Econometrica, Vol. 59, No. 6 (November, 1991), 1591–1599

UNDERSTANDING UNIT ROOTERS: A HELICOPTER TOUR

BY CHRISTOPHER A. SIMS[1] AND HARALD UHLIG

While technically p-values should not be interpreted as probabilities, they often are, and their usual asymptotic equivalence to Bayesian posterior tail probabilities provides an approximate justification for doing so. In inference about possibly nonstationary dynamic models the usual asymptotic equivalence fails, however. We show with three-dimensional graphs how it is possible that in autoregressive models the distribution of the estimator is skewed asymptotically, while the likelihood and hence the posterior pdf remains symmetric. We show that no single prior can rationalize treating p-values as probabilities in these models, and we display examples of the sample-dependent "priors" that would do so. We argue that these results imply at a minimum that the usual test statistics and covariance matrices for autoregressions, which characterize the likelihood shape in dynamic models just as in static regression models, should be reported without any corrections for the special unit root distribution theory, even if the corrected classical p-values are reported as well.

KEYWORDS: Unit roots, likelihood principle, Bayesian methods, autoregressive models.

THE USUAL SITUATION IN ECONOMETRIC INFERENCE is that, at least asymptotically, Bayesian probability statements about the unknown parameters conditional on the data are very similar to classical confidence statements about the probability of random intervals covering the true value of the parameter. In time series models with possible unit roots this is not true. In an earlier paper (Sims (1988)) one of the authors of this paper made this point and argued that Bayesian inference for such models was more sensible, as well as much easier to handle analytically, than the classical confidence statements.

Many economists are not used to having to make careful distinctions between probability statements about the location of unknown parameters conditional on the data (Bayesian, or posterior statements) and probability statements about the behavior of statistics in repeated samples conditional on the parameter values (classical confidence, or pre-sample statements). The earlier paper included an example that aimed at guiding intuition about these distinctions, but the example used discrete data and had no evident connection to the unit root time series context. This paper explores in more detail the distinction between confidence statements and probability statements about parameters, in a simple time series model that may show a unit root.

We first summarize graphically the results of a Monte Carlo study of the joint p.d.f. of an unknown autoregressive coefficient ρ and its least squares estimate $\hat{\rho}$, when ρ is treated as uniformly distributed. Bayesian conditional p.d.f.'s for ρ are cross sections of this joint p.d.f. along a fixed-$\hat{\rho}$ line, while classical distributions for $\hat{\rho}$ are sections of the joint p.d.f. along a fixed-ρ line. We display several views of the joint p.d.f., sliced in various ways (the helicopter tour of the title).

[1] This research supported by NSF Grant 8608078.

One-tailed tests of a unit-root null hypothesis against $\rho < 1$ using the appropriate classical distribution theory always accept the null more easily than would tests based on the usual t distribution for the same statistic. Flat-prior Bayesian analysis leads to the usual t tests for generation of posterior probabilities even in dynamic nonstationary models. Therefore use of marginal significance levels under the correct classical distribution theory as if they were posterior probabilities corresponds to using a Bayesian prior distribution which favors larger values of ρ. Since much applied work presents p-values computed with the special unit-root distribution theory, it is useful to know the nature of the prior required to justify treating them as posterior probabilities. We compute these priors—there are many, one for each possible observed value of $\hat{\rho}$.

We discuss the implications for practice of the need to distinguish pre-sample from posterior probabilities in dynamic regression models. Our conclusion is that in reporting results or in making decisions about whether to simplify models by differencing, allowing for co-integration, etc. there is no reason to use the special sampling distribution theory generated under the null hypothesis that unit roots are present. The conventional test statistics and distributions for them retain the same interpretation as descriptors of the likelihood in dynamic as in static regressions. Even statisticians and econometricians uncomfortable with this conclusion should agree that the shape of the likelihood is interesting and that therefore conventional "uncorrected" p values should be reported alongside the special ones based on unit root null hypotheses.

