Bayesian Model Comparison and Validation

John Geweke, University of Iowa
john-geweke@uiowa.edu

December 22, 2006

Models are the venue for much of the work of the economics profession. We use them to express, compare and evaluate alternative ways of addressing important questions. Applied econometricians are called upon to engage in these exercises using data and, often, formal methods whose properties are understood in decision-making contexts. This is true of work in other sciences as well. There is an enormous literature on alternative formal approaches to these tasks, and in particular on the relative advantages of Bayesian and frequentist methods. By "Bayesian" I mean statistical inference that reaches a conclusion by means of a conditional distribution of unknown quantities given known quantities and model specifications. This conditional distribution follows from applying Bayes's theorem to the joint distribution of known and unknown quantities in the model specifications. Known quantities, including data, are treated as observed values of random variables. Unknown quantities, including functions of parameters and as yet unobserved data, are treated as unobserved random variables. By "frequentist" I mean inference derived from the distribution of statistics in repeated sampling. This includes the use of hypothesis tests, confidence intervals and p-values.

The literature on the relative advantages of Bayesian and frequentist methods is also very large. Bayesian procedures coincide with models of rational behavior, especially updating information and behavior in uncertain environments, and Bayesian learning is the dominant paradigm in formal modeling of this behavior. It is easy to concoct interesting situations in which Bayesian methods yield reasonable outcomes and frequentist methods do not. On the other hand, Bayesian learning assumes costless cognition, whereas in fact reasonable specification of the joint distribution of known and unknown quantities can be very demanding. Frequentist methods avoid many of these demands, and for this reason they are likely to enjoy continued widespread use even as advances in simulation methods continue to widen the scope of application for Bayesian methods in econometrics and statistics. In this context the relative advantages of Bayesian and frequentist approaches seem clear. This has been reaffirmed in a rich literature that includes Box (1980), Dawid (1982), Rubin (1984), Gelman et al. (1996) and Little (2006). An informal statement of these conclusions helps to explain the organization of the rest of this essay, which will then provide a more precise rendering.

Suppose, first, that conclusions are to be drawn based on a workably small number of well-articulated models. Then the situation for Bayesian inference is not fundamentally altered whether that number is one or several, so long as prior distributions are extended to assign prior probabilities to the respective models. One must (and can) deal with some interesting technical details, but the paradigms of rational behavior and Bayesian learning emerged unscathed. Indeed, they differ little operationally, in

marked contrast to frequentist methods in which non-nested models are particularly awkward and the problem is usually cast artificially as one of model choice. Section I briefly reviews and illustrates Bayesian model comparison in this setting, involving the familiar devices of Bayes factors and posterior odds. Poirier (1988) provides a more extensive introduction.

A significant limitation of this setting is the conditioning on a particular set of well-articulated models. It may well be the case that none of these models account well for "stylized facts" or other aspects of the data that are thought to be important for the purposes at hand. Bayesian methods in particular place probability distributions on all knowable quantities, and these distributions imply restrictions on what may happen, or could have happened. If events transpire that are improbable under all of the well-articulated models, then one has little confidence in the set of models being used. This set of circumstances can be revealed by non-Bayesian methods, for example in frequentist testing of restrictions against unspecified alternatives. A Bayesian model in which these deficiencies do not emerge is said to be well calibrated (Dawid, 1984; Little, 2006). Section II describes and illustrates this process, utilizing prior and posterior predictive distributions.

One conclusion might be that inference under a specified set of models should be Bayesian, but that assessment of these models can and should involve frequentist ideas. (For an alternative conclusion see Poirier, 1988, especially pp. 138-141.) Such a synthesis of Bayesian and frequentist ideas has been advocated by many, including prominent Bayesians like Berger (2000). This is an interesting challenge to the conjecture that economists and other scientists act like Bayesians, rather than frequentists. There is no doubt that these investigators look for deficiencies in well specified models using frequentist methods, using hypothesis tests and p-values as well as less formal steps. But judicious application of frequentist methods can also amount to Bayesian learning using available technology, an idea familiar to econometricians from Leamer (1978). Section III develops this idea by augmenting the set of complete models with incomplete models, and provides an illustration. The result is a universe of ideas that (to me) is very much like the one in which economists live and develop new ideas using a process that is fundamentally Bayesian. It has some interesting practical implications for model specification checking, discussed in the concluding section.

