ECO 513 Spring 2015

Asymptotics

Christopher A. Sims
Princeton University
sims@princeton.edu

April 18, 2018

Frequentist asymptotics in the simplest AR case

$$y_t = \rho y_{t-1} + \varepsilon_t \,, \tag{$*$}$$

with ε_t the innovation in y.

$$\hat{\rho}_T = \rho + \frac{\sum_1^T \varepsilon_t y_{t-1}}{\sum_1^t y_{t-1}^2}$$

Martingale Central Limit Theorem

If z_t is stationary, ergodic, has finite variance σ^2, and satisfies $E_t z_{t+1} = 0$ (so $\sum_1^t z_s$ is a martingale), then

$$T^{-1/2} \sum_1^T z_t \xrightarrow[T \to \infty]{\mathcal{D}} N(0, \sigma^2)\,.$$

So if ε_t is stationary and ergodic, $|\rho| < 1$ and y_0 is drawn from the stationary distribution, the numerator of $\sqrt{T}(\hat{\rho}_T - \rho)$ is asymptotically normal. Ergodicity implies

$$\frac{1}{T} \sum_1^T y_{t-1}^2 \xrightarrow[T \to \infty]{P} E[y_t^2]\,.$$

Therefore

$$T^{1/2}(\hat{\rho}_T - \rho) \xrightarrow[T\to\infty]{D} N\left(0, \frac{\sigma^2}{1-\rho^2}\right) .$$

The unit root case

This is the $\rho = 1$ case. The numerator still converges under weak regularity conditions. We can no longer normalize by $\sqrt{T}$, though. The elements in the sum in the numerator of $\hat{\rho} - \rho$ are $\varepsilon_t y_{t-1}$ and the variance of the t'th term in the sum, assuming the ε_t's are martingale-differences and have constant variance σ^2, is $\sigma^2(E[y_0^2] + t\sigma^2)$. The sum of these terms over $t = 1, \ldots, T$ is $\sigma^2(TE[y_0^2] + \frac{1}{2}(T^2 + T))$. Thus

$$\operatorname{Var}\left(\frac{1}{T\sigma} \sum_{t=1}^{T} \varepsilon_t y_{t-1}\right) \xrightarrow[T \to \infty]{} 1\,.$$

The regularity conditions that need to be checked can be seen in detail in Brown (1971). Sufficient conditions are that ε_t itself is stationary and ergodic.

Why asymptotic theory is different when $\rho = 1$ for frequentists

We have just verified that we still get frequentist asymptotic normality for the numerator of $\hat{\rho} - \rho$ when $\rho = 1$. So the problem has to be the denominator, $\sum y_{t-1}^2$. For each t, $E[y_{t-1}^2] = t\sigma^2$. Thus $E[\sum_1^T y_{t-1}^2] = T \cdot (T-1)/2$. If we divide the denominator by T^2, then, we get an expression whose expectation converges to a constant, but unlike the stationary case, the expression does not converge in probability to a constant. It converges in distribution to the distribution of $\int_0^1 W_s^2 \, ds$, where W_s is a Wiener process. Furthermore, the random variable is not independent of the numerator. Correct frequentist distribution theory therefore must derive the asymptotic joint distribution of the two random variables $T^{-1} \sum \varepsilon_t y_{t-1}$ and $T^{-2} \sum y_{t-1}^2$ and use this joint limit to determine the asymptotic distribution of $T(\hat{\rho} - \rho)$.

A curious point: it might not be different after all.