1. THE MODEL

We consider the simple univariate autoregressive model

$$y(t) = \rho y(t-1) + \varepsilon(t), \tag{1}$$

with i.i.d. $\varepsilon(t) \sim N(0, \sigma^2)$. If we observe $y(t)$, $t = 0, \ldots, T$, we can form the least squares estimate $\hat{\rho}$ of ρ. In this model $\sqrt{T}(\hat{\rho} - \rho)$ is asymptotically normal if, say, ε is i.i.d. with finite variance and mean zero and $|\rho| < 1$. When $\rho = 1$, $\hat{\rho}$ is not asymptotically normal. The likelihood, conditional on the initial observation $y(0)$, is Gaussian in shape as a function of ρ, however, and this result does not depend on whether the data is actually generated by a process with a unit root or not.

Because the likelihood depends on both $\hat{\rho}$ and

$$\sigma_\rho^2 = \frac{\sigma^2}{\sum_{t=1}^{T} y(t-1)^2}, \tag{2}$$

there is no one-dimensional way to summarize the sample evidence. In order to develop insight into the relation between Bayesian and classical inference, it is helpful to simplify the situation further. We will consider the situation where

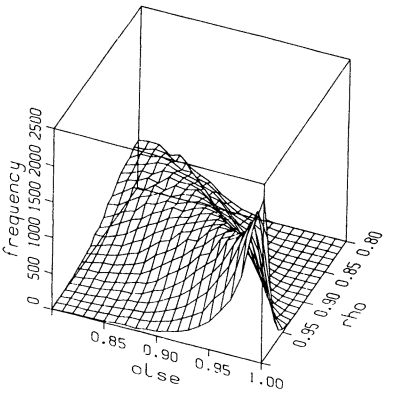

FIGURE 1.—Joint frequency distribution of $\hat{\rho}$ and ρ.

one cannot observe the full sample—only $\hat{\rho}$. We also assume $\sigma^2 = 1$ and is known.[2]

These simplifying assumptions make the shape of the likelihood nonnormal and difficult to derive. Their appeal is only that they make the Bayesian analytical framework consist of a two-dimensional joint p.d.f., that of ρ and $\hat{\rho}$. A function of two arguments is easily visualized as a surface in three dimensions, while a function of three arguments is much harder to visualize.

We can be sure in advance that the likelihood will remain symmetric in ρ around a peak $\hat{\rho}$, because conditional on $\hat{\sigma}$ it has these properties and it therefore will not lose them when $\hat{\sigma}$ is integrated out.

In the next section we will proceed to construct, by Monte Carlo, an estimated joint p.d.f. for ρ and $\hat{\rho}$ under a uniform prior p.d.f. on ρ. We choose 31 values of ρ, from .80 to 1.10 at intervals of .01. We draw 10000 100×1 i.i.d. $N(0,1)$ vectors of random variables to use as realizations of ε. For each of the 10000 ε vectors and for each of the 31 ρ values, we construct a y vector with $y(0) = 0$, $y(t)$ generated by equation (1). For each of these y vectors, we construct $\hat{\rho}$. Using as bins the intervals $[-\infty, .795)$, $[.795, .805)$, $[.805, .815)$, etc. we construct a histogram that estimates the p.d.f. of $\hat{\rho}$ for each fixed ρ value. When these histograms are lined up side by side, they form a surface that is the joint p.d.f. for ρ and $\hat{\rho}$ under a flat prior on ρ.

2. THE HELICOPTER TOUR

Figures 1–5 display different views of the same surface, the estimated joint p.d.f. for ρ and $\hat{\rho}$. Figure 1 shows the surface sliced along the $\hat{\rho} = 1$ and $\rho = 1$ planes. This angle gives a good view of the surface shape. The view from lower down, centered on the corner of the viewing box, is shown in Figure 2. Observe

[2] Because in our Monte Carlo experiments the initial y value on which we condition is $y(0) = 0$ for all samples, the assumption of a particular known value for σ^2 plays no role. We assert it only because the argument as to why it plays no role would take up space and because our choice of σ^2 could make some numerical difference to someone trying to replicate our results.

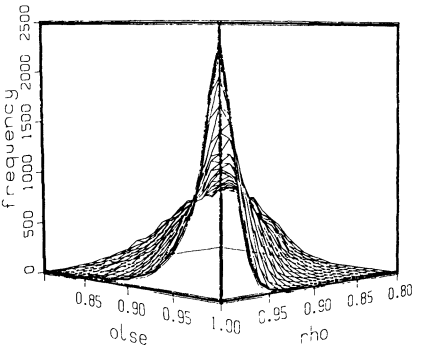

FIGURE 2.—Joint frequency distribution of $\hat{\rho}$ and ρ.