I. Complete model comparison and averaging

There are J models, indexed $j = 1, \ldots, J$. A complete model A_j specifies a density of a vector of observable data $\mathbf{y}$ in terms of a vector of unobservable parameters and/or latent variables $\boldsymbol{\theta}_{A_j}$, $p\left(\mathbf{y} \mid \boldsymbol{\theta}_{A_j}, A_j\right)$. For example, suppose $\mathbf{y}$ is a sequence of binary outcomes (as in a coin-tossing experiment). Then A_1 could specify an i.i.d. Bernoulli process, $p\left(y_t = 1 \mid \theta_{A_1} = p\right) = p$; A_2 could specify a first-order Markov process with $\boldsymbol{\theta}_{A_2} = (p_1, p_2)'$ and $p\left(y_t = i \mid y_{t-1} = i\right) = p_{i+1}$ $(i = 0, 1)$. A complete model also

specifies a prior distribution of the unobservables, $p\left(\boldsymbol{\theta}_{A_j} \mid A_j\right)$. For example, the prior distribution of p could be uniform on $[0,1]$ in A_1, and the prior distribution of p_1 and p_2 could be independent uniform on $[0,1]$ in A_2. The model A_j thus implies a joint distribution for observables $\mathbf{y}$ and unobservables $\boldsymbol{\theta}_{A_j}$, $p\left(\boldsymbol{\theta}_{A_j}, \mathbf{y} \mid A_j\right) = p\left(\boldsymbol{\theta}_{A_j} \mid A_j\right) \cdot p\left(\mathbf{y} \mid \boldsymbol{\theta}_{A_j}, A_j\right)$. A complete collection of J models provides probabilities $p\left(A_j\right)$, which sum to 1. For example, the prior probabilities for the Bernoulli and first-order Markov process could each be one-half. The complete collection of the J complete models thus implies

$$p\left(\left\{A_j, \boldsymbol{\theta}_{A_j}\right\}_{j=1}^{J}, \mathbf{y}\right) = \sum_{j=1}^{J} p\left(A_j\right) p\left(\boldsymbol{\theta}_{A_j} \mid A_j\right) p\left(\mathbf{y} \mid \boldsymbol{\theta}_{A_j}, A_j\right). \tag{1}$$

Denote the data – i.e., the value of $\mathbf{y}$ actually observed – by $\mathbf{y}^o$; this is a number rather than a random variable. The *likelihood function* for model A_j is $L\left(\boldsymbol{\theta}_{A_j}; \mathbf{y}^o, A_j\right) \propto p\left(\mathbf{y}^o \mid \boldsymbol{\theta}_{A_j}, A_j\right)$, and the *marginal likelihood* is

$$p\left(\mathbf{y}^o \mid A_j\right) = \int_{\boldsymbol{\theta}_{A_j}} p\left(\boldsymbol{\theta}_{A_j} \mid A_j\right) p\left(\mathbf{y}^o \mid \boldsymbol{\theta}_{A_j}, A_j\right).$$

The posterior (i.e., conditional on $\mathbf{y}^o$) model probabilities are, using (1), $p\left(A_j \mid \mathbf{y}^o\right) \propto p\left(A_j, \mathbf{y}^o\right) = p\left(A_j\right) p\left(\mathbf{y}^o \mid A_j\right)$. Corresponding to this decomposition, in comparing models i and j, $p\left(A_i \mid \mathbf{y}^o\right) / p\left(A_j \mid \mathbf{y}^o\right)$ is the *posterior odds ratio* in favor of model i, and it decomposes as the product of the *prior odds ratio* $p\left(A_i\right) / p\left(A_j\right)$ and the *Bayes factor* $p\left(\mathbf{y}^o \mid A_i\right) / p\left(\mathbf{y}^o \mid A_j\right)$. The posterior probabilities $p\left(A_j \mid \mathbf{y}^o\right)$ are directly relevant for inference about any common property ω of the models, for instance the longest succession of identical outcomes in 100 future trials in the binary outcomes example, so long as the models specify $p\left(\omega \mid \boldsymbol{\theta}_{A_j}, A_j\right)$, as is the case in that example. Each model implies a posterior distribution $p\left(\boldsymbol{\theta}_{A_j} \mid \mathbf{y}^o, A_j\right) \propto p\left(\boldsymbol{\theta}_{A_j} \mid A_j\right) p\left(\mathbf{y}^o \mid \boldsymbol{\theta}_{A_j}, A_j\right)$, and therefore a distribution