Suppose sample size T is not fixed in advance, but instead is chosen by a rule that depends on the observed data. The likelihood then is still $p(Y_T \mid \theta)$. Since T is a known function of Y_T, the fact that it is random across samples makes no difference to likelihood-based inference or, therefore, to the implications of the data for decisions. This is in a way intuitive: there is no extra information, beyond that in Y_T, in T. So suppose our sample $Y_T = \{y_1, \ldots, y_T\}$ has been generated from the autoregressive model $(*)$ by increasing T until $\sum y_{t-1}^2 = K$, and then stopping. This will not affect the likelihood or Bayesian inference; the likelihood is the same as if T were fixed non-randomly a priori. But the frequentist distribution of the estimator is changed. The denominator of $\hat{\rho} - \rho$ is "almost" non-random, and has less and less randomness in it as we choose K larger. The numerator is still the sum of a martingale difference sequence and hence converges to normality. (It would not be a martingale difference sequence conditional on knowledge of T, but T depends on the whole sample, and is thus not in the information set at $t < T$. At each date t, given that the sum of y_{t-1}^2 values to that date is still below K, $E_t[\varepsilon_{t+1} y_{t-1}] = 0$.) So in this case the frequentist asymptotic distribution theory for $\hat{\rho}$ is back to the normal distribution theory that coincides with the Gaussian likelihood-based posteriors for $\rho \mid Y_T$.

Arguably in practice sample sizes are in fact random. Researchers commonly first collect easily available data and, if the data turn out to contain inadequate variation to sharply estimate the parameters of interest, extend the data set. This does not amount to a fixed rule connecting sample size to Y_T, but it also does not amount to fixing T non-randomly in advance. Frequentists should (but do not in practice) worry about this. For Bayesians it makes no difference to inference.

Bayesian Asymptotics

Full information Posterior distributions, under reasonable regularity conditions, converge to normality.

Limited information Posterior distributions, conditioned on statistics that have limiting normal sampling distributions, under reasonable regularity conditions, converge to the distribution obtained by treating the limiting distributions as exact.

Full information

One of the functions of frequentist asymptotics is to deliver simple, usable distributions where finite-sample distributions are inconvenient or hard to derive. Full information Bayesian asymptotics has the same function. The Bayesian asymptotics asserts that with high pre-sample probability the second order Taylor expansion of the likelihood function becomes increasingly accurate as sample size increases. This suggests maximizing likelihood, forming the second derivative of the log likelihood at the peak, and using minus the inverse of the second derivative matrix as the covariance matrix of the parameter about the MLE, which is the same sequence of calculations implied by use of frequentist asymptotics for the MLE.

The Bayesian asymptotics, though, are just a computational approximation and can be checked for validity in any given sample. That is, one can explore the actual shape of the likelihood to check that the normal approximation is accurate. This is not possible with frequentist asymptotics, because frequentist asymptotics makes no assertion about any particular sample — only about behavior of the estimator across multiple samples. Showing that the likelihood in a particular sample is, say, multiple-peaked, cannot be interpreted as evidence that frequentist asymptotic theory does not apply in that sample, whereas it is evidence that the Bayesian asymptotic computational approximation does not apply.

Kim (1994) shows that the shape of the likelihood converges to Gaussian form in autoregressive models, even when the model contains unit roots. (When the model contains roots larger than one in absolute value, so that it is exponentially explosive, the shape of the likelihood and the distribution of the estimators depends on the distribution of the shocks, even asymptotically.) Sims, Stock, and Watson (1990) displays the form of the frequentist asymptotics for a general multivariate autoregressive model.

Limited information

Full information Bayesian asymptotics assumes that the likelihood function is known and corresponds to the actual data generating process. Frequentist asymptotics often is used to justify a claim that frequentist asymptotic distributions can be used under “weak assumptions”. For example, in the standard linear model, the frequentist distribution theory that arises with i.i.d. zero-mean normal residuals is also asymptotically correct under much weaker assumptions, e.g. that the residuals are stationary martingale differences with finite variance. Bayesian full-information asymptotics applies only to likelihood-based inference and does not in itself allow any weakening of assumptions.

