Queen's Economics Department Working Paper No. 903

Graphical Methods for Investigating the Size and Power of Hypothesis Tests

Russell Davidson James G. MacKinnon

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

6-1994

Discussion Paper #903

Graphical Methods for Investigating the Size and Power of Hypothesis Tests

by

Russell Davidson
Queen's University
and GREQE-EHESS

and

James G. MacKinnon
Queen's University

June 1994

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

GREQE–EHESS
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 are used to illustrate these figures. Tests based on the OPG regression are found to perform poorly in terms of both size and power.

This research was supported, in part, by the Social Sciences and Humanities Research Council of Canada. We are grateful to participants in workshops at Cornell University, York University, and GREQE for helpful comments.

June, 1994

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. Even so, little effort seems to have been devoted to devising good ways of presenting such results. This paper discusses some simple graphical methods that appear to be very useful for characterizing both 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 2), 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, and variance reduction techniques may be used to improve the efficiency with which the quantities of interest are estimated; see Davidson and MacKinnon (1992b). For simplicity of notation, we will ignore these possibilities here.

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. 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. There are many ways to present graphically the sort of information that is usually presented in tabular form. The ones we advocate here are easy to implement and yield graphs that are easy to interpret.

All of the graphs we discuss are based on the empirical distribution function, or EDF, of the P values of the τ_j's. The P value of τ_j is the probability of observing a value of τ 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's 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. In order to conserve storage space (since N will often be very large), we choose to evaluate the EDF (1) only at m points $x_i, i = 1, \ldots, m$, which should be chosen in advance. The x_i's must be chosen 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's should be chosen. A quite parsimonious set of x_i's is

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

Another choice which should 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 5, it may be necessary to add additional points in certain cases.

The simplest graph that we will discuss is simply a plot of $\hat{F}(x_i)$ against x_i. We shall refer to such plots as ***P* value plots**. If the distribution of τ used to compute the p_j's is correct, each of the p_j's 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 3, 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 all test statistics that behave approximately the way they should will look roughly like 45° lines, P value 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 these graphs as ***P* value discrepancy plots**. 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 4, 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 very useful for dealing with test power. In Section 5, 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's instead of the EDF of their P values. However, 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.

Another fairly standard graphical approach, at least in the statistics literature, is to use what are called **quantile-quantile plots** or **QQ plots**; see, for example, Chesher and Spady (1991). In such a plot, the empirical quantiles of the τ_j's 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. Our approach has several advantages over using QQ plots. For a QQ plot, there is no natural scale for the axes: If the hypothesized distribution changes, so will that scale. This makes it impossible to plot on the same axes test statistics which have different distributions under the null. It is also much more difficult to interpret a QQ plot than it is to interpret a P value plot when the plot does not lie on the 45° line, since there is no way to see how actual test size is related to nominal test size.

Of course, graphical methods by themselves are not always enough. When test performance depends on a number of factors, the two-dimensional nature of both graphs and 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).

In order to illustrate and motivate these procedures, we use them to present the results of a study of the properties of alternative forms of the information matrix (IM) test proposed by White (1982). We compare tests based on the OPG regression, which was proposed as a way to compute IM tests by Chesher (1983) and Lancaster (1984), with other forms of the IM test. We find that the OPG variant performs relatively poorly in terms of both size and power. The result that the OPG form of the IM test frequently tends to over-reject in finite samples is not new: See, among others, Taylor

(1987), Chesher and Spady (1991), and Davidson and MacKinnon (1992a). However, the result that, for a given size, OPG IM tests have lower power than other forms of the IM test does not appear to be well-known.

The plan of the paper is as follows. In the next section, we briefly discuss several forms of the IM test statistic. In Section 3, we present a number of Monte Carlo results to illustrate the use of P value plots and P value discrepancy plots. In Section 4, we discuss and illustrate methods for smoothing P value discrepancy plots. Finally, in Section 5, we discuss and illustrate the use of size-power curves.

2. Alternative Forms of the Information Matrix Test

In Davidson and MacKinnon (1992a), we derived a new variant of the IM test and presented a number of Monte Carlo results comparing it with two other variants. Table 1 of that paper, which occupies a page and half, is a particularly striking example of the disadvantages of presenting simulation results for test statistics in a non-graphical way. In this paper, we extend the experiments of the earlier paper and present the results graphically.

We shall deal with three variants of the IM test: the OPG variant, the DLR variant, and the efficient score variant. These are derived for the linear regression model:

$$y_t = \beta_1 + \sum_{i=2}^{k} \beta_i X_{ti} + u_t, \qquad u_t \sim \text{NID}(0, \sigma^2). \tag{4}$$

The regressors X_{ti} are normal random variables, independent across observations, and equicorrelated with correlation coefficient one-half. All versions of the IM test are independent of the specific values of the β_i and σ^2, and so those values 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}'s. 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 $\chi^2\left(\frac{1}{2}(k^2 + 3k)\right)$.

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 test, which is not as widely available as the other two but is available for linear regression models, is the efficient score, or ES, form of the IM test. 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:

$$\frac{1}{2}\boldsymbol{h}_2^\top \boldsymbol{Z}\big(\boldsymbol{Z}^\top\boldsymbol{Z}\big)^{-1}\boldsymbol{Z}^\top\boldsymbol{h}_2 + \frac{1}{6}\boldsymbol{h}_3^\top \boldsymbol{X}\big(\boldsymbol{X}^\top\boldsymbol{X}\big)^{-1}\boldsymbol{X}^\top\boldsymbol{h}_3 + \frac{1}{24n}\sum_{t=1}^{n}\big(\tilde{e}_t^4 - 3\big)^2, \quad (5)$$

where $\boldsymbol{Z}$ is a matrix with 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).

It is well-known that the OPG form of the LM statistic has very poor finite-sample properties under the null. One obvious way to improve these is to obtain P values by bootstrapping instead of by using the test's asymptotic distribution. This idea was investigated by Horowitz (1994), who found that it worked very well. The methodology is as follows. After the OPG test statistic, say τ, has been obtained, B sets of simulated data are generated and B bootstrap test statistics are computed. Suppose that B^* of these test statistics are greater than τ. Then the estimate of $p(\tau)$ is B^*/B. In our experiments, we used $B = 1000$. Note that the error terms for the bootstrap samples must be obtained from a normal distribution rather than by resampling from the residuals, as is commonly done, since the residuals will not be normally distributed, and the IM test is sensitive to non-normality.