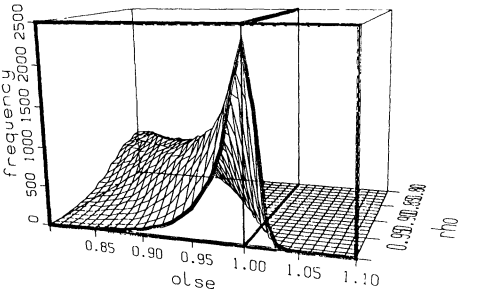

FIGURE 3.—Joint frequency distribution of $\hat{\rho}$ and ρ sliced along $\rho = 1$.

that the distribution of $\hat{\rho}|\rho = 1$, one side of which is the section generated by the left-hand panel in Figure 2, really is more skewed toward lower values than the conditional distribution of $\rho|\hat{\rho} = 1$, one side of which is the section generated by the right-hand panel in Figure 2.

Figure 3 is sliced only along the $\rho = 1$ plane, so the section is just the p.d.f. of $\hat{\rho}|\rho = 1$. Here the well known result that $\hat{\rho}$ is asymmetric, with much more probability below than above one, is easily visible.[3] The section along the $\hat{\rho} = 1$ plane shown in Figure 4 confirms the theoretical result that this p.d.f. is symmetric about $\rho = 1$. Figure 5 shows that the distribution remains symmetric along the $\hat{\rho} = .95$ plane, though it is more dispersed. This result, that the ρ distributions spread out as $\hat{\rho}$ get smaller, is what generates the skewness when the joint p.d.f. is sliced in the other direction. The two sections shown in Figures 3 and 4 are displayed on top of each other in a two-dimensional graph in Figure 6, with both normalized to have the same integral.

[3] The bias in least squares estimates of ρ in this model was noted by Hurwicz (1950). A plot of the asymptotic distribution for $\hat{\rho}$ when $\rho = 1$ reveals that the peak of that density is slightly to the right of 1. Thus, it is probably not the case that the true finite sample distribution for $\hat{\rho}|\rho$ peaks exactly at $\hat{\rho} = \rho$, although this is impossible to determine in our figures.

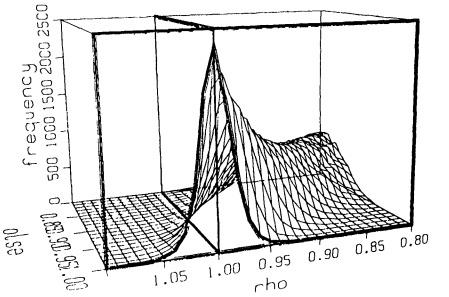

FIGURE 4.—Joint frequency distribution of $\hat{\rho}$ and ρ sliced along $\hat{\rho} = 1$.

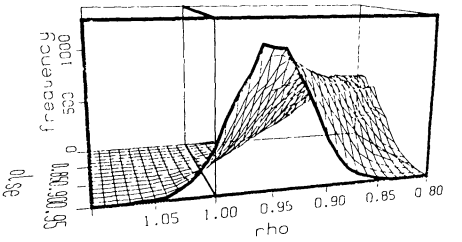

FIGURE 5.—Joint frequency distribution of $\hat{\rho}$ and ρ sliced along $\hat{\rho} = .95$.

Suppose we were studying many instances of the model (1), with true values of ρ drawn at random from a distribution that was uniform over $(.84, 1.06)$[4] and possibly nonuniform, but not too wildly behaved outside that interval. Then any reasonable person would have to agree that what the data imply about the likely location of ρ once we observe $\hat{\rho} = 1$ is given by taking the dotted line in Figure 6 as a p.d.f. for the unknown ρ. The difference between Bayesian and classical statistics is not over the logic of Bayes' rule, but over whether it can legitimately be applied when there is no "objective" source of randomness on which to base the notion of a probability distribution for ρ.