Bayesian limited information asymptotics hypothesizes that the sample data Y_T is not directly available, but instead a vector of statistics $S(Y_T)$

are available and it is known that asymptotically $S(Y_T) \sim N(\theta, \Sigma)$. For a Bayesian, then, this is just an approximately correct model, and the natural way to form a posterior is to treat θ as $N(S(Y_T), \Sigma)$. This can be justified under some regularity conditions. The most important regularity conditions require that the convergence in distribution for $S(Y_T)$ be uniform in θ. This condition is what fails in the unit-root AR case. The details of the regularity conditions for stationary cases are worked out in Kwan (1998).

This kind of Bayesian asymptotics "frees" Bayesian inference from assumptions in the same way that it "frees" frequentist inference from assumptions. But of course basing inference on $S(Y_T)$ rather than on Y_T itself can result in an arbitrarily large penalty on accuracy of inference in a given sample, even when the asymptotic theory applies. That is, even when the asymptotic distribution for $S(Y_T)$ is a very good approximation to its exact finite-sample distribution, a Bayesian who knew the true model might find the posterior given Y_T to be very different from the posterior given

$S(Y_T)$. For example, suppose $y_t = \mu + \varepsilon_t$, with ε_t i.i.d. and equal to σ with probability .5 and equal to $-\sigma$ with probability .5. In sample sizes of a few hundred, the normal approximation to the distribution of the sample mean will be pretty good, and a usual t statistic can generate a confidence interval with accurate coverage probability. But of course anyone knowing the true model and able to observe the full sample would know μ exactly after a finite number of draws, by taking the mean of the largest and smallest observed value of y_t.

Also, the frequentist asymptotic theory can be an arbitrarily bad approximation in finite samples of any given size. There may be, as in the standard linear model, an assumption about the model that would make the asymptotic theory exactly correct (i.i.d. zero-mean normal residuals in the case of the standard linear model). But the asymptotic theory may be a good approximation only in extremely large samples when we deviate far from these assumptions.

For (an extreme) example, suppose $y_t = \mu + \varepsilon_t$, with μ unknown and to be estimated and ε_t i.i.d. Suppose further that the true value of μ is 1.0 and $\varepsilon_t = -1$ with probability .999 and $\varepsilon_t = 999$ with probability .001. The sample mean of observed y's is asymptotically normal here, and the usual t-statistic is justified asymptotically for testing the null hypothesis $\mu = 0$. But of course in sample sizes of less than 1000, it is very likely that every observed value of y_t is zero, so that the sample mean is zero and the sample variance is zero. Inference based on the asymptotic normal theory would be very bad because the asymptotic theory is a poor approximation to the actual finite sample distribution. Bayesian limited information inference, which assumes the asymptotic theory is a good approximation, would be equally bad. So Bayesians or frequentists who rely on asymptotic theory to claim that they are making only "weak" assumptions here are simply avoiding discussion of what range of deviations from the standard normal linear model they think are plausible. At a given sample size, some bounds on the deviations are necessary in order to make the claim that the asymptotic theory applies despite deviations from the normality assumption.

Reconciling Bayesian and frequentist asymptotics when they differ

Bayesian and frequentist asymptotics agree in the "regular" cases covered by our discussion above of limited information asymptotics. That is, the two approaches result in distributions for $\hat{\theta} \mid \theta$ (frequentist) and $\theta \mid \hat{\theta}$ (Bayesian) that have the same form. But there are some important cases that are not regular, mainly where frequentist convergence to the limiting distribution is not uniform over the parameter space. The univariate autoregression, and time series models that allow for non-stationarity more generally, are examples.

Scale changing with location: the archery example

There are N targets numbered left to right, 1 to N, at an archery range. Archers are assigned to targets according to their records, with those known to be most accurate at the right-hand end, the beginners at the left hand end. Each archer j has a probability p_j of hitting the target, and equal probabilities $(1 - p_j)/2$ of missing to the left or right. If an archer misses, her arrow lands between her target and the next one over. p_j is therefore increasing in j. Common sense tells us that if, after all archers have shot an equal number of arrows, we find an arrow between target j and $j + 1$, it most likely belongs to the archer at station j. Her shots have a higher probability of missing, and are thus more likely to be in the grass position than those of archer $j + 1$.