In the case of the linear regression model (4), the finite-sample distributions of all forms of the IM test statistic do not depend on the unknown parameters of the regression function. Such test statistics are said to be **pivotal**. This implies that, as $B \to \infty$, the EDF of the bootstrapped P values must tend to the 45° line. Therefore, in this case, we do not really need to do a Monte Carlo experiment to see how the bootstrap works. Most test statistics, including most variants of the IM test, do not share this special property, however; see Horowitz (1994).

In the remainder of the paper, we present results for four variants of the IM test for the model (4): the OPG form, the DLR form, the ES form, and what we call the bootstrap OPG form, which is the OPG form with P values computed by the bootstrapping technique just described. Of course, bootstrap P values could also be computed for the DLR and ES forms of

the IM test, but we did not do this. As with the OPG form, bootstrapping these variants of the IM test must yield tests with exactly the right size, except for sampling error in the bootstrap procedure.

3. *P* Value Plots and *P* Value Discrepancy Plots

Figure 1 shows P value plots for the OPG, DLR, and ES variants of the IM test for the case $n = 100, k = 2$. These are based on an experiment with 5000 replications. The x_i's were chosen as in (3), so that $m = 215$. The figure makes it dramatically clear that the OPG form works very badly under the null. In this case, it rejects almost half the time at the nominal 5% level. In contrast, the DLR form seems to work quite well, and the ES form works reasonably well in the left tail but tends to under-reject elsewhere. Results for the bootstrap OPG form are not shown, because they would have been almost indistinguishable from the 45° line.

Figure 1 illustrates some of the advantages and disadvantages of P value plots. On the one hand, these plots make it very easy to distinguish tests that work well, such as DLR, from tests that work badly, such as OPG. On the other hand, they 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 and over-rejects for larger test sizes.

Another disadvantage of P value plots is that they can take up a lot of space. Since we are primarily interested in reasonably small test sizes, it makes 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 $x = 0.4$. 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 size improves as n increases but is still quite unsatisfactory for $n = 1000$. From Figure 2b, we see that the performance of these tests deteriorates dramatically as k (and hence the number of degrees of freedom) increases. These figures tell us all we really need to know about the performance of the OPG IM test under the null.

For the DLR and bootstrap OPG forms of the IM test, P value plots are not very informative because the tests perform so well. Therefore, Figure 3 shows P value discrepancy plots for these two tests for the same case as Figure 1 (namely, $n = 100$ and $k = 2$). From this figure, it is clear that the bootstrap OPG test performs just about the way it should: The discrepancies between $\hat{F}(x_i)$ and x_i are small, change sign often, and can easily be explained by experimental randomness. On the other hand, the DLR test, although it performs quite well, seems systematically to under-reject in the left-hand part of the figure and over-reject elsewhere.

It is natural to ask whether the discrepancies 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

$$\max_{j=1,\ldots,N} \left|\hat{F}(p_j) - p_j\right|. \tag{6}$$

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

$$\max_{i=1,\ldots,m} \left|\hat{F}(x_i) - x_i\right|. \tag{7}$$

However, expression (6) will almost always be slightly bigger than expression (7) because the maximum is being taken over a larger number of points.

As an example, in the case of Figure 3, the correctly computed KS tests, based on (6), are .0046 for bootstrap OPG and .0187 for DLR, while the approximate ones, based on (7), are .0044 and .0184. These tests are highly insignificant in the case of bootstrap OPG ($p > .99$) and marginally significant in the case of DLR ($p = .06$). The fact that (7) is available can be convenient if one decides to compute a KS test after an experiment has been completed. Because both formulae yield essentially the same answers, there is generally no great harm in using (7) rather than (6).

Figure 4 looks at several other cases. Figure 4a makes it clear that the tendency for DLR systematically to under-reject for small test sizes when $k = 2$ is not an artifact of experimental error. However, as we see from Figure 4b, things change as k increases. For $n = 200$, DLR over-rejects for all test sizes whenever $k \geq 5$. For large values of k and small values of n, the DLR test is sufficiently badly behaved that a P value plot might be more appropriate than a P value discrepancy plot.

Figures 3 and 4 illustrate one serious problem with P value discrepancy plots: They tend to be quite jagged, reflecting experimental error. This is true even when N is much larger than the value of 5000 used for these experiments. Therefore, it is natural to think about how to obtain smoother plots that may be easier to interpret. This is the topic of the next section.

4. Smoothing P Value Discrepancy Plots

One natural 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

$$v_i = \sum_{l=1}^{L} \gamma_l f_l(x_i) + u_i, \qquad i = 1, \ldots, m. \tag{8}$$

There are two difficulties with this approach. One is that the u_i's are neither homoskedastic nor serially uncorrelated. However, it turns out to be relatively easy to derive a feasible GLS procedure. Another problem is how to choose L and the functions $f_l(x_i)$. There are several ways to do this, and our experience suggests that there is no single best way.

If the regression function in (8) 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,

$$\begin{aligned} \mathrm{Var}(\hat{F}(x)) &= N^{-1} F(1 - F), \text{ and} \\ \mathrm{Cov}(\hat{F}(x), \hat{F}(x')) &= N^{-1}\big(\min(F, F') - FF'\big). \end{aligned} \tag{9}$$

Here $F \equiv F(x)$ and $F' \equiv F(x')$. Notice that the first line of (9) is just a restatement of the well-known result about the variance of the mean of N Bernoulli trials.

Expression (9) makes it clear that the $m \times m$ covariance matrix of the u_i's in (8), which we shall call $\boldsymbol{\Omega}$, exhibits a moderate amount of heteroskedasticity and a great deal of serial correlation. The standard deviation of u_i is greatest when $F_i = 0.5$ and declines as F_i approaches 0 or 1. For example, suppose that $N = 5000$. In this case, the standard deviation of u_i^2 is 0.0071 for $F_i = 0.5$ and 0.0031 for $F_i = 0.05$. The correlation between u_i and u_{i-1} is also greatest when $F_i = 0.5$. For example, if $F_i = 0.5$ and $F_{i-1} = 0.49$, the correlation between u_i and u_{i-1} is 0.9802; if $F_i = 0.05$ and $F_{i=1} = 0.04$, that correlation is 0.8898.