So let us suppose that we really have an application where, say, someone is generating ρ's uniformly by flipping coins or drawing numbers out of a hat. Everyone should agree that, on observing $\hat{\rho} = .95$, our uncertainty about ρ is symmetric about $\rho = .95$, characterized by the p.d.f. shown as the dotted line in Figure 7. What if we nonetheless try comparing the p-values of the null hypotheses $\rho = .9$ and $\rho = 1.0$ by classical procedures? The natural classical test of $\rho = .9$, assuming we can see only $\hat{\rho}$ and not the whole sample, is obtained by normalizing the $\rho = .9$ section of our $\rho, \hat{\rho}$ p.d.f. to integrate to one, then computing the area under the curve to the right of the observed $\hat{\rho}$. This area is the p-value, and one would reject $\rho = .9$ if it fell below some critical level, say $\alpha = .05$. Our Monte Carlo joint p.d.f. implies that the p-value for $\rho = .9$ given an

[4] This is the interval over which the likelihood, normalized to integrate to one, is perceptibly different from zero in Figure 6.

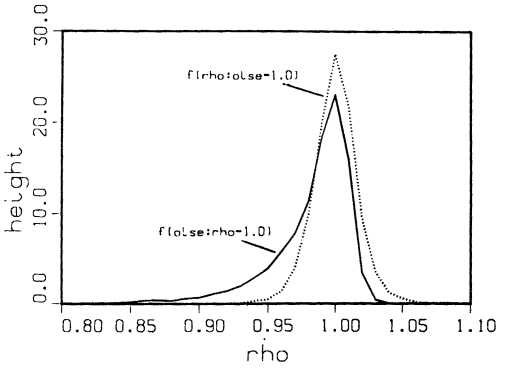

FIGURE 6.—Probability densities for $\rho|\hat{\rho} = 1$ (. . .) and $\hat{\rho}|\rho = 1$ (—).

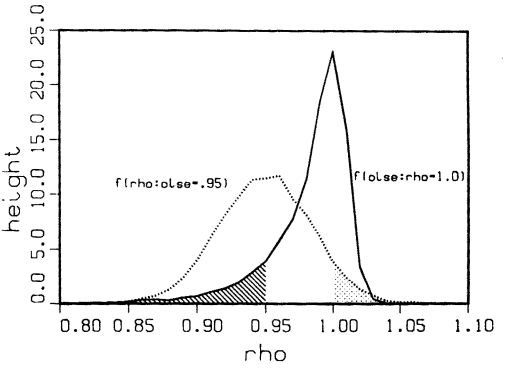

FIGURE 7.—*P*-value vs. posterior probability.

observed $\hat{\rho} = .95$ is .04, while the *p*-value for $\rho = 1.0$ given $\hat{\rho} = .95$ is .12.[5] (The *p*-value for $\rho = 1.0$ when $\hat{\rho} = .95$ is observed is the shaded area under the solid line in Figure 7.) We can reject .9 at the .05 level, in other words, while easily accepting 1.0. The actual conditional probability of $\rho > 1.0$ given observed $\hat{\rho} = .95$ is .07, which is the same as the conditional probability of $\rho < .9$ given $\hat{\rho} = .95$.[6] (This probability is the shaded area under the dotted line in Figure 7.)

How can this be, given that we are already sure that any reasonable person must agree that our beliefs about ρ are symmetrically distributed about $\rho = .95$? The answer is that the *p*-values are distorted by some irrelevant information. It is indeed about equally likely that an observed $\hat{\rho} = .95$ is generated by a true

[5] Here we are using one-sided tests in each case, but with the lower tail the rejection region for $\rho = 1$ and the upper tail the rejection region for $\rho = .9$. All the calculations reported in this paragraph are based on a second Monte-Carlo simulation, different from the one used to generate the graphs. There are, as in the other simulation, 10000 ε vectors, but the $\hat{\rho}$ bins used here are (.949, .950], (.950, .951], etc.

[6] This computed probability involves some interpolation. It averages the two separate probabilities from the Monte Carlo study for $\rho \leq .9$ and $\rho > 1.0$. The raw Monte Carlo data gives $P[\rho \leq .9|\hat{\rho}$ in (.949, .950]] = .070, $P[\rho \leq .9|\hat{\rho}$ in (.950, .951]] = .065, $P[\rho > 1.0|\hat{\rho}$ in (.949, .950]] = .067, $P[\rho > 1.0|\hat{\rho}$ in (.950, .951]] = .067.