$$p\left(\omega \mid \mathbf{y}^o, A_j\right) = \int_{\boldsymbol{\theta}_{A_j}} p\left(\boldsymbol{\theta}_{A_j} \mid \mathbf{y}^o, A_j\right) p\left(\omega \mid \boldsymbol{\theta}_{A_j}, A_j\right) d\boldsymbol{\theta}_{A_j}.$$

Then for the complete collection of models $p\left(\omega \mid \mathbf{y}^o\right) = \sum_{j=1}^{J} p\left(A_j \mid \mathbf{y}^o\right) p\left(\omega \mid \mathbf{y}^o, A_j\right)$, a process known as *Bayesian model averaging*.

In the example, suppose that there are 20 observations of the outcome $\mathbf{y}$, and that

$$\mathbf{y}^o = (0,0,0,0,0,0,1,1,1,1,1,1,1,1,1,1,0,0,0,0,0)'. \tag{2}$$

Straightforward calculation (Geweke, 2005, Sections 6.3 and 7.2) yields $p\left(A_1 \mid \mathbf{y}^o\right) = 0.0051$, $p\left(A_2 \mid \mathbf{y}^o\right) = 0.9949$. Since the posterior densities $p\left(\boldsymbol{\theta}_{A_j} \mid \mathbf{y}^o, A_j\right)$ are the

products of beta densities in both models the distribution of ω (the maximum run length in a future sample of size 100) can be determined by straightforward simulation.

II. Model validation with prior and posterior predictive distributions

Box (1980) introduced the idea of checking functions of the form $z_i = g_i(\mathbf{y})$, to be used in model evaluation. Each complete model j induces a prior predictive distribution for z_i, $p(z_i \mid A_j)$, and Box (1980) derives these distributions analytically in some simple examples. With modern (but typically simple) simulation methods these properties can be studied by means of the repeated simulations $\boldsymbol{\theta}_{A_j}^{(m)} \sim p\left(\boldsymbol{\theta}_{A_j} \mid A_j\right)$, $\mathbf{y}^{(j,m)} \sim p\left(\mathbf{y} \mid \boldsymbol{\theta}_{A_j}^{(m)}, A_j\right)$, $z_i^{(j,m)} = g_i\left(\mathbf{y}^{(j,m)}\right)$, $m = 1, 2, \ldots$. In the binary outcomes example, one checking function z_1 might be the sample mean, while a second checking function z_2 might be the maximum run length in $\mathbf{y}$; in (2), $z_1^o = 0.5$ and $z_2^o = 10$. Since model validation is carried out one model at a time, henceforth the notation assumes a single model A, and the illustration for A will be the i.i.d. Bernoulli model with a flat prior and the data (2).