Equation (8) can be rewritten using matrix notation as

$$\boldsymbol{v} = \boldsymbol{Z}\boldsymbol{\gamma} + \boldsymbol{u}, \quad E(\boldsymbol{u}\boldsymbol{u}^\top) = \boldsymbol{\Omega}, \tag{10}$$

where $\boldsymbol{Z}$ is an $n \times L$ matrix, the columns of which are the regressors in (8). The GLS estimator of $\boldsymbol{\gamma}$ is

$$\tilde{\boldsymbol{\gamma}} = \big(\boldsymbol{Z}^\top \boldsymbol{\Omega}^{-1} \boldsymbol{Z}\big)^{-1} \boldsymbol{Z}^\top \boldsymbol{\Omega}^{-1} \boldsymbol{v}. \tag{11}$$

The smoothed discrepancies are the fitted values $\boldsymbol{Z}\tilde{\gamma}$ from (10), and the covariance matrix of these fitted values is

$$\boldsymbol{Z}(\boldsymbol{Z}^{\top}\boldsymbol{\Omega}^{-1}\boldsymbol{Z})^{-1}\boldsymbol{Z}^{\top}. \tag{12}$$

Confidence bands can be constructed by using the square roots of the diagonal elements of (12).

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,

$$\begin{aligned}
\boldsymbol{\Omega}^{-1}_{ii} &= N(F_{i+1} - F_i)^{-1} + N(F_i - F_{i-1})^{-1}, \\
\boldsymbol{\Omega}^{-1}_{i,i+1} &= -N(F_{i+1} - F_i)^{-1}, \\
\boldsymbol{\Omega}^{-1}_{i,i-1} &= -N(F_i - F_{i-1})^{-1},
\end{aligned} \tag{13}$$

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 $\boldsymbol{Z}^{\top}\boldsymbol{\Omega}^{-1}\boldsymbol{Z}$ and $\boldsymbol{Z}^{\top}\boldsymbol{\Omega}^{-1}\boldsymbol{v}$ directly. For example, the lj^{th} element of $\boldsymbol{Z}^{\top}\boldsymbol{\Omega}^{-1}\boldsymbol{Z}$ is

$$\sum_{i=1}^{m} Z_{il}Z_{ij}\boldsymbol{\Omega}^{-1}_{ii} + \sum_{i=2}^{m} Z_{i-1,l}Z_{ij}\boldsymbol{\Omega}^{-1}_{i,i-1} + \sum_{i=1}^{m-1} Z_{i+1,l}Z_{ij}\boldsymbol{\Omega}^{-1}_{i,i+1}, \tag{14}$$

where the needed elements of $\boldsymbol{\Omega}^{-1}$ were defined in (13). Thus, by using (14), it is straightforward to compute the GLS estimates (11).

As is generally the case, 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 (13). 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's in (13) 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 whatever values are used to estimate the F_i's must be positive, as must be the estimates of $F_i - F_{i-1}$. Since these conditions may not always be satisfied by the $\acute{F}_i$'s, it may be necessary to modify them slightly before computing the feasible GLS estimates and the final estimates $\tilde{F}_i$.

For completeness, we note that the determinant of $\boldsymbol{\Omega}^{-1}$ is

$$N^m\Big(\prod_{i=0}^{m} a_i \sum_{j=0}^{m} 1/a_j\Big), \tag{15}$$

where $a_i \equiv (F_{i+1} - F_i)^{-1}$, for $i = 0, \ldots, m$, and, as before, $F_0 = 0$ and $F_{m+1} = 1$. Expression (15) is needed if we wish to compute the value of the loglikelihood function associated with GLS estimation of (10).

We have not yet said anything about how to specify the regression function in (8), that is, the matrix $\boldsymbol{Z}$. 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 (i.e., the value of the loglikelihood function minus L). It is important to make sure that (8) 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:

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

where $\tilde{\boldsymbol{u}}$ is the vector of (feasible) GLS estimates from (10). If $\boldsymbol{Z}$ has been specified correctly, s should be approximately equal to unity.

Figure 5 illustrates the smoothing procedure we have just described for the case of the DLR test with $n = 200$ and $k = 5$. In this case, only four regressors—$\sin(l\pi x_i)$ for $l = 1, \ldots, 4$—were needed to fit as well as we would expect ($s = 0.98$). A polynomial approximation with two more regressors worked equally well. The confidence bands were obtained by adding to and subtracting from the $\tilde{F}_i$'s twice the square roots of the diagonal elements of (12). Because of the scale of the vertical axis, these bands appear to be quite wide. What is more interesting is that they are distinctly wider in the far left-hand part of the figure than in the far right-hand part. This is because the DLR test is tending to over-reject everywhere. For example, $\tilde{F}(.05) = 0.0797$ and $\tilde{F}(.95) = 0.9622$. This means that $\tilde{F}_i(1 - \tilde{F}_i)$, to which the variance of $\hat{F}_i$ is assumed to be proportional, is equal to 0.0733 in the former case and 0.0364 in the latter. The fitted values $\tilde{F}_i$ naturally tend to be more precisely estimated where the variance of $\hat{F}_i$ is smaller.

Figure 6 shows two sets of truncated, smoothed P value discrepancy plots for the ES form of the IM test. Because $\hat{F}(.001)$ was always substantially greater than zero, the trigonometric approximations did not work at all well, and the smoothing was done using polynomial approximations. It is clear that the ES form over-rejects for tests at the conventional .05 level but under-rejects when the nominal size is large enough. These tendencies

become more pronounced as n falls and k rises, although the effects of reducing n and increasing k are by no means the same. In order to keep the figure readable, confidence bands are not shown. Since the standard error of $\tilde{F}(.05)$ was between .0032 and .0039, the basic shape of these smoothed curves is certainly reliable, but one should not take every wiggle seriously. The main advantage of smoothing here is that it makes the figures much easier to read.

5. 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 done, except that, in most cases, only a few points on the EDF are 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 null hypothesis holds, 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. As before, these are to be evaluated at a prechosen set of points $x_i, i = 1, \ldots, m$. As we have seen, $F(x)$ 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 of course 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; see, for example, Davidson and MacKinnon (1993, Chapter 12), who call them size-power tradeoff curves. However, the method of doing so using EDFs of P values that we have just proposed does appear to be new. It is also remarkably simple.

Figure 7 shows size-power curves for the OPG, bootstrap OPG, DLR, and ES forms of the IM test, for $n = 100$ and $k = 2$. The error terms in the non-null data generating process were generated as a mixture of normals

with different variances, and 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 it actually has power less than its size for true sizes that are small enough to be interesting. As a consequence of the fact that IM tests for linear models are pivotal, bootstrapping the OPG form has no effect on its size-power curve. This theoretical result is illustrated in the figure.