$\rho = 1.0$ or a true $\rho = .9$. However $\hat{\rho}$'s much below .95 are much more likely given $\rho = 1.0$ than are $\hat{\rho}$'s much above .95 given $\rho = .9$. In this particular sample we have observed $\hat{\rho} = .95$, not $\hat{\rho}$ much above or much below .95; for deciding what this sample tells us about ρ, the implications of the competing hypotheses about $\hat{\rho}$'s we have not observed are irrelevant.

3. IMPLICIT PRIORS

In the standard normal linear regression model, and asymptotically in most econometric applications, Bayesian probability statements about the location of ρ approximately coincide with corresponding p-values. It is well known that exact coincidence of one-tailed p-values with Bayesian posterior probability statements occurs in the standard normal linear regression model when the prior is flat both on the regression parameters and on the log of residual variance. Pratt (1965) and the following discussion considers conditions under which p-values and posterior probabilities will approximately coincide. Some econometricians, including the authors of this paper, think that it is quite common for applied researchers, even when trained classically, to interpret p-values as if they were posterior probabilities. In any case, since p-values are so widely reported, Bayesian statisticians will be interested in the conditions under which they can be interpreted as approximate posterior probabilities.

In our application the question is, under what conditions does an observation, say, that $\hat{\rho} = .95$, which has a one-sided p-value of .12 on the null hypothesis $\rho = 1$, suggest that the probability of $\rho \geq 1$ is about .12? If we systematically interpreted p-values as estimated posterior probabilities in this way, having observed $\hat{\rho}$ we would for each ρ^* use as an estimate of the probability that $\rho \geq \rho^*$, the marginal significance level for the observed $\hat{\rho}$ in a one-tailed test of $\rho = \rho^*$ as the null hypothesis. This amounts to summing the joint $\rho, \hat{\rho}$ p.d.f. along each constant-ρ line to form a family of c.d.f.'s, but then treating the values of these c.d.f.'s along a constant-$\hat{\rho}$ line as if they formed a c.d.f. for ρ conditional on the observed $\hat{\rho}$.[7] As we have already seen, for $\hat{\rho}$ near one the resulting c.d.f. puts much more weight on $\rho > 1$ than $\rho < 1$, even though a flat prior would imply a conditional c.d.f. symmetric about $\rho = 1$. It may be of interest to see what prior is implicit in inference based on treating p-values as generating a c.d.f. and how the implied prior shifts as the observed $\hat{\rho}$ changes.

Letting $h(\hat{\rho}|\rho)$ be the conditional p.d.f. of $\hat{\rho}$ given ρ, the pseudo-p.d.f. for ρ we are considering is

$$(3) \qquad g(\rho|\hat{\rho}) = \frac{\partial}{\partial \rho} \int_{\hat{\rho}}^{\infty} h(s|\rho)\, ds.$$

[7]In general, it would not necessarily emerge that the resulting pseudo-c.d.f. satisfies the basic properties of a c.d.f.—that $F(\rho) \to 0$ as $\rho \to -\infty$, $F(\rho) \to 1$ as $\rho \to +\infty$, and F be monotone increasing. The first two conditions are clearly met here, since $P[\hat{\rho} < \rho^*|\rho] \to 1$ as $\rho \to -\infty$ and $P[\hat{\rho} < \rho^*|\rho] \to 0$ as $\rho \to \infty$. The monotonicity condition would follow if one-tailed tests which reject $\rho = \rho^*$ in favor of $\rho < \rho^*$ when $\hat{\rho} < \pi$ were unbiased in the classical sense for every ρ^* and π. This condition seems very likely to be true, and certainly our calculations provide no evidence against it.

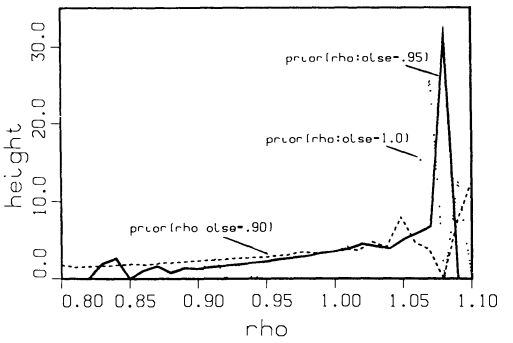

FIGURE 8.—Implicit prior probability densities.