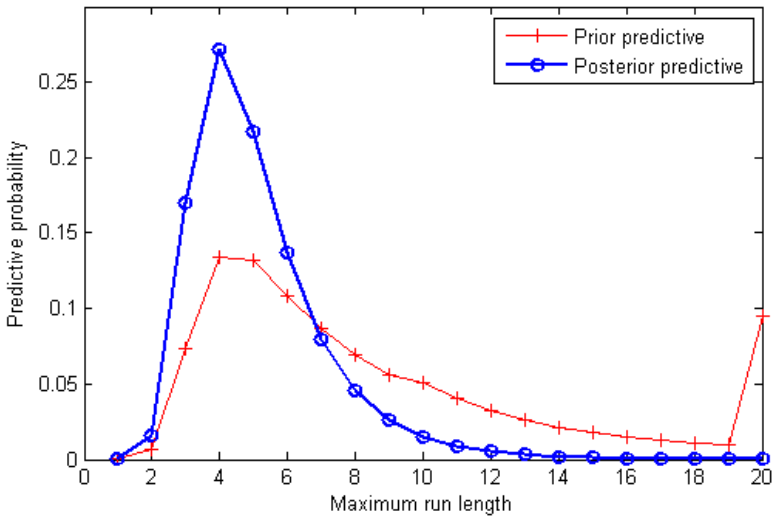
The quantile, or p-value, of the observed outcome $z_i^o = g_i(\mathbf{y}^o)$ can be computed directly from the simulations. In model validation with prior predictive distributions, quantiles near zero or one are taken as cause for concern. Box (1980) was explicit about this use of quantiles, and successive works have been as well, including the recent Bayesian econometrics textbooks by Lancaster (2004) and Geweke (2005). The prior predictive distribution has the same Bayesian interpretation as does the prior distribution itself, and because it elicits the implications and limitations of models without the cost of developing methods of formal inference (either Bayesian or frequentist) it can be very useful at an early stage of model development. On the other hand, the use of quantiles is decidedly frequentist in flavor, invoking the probabilities of unobserved events to assess the quality of the model.

In the binary outcomes example there are 21 possible realizations of z_1, and because the prior distribution of p is uniform these outcomes are equally probable. There are 20 possible outcomes for maximum run length. The crosses connected by thin lines in Figure 1 indicate the prior predictive distribution of this discrete random variable. The distribution is bimodal, because the prior distribution makes reasonable low and high values of p that, it turn, generate run lengths equal to sample size with substantial probability. The observed maximum run length of 10 is well inside the support of this distribution: $P(z_2 \leq z_2^o \mid A) = 0.718$.

Box (1980) also examined the posterior distribution of the checking functions $z_i = g_i(\mathbf{y})$. Note that $\mathbf{y}$, here, is random, corresponding to the outcome of an independent repetition of the same experiment that generated $\mathbf{y}^o$, as contrasted with $z_i^o = g_i(\mathbf{y}^o)$ which is degenerate. The posterior distribution of z_i can be studied by means of the same type of simulation, except that simulation from the posterior distribution, $\boldsymbol{\theta}_A^{(m)} \sim p(\boldsymbol{\theta}_A \mid \mathbf{y}^o, A)$, replaces simulation from the prior distribution.

(Modern Bayesian inference is usually carried out with such posterior simulators.) The same literature is also clear about using the quantile of $p(z_i \mid \mathbf{y}^o, A)$ to validate a model, very low and high quantiles again being cause for concern. Given the posterior simulator the exercise is straightforward and does not require the development of any new distribution theory. The intuition behind the posterior predictive distribution is compelling: if the model and data predict that something very different will nearly always happen to z_i in an independent repetition of the experiment, that should be cause for concern about the specification of the model. Exercises with posterior predictive distributions are frequentist at their core, however, whether one uses quantiles or not.

Figure 1. Predictive distributions of maximum run length



In the i.i.d. Bernoulli example, the posterior predictive distribution of z_1 is centered at $z_1^o = 0.5$, and is very close to a discrete approximation of a Gaussian distribution. Consequently the posterior predictive quantile of z_1^0 includes 0.5. The posterior predictive distribution of the maximum run length z_2 is indicated by the circles connected with solid lines in Figure 1. It differs substantially from the prior predictive distribution: $p(z_2 \geq z_2^o = 10) = 0.037$, an indication that there might be difficulties with the i.i.d. Bernoulli specification. Whereas the prior distribution for p is uniform, the posterior distribution of p is centered at $p = 0.5$ and assigns very low probability to those values of p near 0 and 1 that generated substantially larger prior probabilities for long maximum run lengths in Figure 1.