There is one potentially serious 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 with $n = 100$ and $k = 5$, $\hat{F}(.001) = .576$ and $\hat{F}^*(.001) = .405$. Therefore, if the size-power curve were plotted with $x_1 = .001$, there would be a long straight segment extending from $(0, 0)$ to $(.576, .405)$. 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's include enough very small numbers. Experience suggests that adding the following 15 points to (3) will produce reasonably good results even in extreme cases:

$$x_i = .1 \times 10^{-7}, .2 \times 10^{-7}, .5 \times 10^{-7}, \ldots, .1 \times 10^{-3}, .2 \times 10^{-3}, .5 \times 10^{-3}.$$

Of course, N (the number of replications) also matters. Figure 7 is based on just 5000 replications, while Figures 8 and 9 are based on 100,000. The curves in the latter figures are a good deal smoother than those in the former.

Figure 8 shows truncated size-power curves for the OPG IM test. The data under the alternative were generated by the same process as for Figure 7. From Figure 8a, we see that as n becomes larger, the performance of the OPG IM test improves. From Figure 8b, we see that the performance of the test deteriorates dramatically as k increases. Figure 9 shows truncated size-power curves for the ES and DLR forms of the IM test, for the same cases as Figure 8b. It is clear that ES has substantially more power than DLR, and that both of them have dramatically more power

than OPG. Several other experiments were run in which the null hypothesis was false in other ways. The results were always qualitatively similar to those portrayed in Figures 7 through 9.

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 would have essentially the correct size (because the test is pivotal), and it would have better power than any of the other tests. Of course, if the objective were not simply to see whether inferences based on the usual information matrix are reliable, which is what the IM test is designed to do, it might well be better 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.

6. 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 proposed 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. These results, which are entirely presented in graphical form, provide far more information about these tests than the tabular results which are typically presented; it may be instructive to compare them with those in Davidson and MacKinnon (1992a).

References

Chesher, A. (1983). "The information matrix test: simplified calculation via a score test interpretation," *Economics Letters*, **13**, 45–48.

Chesher, A., and R. Spady (1991). "Asymptotic expansions of the information matrix test statistic," *Econometrica*, **59**, 787–815.

Davidson, R., and J. G. MacKinnon (1984). "Model specification tests based on artificial linear regressions," *International Economic Review*, **25**, 485–502.

Davidson, R., and J. G. MacKinnon (1992a). "A new form of the information matrix test," *Econometrica*, 60, 145–157.

Davidson, R., and J. G. MacKinnon (1992b). "Regression-based methods for using control variates in Monte Carlo experiments," *Journal of Econometrics*, 54, 203–222.

Davidson, R. and J. G. MacKinnon, *Estimation and Inference in Econometrics*, Oxford University Press, 1993.

Hall, A. (1987). "The information matrix test for the linear model," *Review of Economic Studies*, **54**, 257–63.

Hendry, D. F. (1984). "Monte Carlo experimentation in econometrics," Ch. 16 in *Handbook of Econometrics*, Vol. II, eds. Z. Griliches and M. D. Intriligator, Amsterdam, North-Holland.

Horowitz, J. L. (1994). "Bootstrap-based critical values for the information matrix test," *Journal of Econometrics*, 61, 395–411.

Lancaster, T. (1984). "The covariance matrix of the information matrix test," *Econometrica*, **52**, 1051–53.

Taylor, L. W. (1987). "The size bias of White's information matrix test," *Economics Letters*, **24**, 63–67.

White, H. (1982). "Maximum likelihood estimation of misspecified models," *Econometrica*, **50**, 1–26.