However, if we think of the observed position of the found arrow as a random variable X, we can construct from it an ubiased (in the frequentist sense) estimator of the archer who shot it: allocate the arrow with equal probability to either archer. This is unbiased, because conditional on the true value of the “parameter” (here the index number of the archer who shot the arrow), this estimator is symmetrically distributed, equally likely to err high or low. This unbiased estimator violates the common sense principle above, which says that arrows between targets j and $j+1$ that have missed the target are more likely to belong to the less accurate archer at j than to the more accurate one at $j+1$. In this situation, unbiasedness is an undesirable property of the estimator, at least if our objective is to get the arrow to its owner with as high probability as possible.

The archery example and univariate AR's

The likelihood for equation ($*$), conditional on the initial condition, has the shape of a normal-inverse-gamma density, just as in a standard normal linear regression model. The variance of $\rho \mid \hat{\rho}$ is proportional to $\sigma^2 / \sum y_{t-1}^2$. When ρ is near one, $\sum y_{t-1}^2$ is likely to be much larger relative to σ^2 than when ρ is small. Thus we can think of ρ as an "unbiased estimator" of $\hat{\rho}$, with variance shrinking as ρ and/or $\hat{\rho}$ increase. This is analogous to the archery example, with ρ playing the role of the arrow, and $\hat{\rho}$ playing the role of the archer. Though ρ varies symmetrically around $\hat{\rho}$, because of the way the scale of variation changes with ρ, $\hat{\rho}$ is more likely to be below than above ρ, if we condition on ρ. This is the frequentist downward bias in $\hat{\rho}$. But if what we observe is $\hat{\rho}$, not ρ, then our beliefs about the unobserved ρ are correctly treated as symmetric about $\hat{\rho}$ — just as if we know who shot the arrow, it is correct to consider it equally likely that the arrow will be found to the left as to the right of the target, even though if we know only where the arrow is — between targets — it is more likely that the arrow was shot by the archer to the left.

This reasoning (without the archery story) is worked out with 3-d graphs in Sims and Uhlig (1991).

A gap in the literature

The articles cited above cover Bayesian limited-information asymptotics for AR models in the stationary case (Kwan) and Bayesian full-information asymptotics for such models in the stationary and non-stationary case (Kim Jae-Young). We know that the posterior conditional on the data is Gaussian in the model with normal residuals (a full-information result), but there does not seem to be any paper showing that $\rho \mid \hat{\rho}$ is asymptotically normal when residuals are not normal. In other words, limited information asymptotics for the non-stationary case is not completely worked out. Kim (2002) and others have considered what models and priors imply optimal Bayesian inference that is asymptotically equivalent (in a frequentist sense) to GMM. These results apply to autoregressive time series models, but they are not “limited information asymptotics” in the sense we discuss above. What is missing is a discussion of Bayesian inference conditional on estimators that are not optimal, like OLS estimators of a model with t-distributed disturbances. For stationary models, Kwan covers this case, but there is no corresponding result for unit-root models.

*

References

BROWN, B. (1971): “Martingale Central Limit Theorems,” *The Annals of Mathematical Statistics*, 42(1), 59–66.

KIM, J. Y. (1994): “Bayesian Asymptotic Theory in a Times Series Model with a Possible Nonstationary Process,” *Econometric Theory*, 10(3), 764–773.

KIM, J.-Y. (2002): “Limited information likelihood and Bayesian analysis,” *Journal of Econometrics*, 107, 175–193.

KWAN, Y. K. (1998): "Asymptotic Bayesian analysis based on a limited information estimator," *Journal of Econometrics*, 88, 99–121.

SIMS, C. A., J. STOCK, AND M. WATSON (1990): "Inference in Linear Time Series Models with Some Unit Roots," *Econometrica*, 58, 113–144.

SIMS, C. A., AND H. D. UHLIG (1991): "Understanding Unit Rooters: A Helicopter Tour," *Econometrica*, 59(6), 1591–1599.