III. Incomplete models and Bayesian model validation

If one were certain that a set of complete models $A_1, \ldots, A_J$ included all possible models for $\mathbf{y}$, no model validation exercise could have a disconcerting outcome. Indeed, it would never be undertaken, but the fact that these exercises are a critical part

of good econometric work reflects the fact that one (or at any rate one's audience) is never so certain. More important, one often has incompletely articulated ideas about reasonable distributions for checking functions z_i $(i = 1, \ldots, n)$. An informal model of some kind for the z_i seems essential to the choice of checking functions, driven by the characteristics that the investigator thinks a good model ought to have. To formalize this notion, associate a model B_i with each checking function z_i, of the form $p(z_i \mid B_i)$. A natural extension of this idea is an incomplete model $p(z_1, \ldots, z_n \mid B)$ for the n checking functions. These models are incomplete, in the sense that they cannot be (or at least have not been) derived as prior predictive distributions from any complete model for $\mathbf{y}$. Therefore the models B_i are not even known to be coherent – i.e., it is not known whether there exists a model A such that $p(\boldsymbol{\theta}_A \mid A)$ and $p(\mathbf{y} \mid \boldsymbol{\theta}_A, A)$ yield $p(z_i \mid A) = p(z_i \mid B_i)$ $(i = 1, \ldots, n)$, or $p(z_1, \ldots, z_n \mid B)$ if a joint incomplete model is being entertained.

The incomplete models B_i allow one to compare the complete models in hand (the A_i) with as-yet incompletely specified alternatives, using formally justified Bayesian methods. The purpose of this exercise is to model what econometricians do implicitly in their day-to-day work, and to elucidate the corresponding formal procedures.

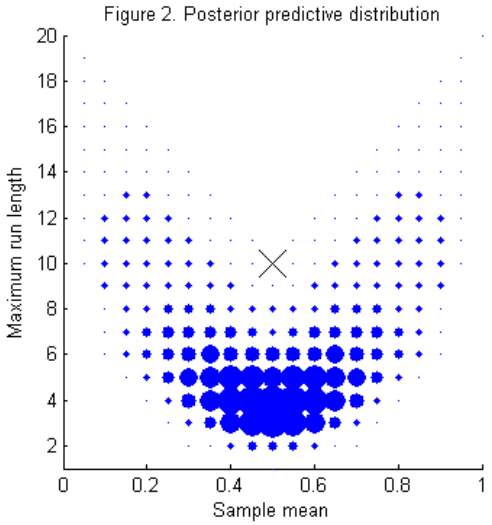
To see how this process works, return to the i.i.d. Bernoulli model of 20 observations on a binary random variable. Suppose that "stylized facts," "theory," or some other set of information credibly independent of the current experiment suggests an incomplete model B_1 for z_1 that assigns equal probability to the 9 outcomes $z_1 = 0.3, 0.35, \ldots, 0.7$ for a sample of size 20. This same information suggests an incomplete model B_2 for z_2 that assigns equal probability to the 5 outcomes $z_2 = 8, 9, 10, 11, 12$ for the same sample size. I intentionally choose incomplete models that are convenient, crude and incoherent, to mimic the behavior of econometricians in more complex settings in which the costs of more elaborate and consistent tentative alternatives make this kind of formulation of incomplete models rational. The reader can easily substitute other densities $p(z_i \mid B_i)$ in the subsequent analysis.

Suppose the econometrician begins with the posterior predictive distribution, which in the i.i.d. Bernoulli example yields $p(z_1 = z_1^o = 10 \mid \mathbf{y}^o, A) = 0.127$. The incomplete-model interpretation of this outcome leads to the Bayes factor $0.127/(1/9) = 1.143$. Any interpretation of this particular result as evidence favoring A over B_1 is misleading: z_1, the sample mean, is also a sufficient statistic in this model, and (in any situation in which the data provide substantial information relative to the prior) the posterior predictive will assign high probability to the observed value of the sufficient statistic relative to other values. This also accounts for the posterior predictive quantile of z_1^o including 0.5. This set of circumstances is characteristic of any function of a vector of sufficient statistics in any model, including models that are much more complex, and stems from the fact that procedures invoking the posterior predictive distribution are inherently non-Bayesian.