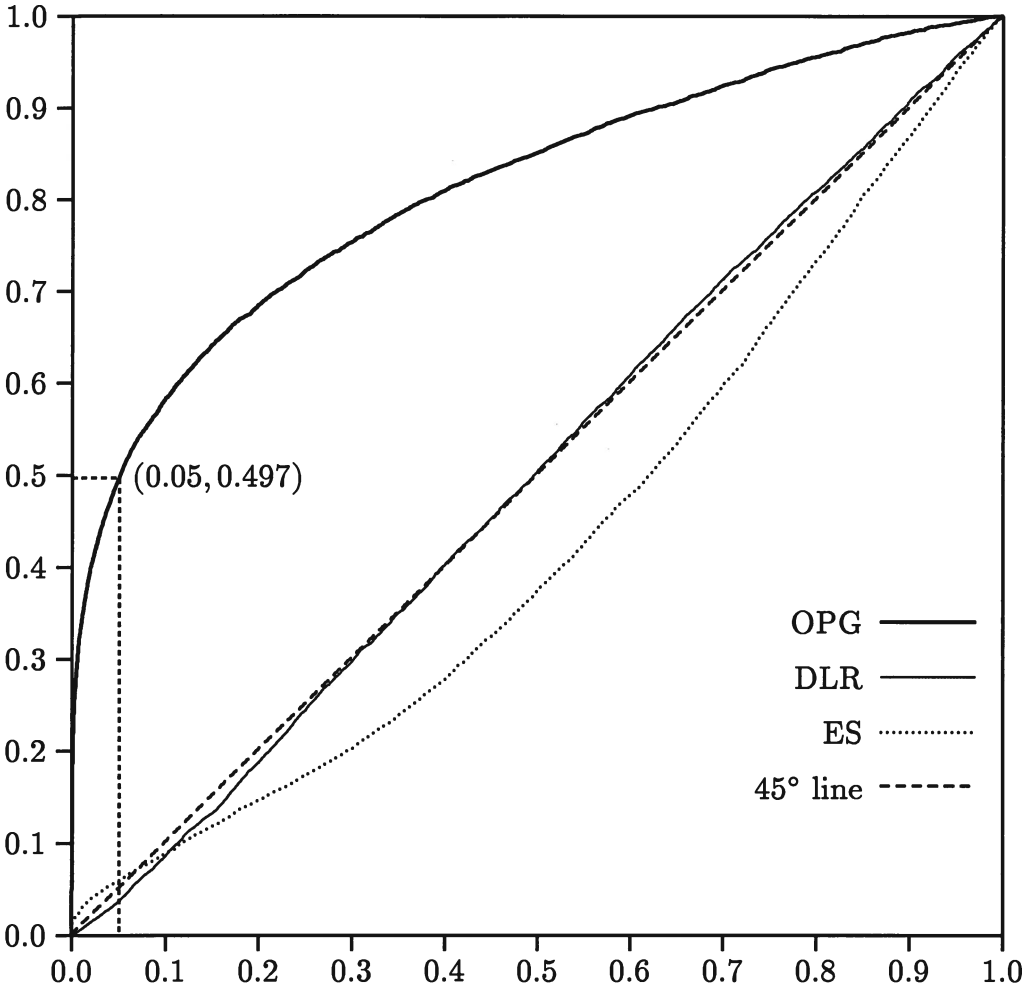
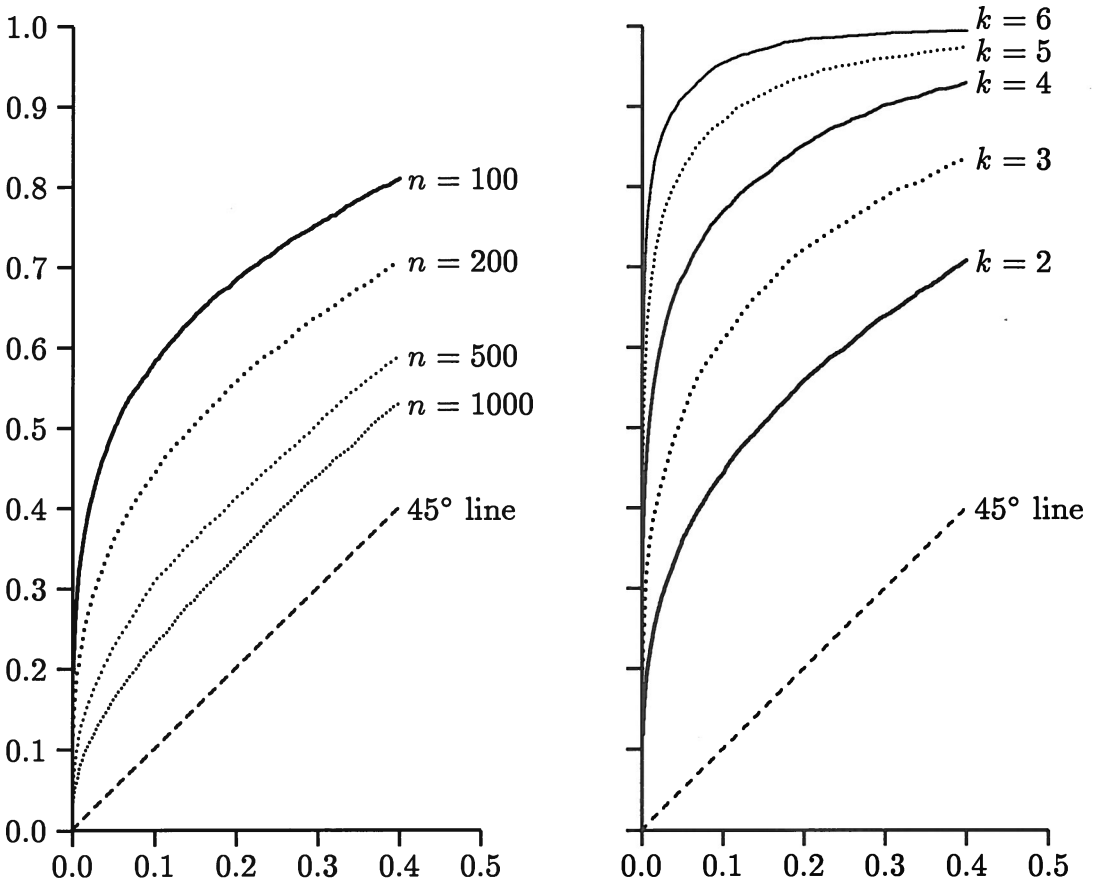


