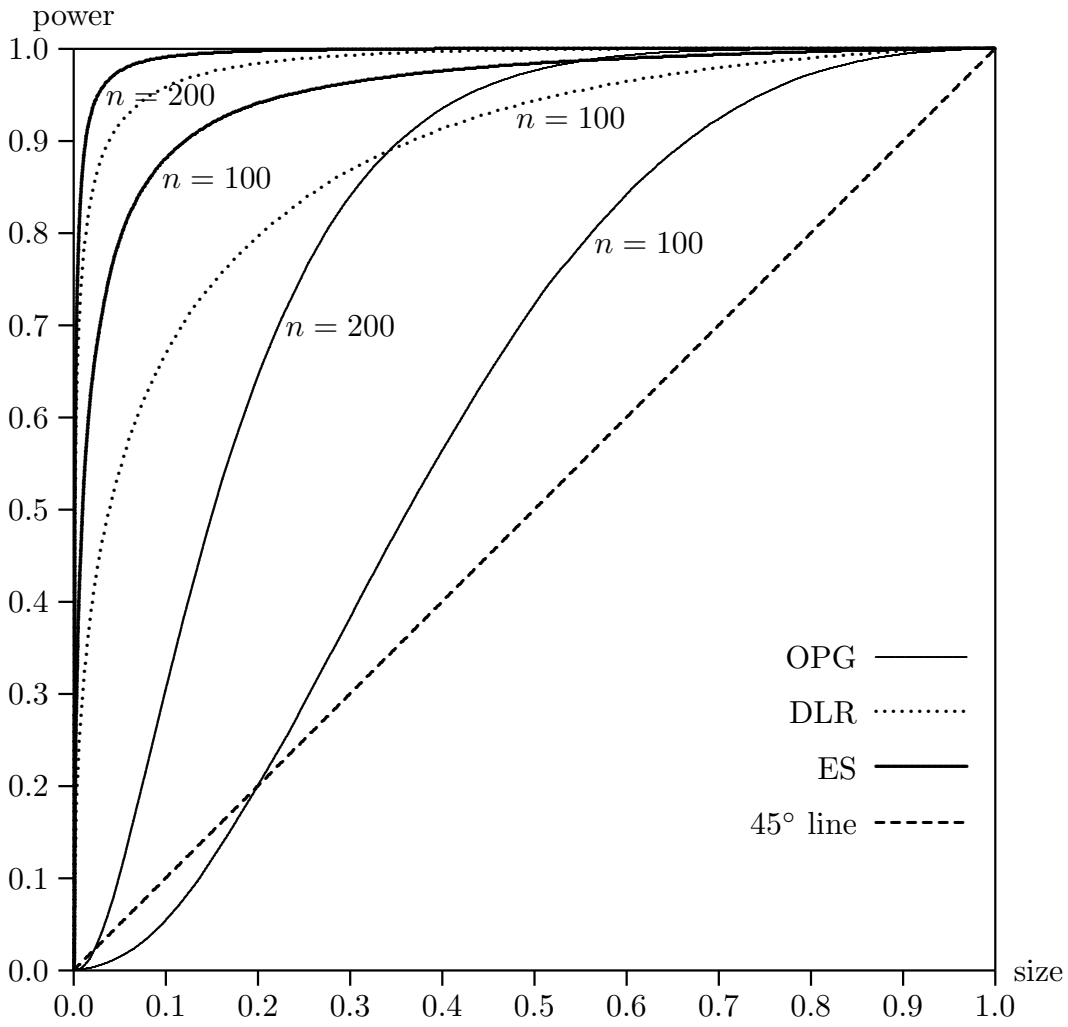
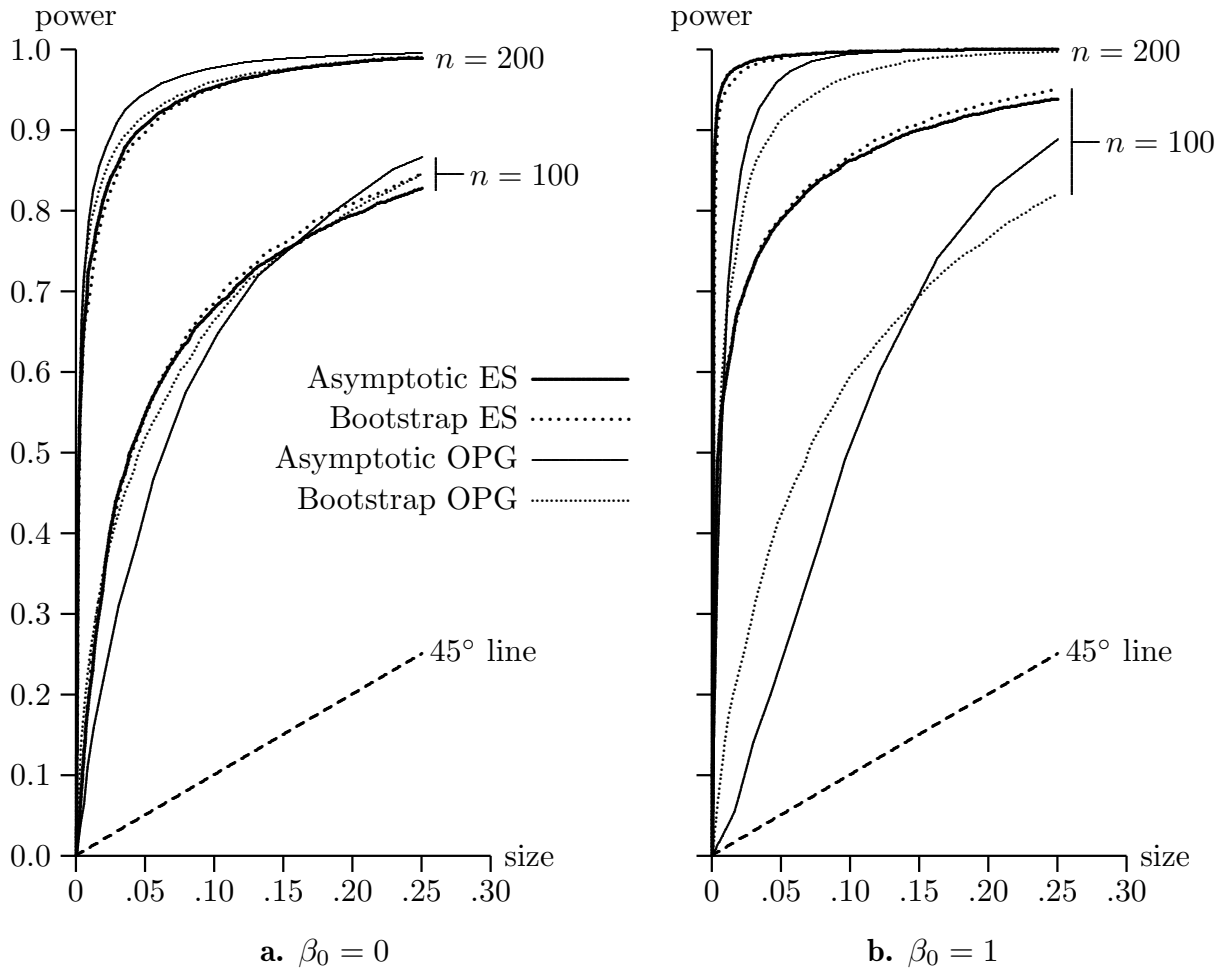


Figure 11. Size-power curves, regression model, kurtosis, $k = 2$



a. $\beta_0 = 0$ **b.** $\beta_0 = 1$

Figure 12. Size-power curves for probit IM tests, $k = 3$