The actual conditional p.d.f. for $\rho|\hat{\rho}$ based on a flat prior over ρ is proportional to $h(\hat{\rho}|\rho)$. For g to emerge as the conditional p.d.f. for $\rho|\hat{\rho}$, therefore, requires that the prior p.d.f. on ρ be proportional to $g(\rho|\hat{\rho})/h(\hat{\rho}|\rho)$. We make an approximate calculation of this implied prior p.d.f. by cumulating our Monte Carlo estimate of h along constant-ρ planes, then differencing the result along constant-$\hat{\rho}$ planes, finally dividing by the original estimated h. For unlikely values of $(\rho, \hat{\rho})$, these estimates are ratios of small numbers with high proportional standard errors. Thus in the tails the estimates are quite erratic.

The results are displayed in Figure 8. One can see that the p.d.f. shifts increasing weight toward the region above $\rho = 1$ as $\hat{\rho}$ gets closer to 1. For $\hat{\rho} = .95$ the implicit prior makes ρ's around 1 two to three times more likely than ρ's around .9. Furthermore, the prior p.d.f.'s for all $\hat{\rho}$ values keep increasing in the region above $\rho = 1$ for as far as the estimates retain any reliability. Thus naive use of classical tests' p-values not only gives special prior weight to $\rho = 1$, it implies a priori belief that a ρ of 1.05 is more likely than a ρ of .95.

4. CONCLUSION

We have illustrated graphically in a simple model how it can be that the likelihood function—the p.d.f. of the data as a function of the parameter, with observed data fixed—is symmetric about its maximum despite pronounced asymmetry in the p.d.f. of the maximum likelihood estimate of the parameter. We have also shown in Section 3 that naive use of p-values as measures of the strength of sample evidence against various parameter values cannot be rationalized as Bayesian under any single choice of prior distribution.

These results reinforce the point that dynamic regression models are a rare instance where Bayesian or other likelihood based forms of inference are not even approximately the same as classical hypothesis testing. Even if the simple context of linearity, Gaussian disturbances, and conditioning on initial conditions is maintained, classical small-sample distribution theory for autoregressions is complex and little used in practice. Classical asymptotic theory breaks discontinuously at the boundary of the stationary region, so the usual simple

normal asymptotic approximations are not available. The likelihood function, however, is well known to be the same in autoregressions and nondynamic regressions, assuming independence of disturbances from lagged dependent variables. Thus inference satistfying the likelihood principle (see Berger and Wolpert (1984)) has the same character in autoregressions whether or not the data may be nonstationary. A t statistic of 3.1 or an F statistic of 1.7 tell us the same thing about the shape of the likelihood in an autoregression as in a regression on exogenous variables.

Many econometricians, ourselves included, will conclude that the complicated apparatus of classical unit root asymptotics is of little practical value. Even econometricians who do not accept this conclusion, however, should agree that the likelihood function's shape is valuable information. It should therefore be standard reporting practice to present information allowing convenient assessment of likelihood shape. In particular, when linear hypotheses on autoregressive systems are being tested, the values or conventional p-values of t and F statistics should be reported, not the classical unit-root asymptotic p-values in isolation.

Dept. of Economics, Yale University, 37 Hillhouse Ave., New Haven, CT 06520-1962, U.S.A.

and

Dept. of Economics, Princeton University, 112 Fisher Hall, Princeton, NJ 08544-1021, U.S.A.

Manuscript received February, 1989; final revision received February, 1991.

REFERENCES

BERGER, JAMES O., AND ROBERT L. WOLPERT (1984): *The Likelihood Principle*, Institute of Mathematical Statistics Lecture Notes—Monograph Series, Volume 6. Hayward, Calif.: Inst. of Math. Stat.

HURWICZ, LEO (1950): "Least Squares Bias in Time Series," in *Statistical Inference in Dynamic Economic Models*, Cowles Commission Monograph No. 10, ed. by T. C. Koopmans. New York: Wiley.

PRATT, JOHN (1965): "Bayesian Interpretation of Classical Inference Statements," *Journal of the Royal Statistical Society*, Series B, 169–192.

SIMS, CHRISTOPHER A. (1988): "Bayesian Skepticism on Unit Root Econometrics," *Journal of Economic Dynamics and Control*, 12, 463–474.