Figure 1. P value plots for IM tests, $n = 100$, $k = 2$



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

Figure 2. ***P*** **value plots for OPG IM tests**

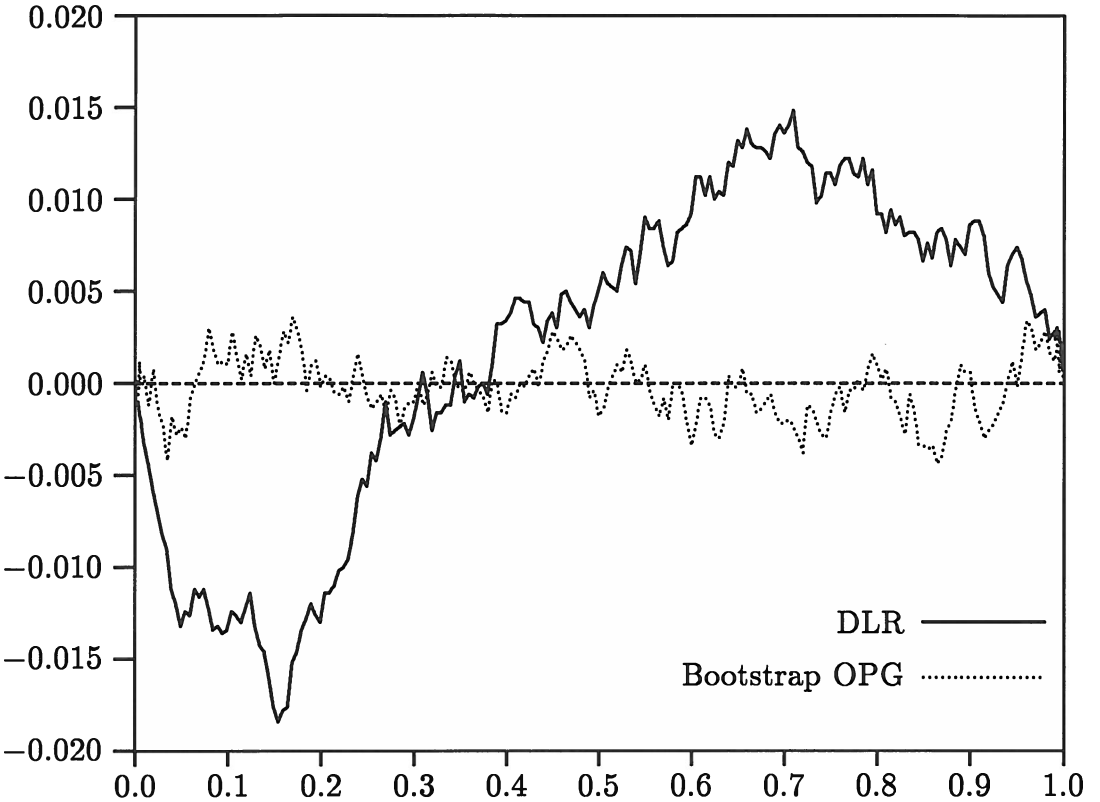


Figure 3. P **value discrepancy plots for IM tests,** $n = 100$, $k = 2$

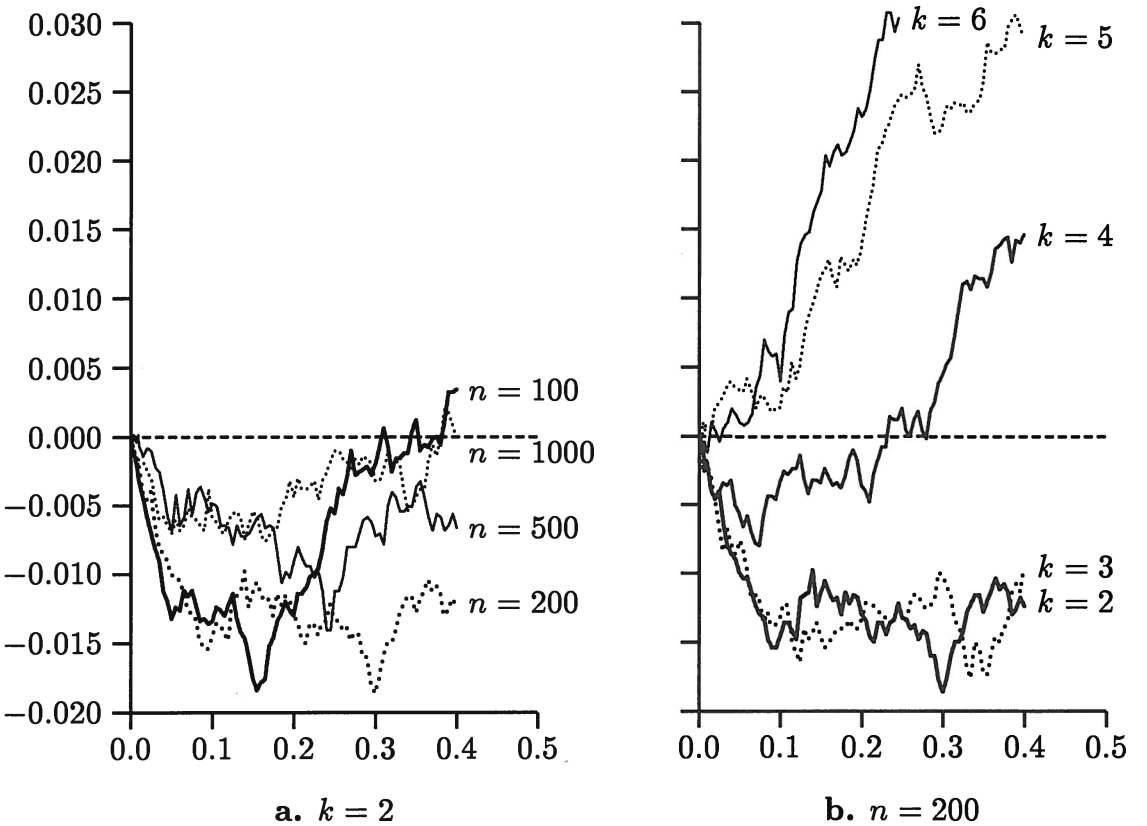


Figure 4. ***P*** **value discrepancy plots for DLR IM tests**

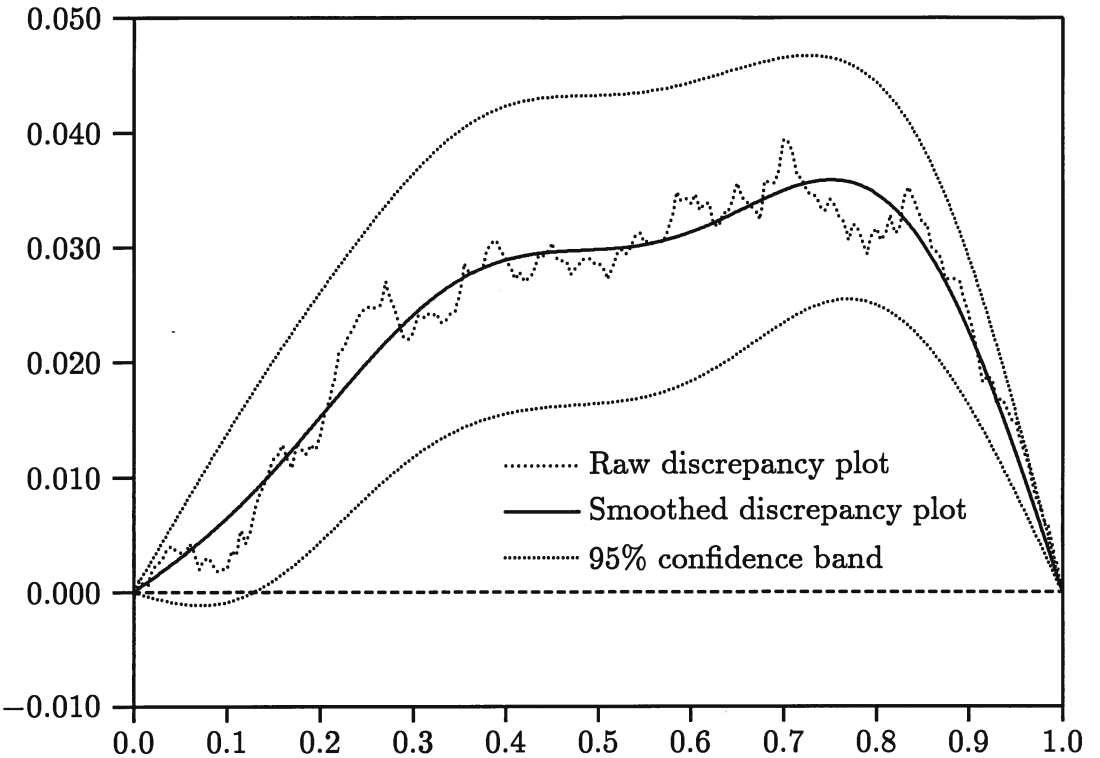


Figure 5. ***P*** **value discrepancy plots for DLR IM test,** $n = 200$**,** $k = 5$

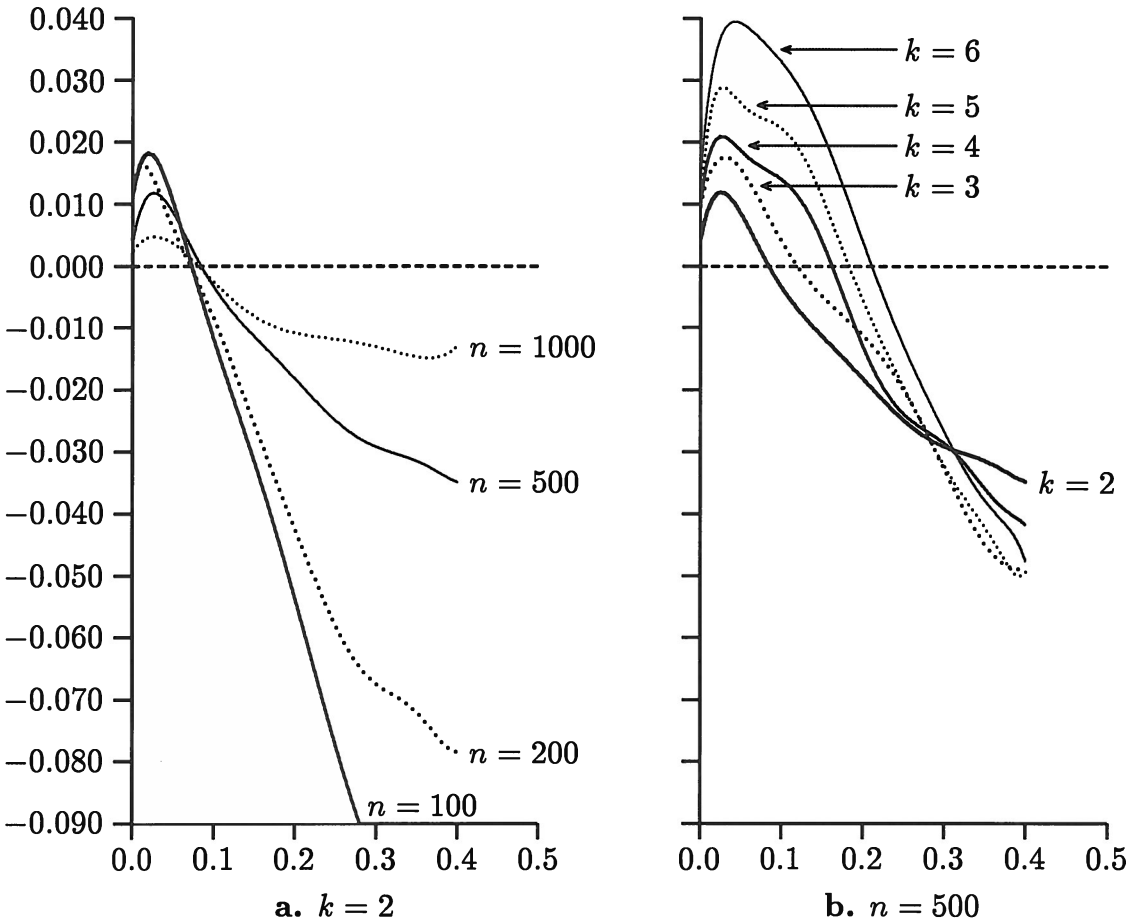


Figure 6. Smoothed P value discrepancy plots for ES IM tests

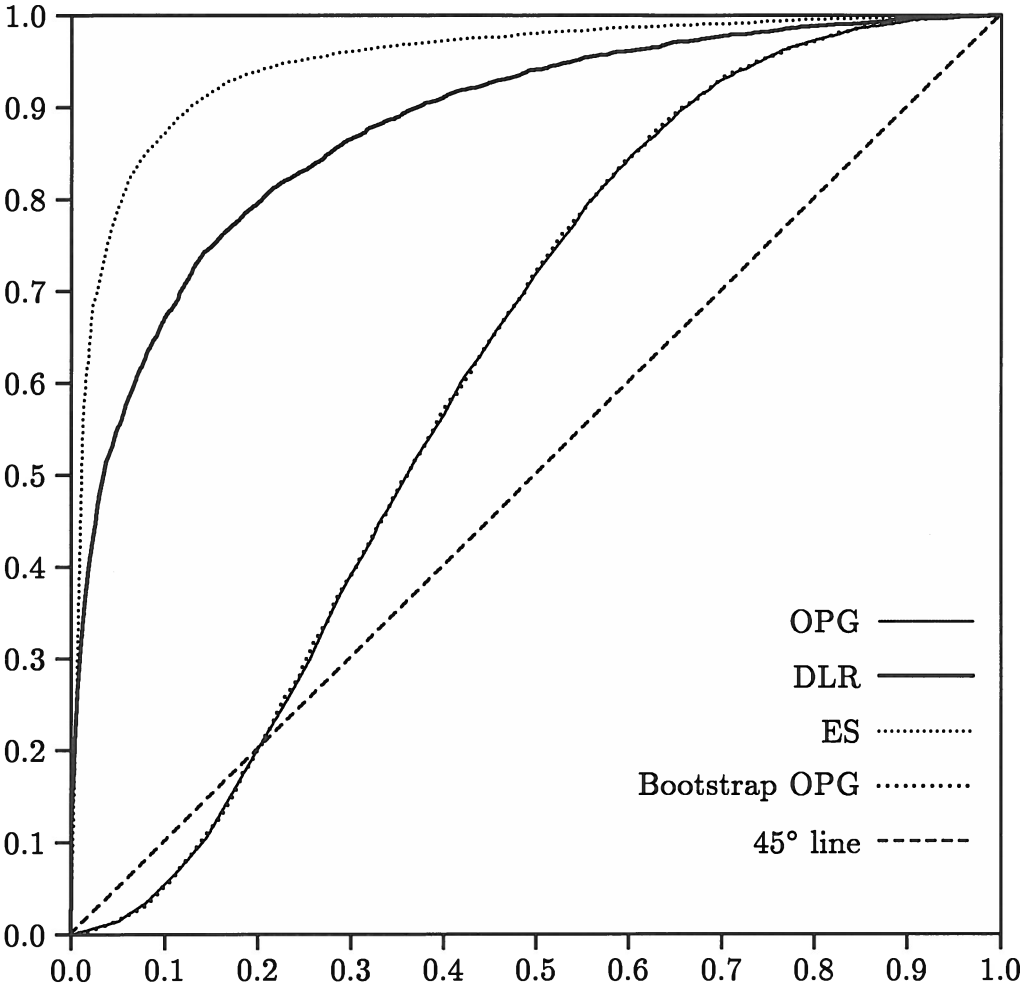


Figure 7. Size-power curves, kurtosis, $k = 2$, $n = 100$

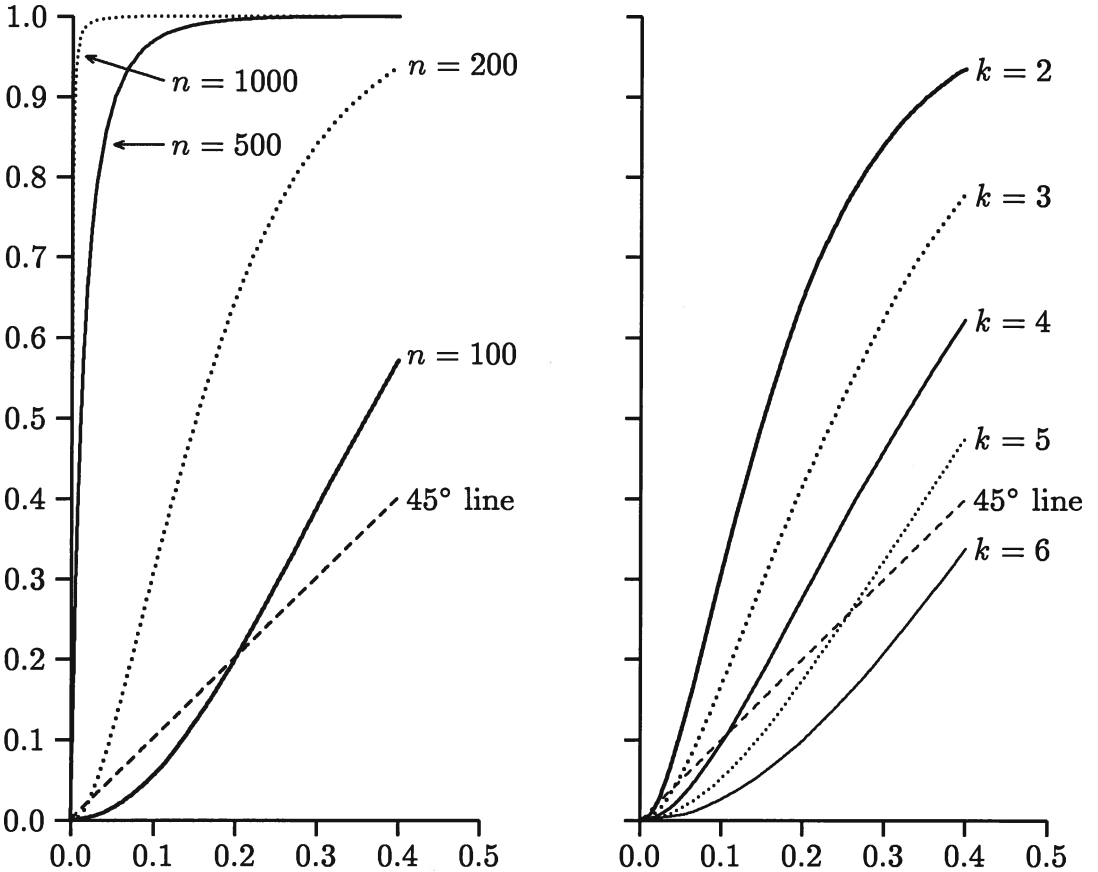
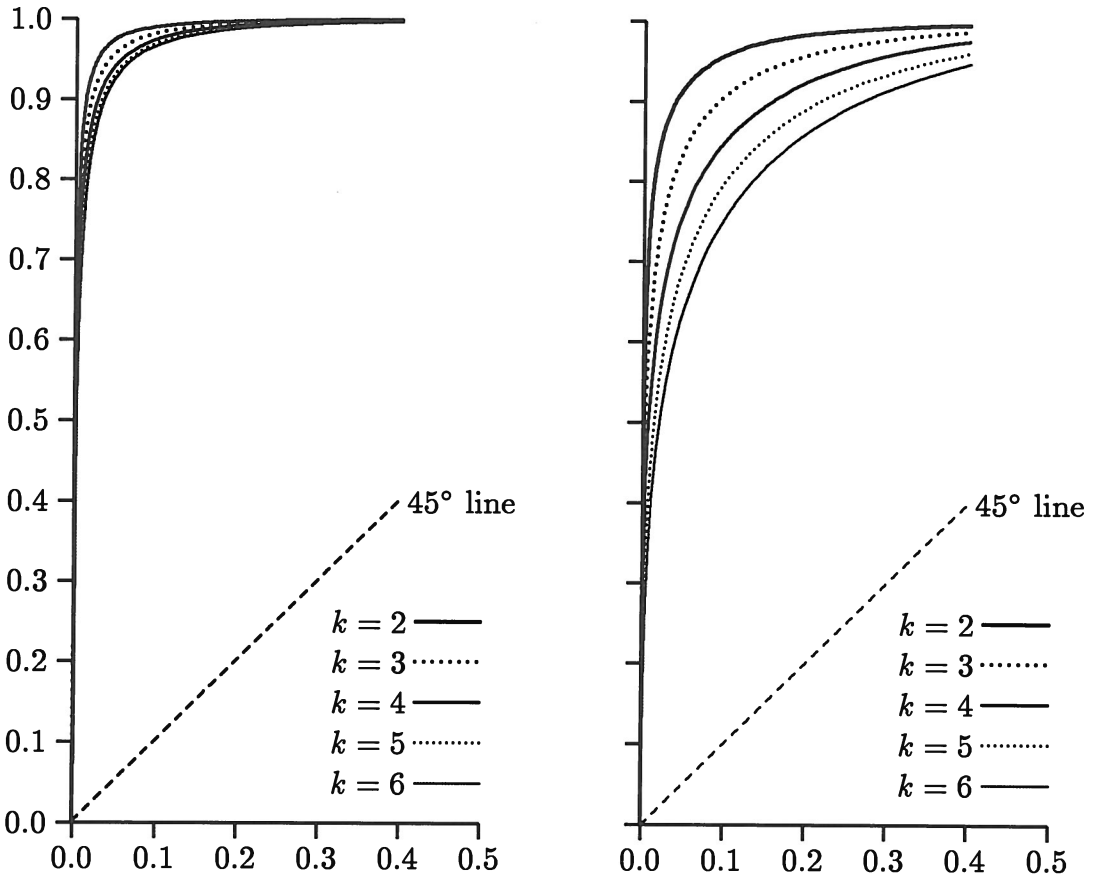


Figure 8. Size-power curves for OPG IM tests



a. ES Form **b.** DLR form

Figure 9. Size-power curves for IM tests, $n = 200$