For z_2 the same exercise leads to $p(z_2 = z_2^o = 10 \mid \mathbf{y}^o, A) = 0.0151$ (see Figure 1), and the Bayes factor $0.0151/(1/5) = 0.076$. Just as in the case of the p-value

(0.037), this result would alert the econometrician to a potential problem. Indeed the econometrician might well make the observation that z_2 is not ancillary for the i.i.d. Bernoulli model, and therefore the caveats of the previous paragraph imply that the posterior predictive distribution might concentrate probability near observed outcomes – again, a consequence of the non-Bayesian character of this distribution. This observation reinforces the interpretation of the Bayes factor for z_2 as a warning about model specification in this example.

The literature's emphasis on p-values in interpreting the posterior predictive concentrates attention on successive individual models $p(z_i \mid B_i)$, and one often observes checks made one-by-one against a list of diagnostics in this fashion. As previously noted, a joint incomplete model may also be natural, and whereas p-values are inherently univariate, Bayes factors are not. In the example, take $p(z_1, z_2 \mid B) = p(z_1 \mid B_1)\, p(z_2 \mid B_2)$; again, intentionally convenient, crude and incoherent. Figure 2 provides the posterior predictive distribution of (z_1, z_2), which indicates that (z_1^o, z_2^o) – the "X" in Figure 2 – is improbable relative to many other outcomes. In fact $p(z_1^o, z_2^o \mid \mathbf{y}^o, A) = 1.4 \times 10^{-5}$, and the Bayes factor in favor of A as opposed to B is 6.2×10^{-4}.



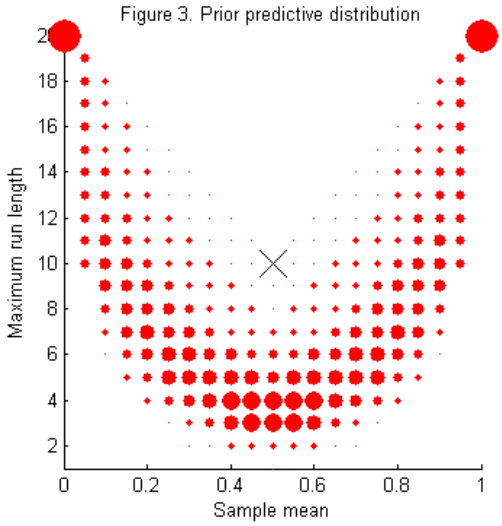


Figure 2 also reveals to the econometrician a systematic relationship between the sample mean z_1 and the maximum run length z_2 that is implied by the posterior distribution of the parameters. The relationship is the distribution

$$p(z_1, z_2) \propto \int_{\boldsymbol{\theta}_A} p(z_1, z_2 \mid \boldsymbol{\theta}_A)\, p(\boldsymbol{\theta}_A \mid A)\, p(\mathbf{y}^o \mid \boldsymbol{\theta}_A, A)\, d\boldsymbol{\theta}_A,$$

where I use $\boldsymbol{\theta}_A$ in place of the parameter p to emphasize the generic character of this kind of exercise. At this point the econometrician very likely suspects that the

difficulty may be in $p(z_1, z_2 \mid \boldsymbol{\theta}_A)\, p(\boldsymbol{\theta}_A)$, which is precisely the prior predictive distribution of z_1 and z_2, shown in Figure 3. (This argument could be streamlined exploiting the fact that z_1 is sufficient for the single parameter p. I avoid that here because this simple model is a stand-in for the more complex models used in most econometric work.) The interpretation of the prior predictive distribution is unambiguously Bayesian (Box, 1980), although the use of quantiles in conjunction with this distribution is not. But quantiles are inessential to the use of any predictive distribution, as illustrated for the posterior predictive, and moreover this is awkward for multivariate distributions in any case. Instead, one can directly compare the prior predictive distribution with the incomplete model or models. The prior predictive distribution, shown in Figure 3, yields $p(z_1^o, z_2^o \mid A) = 5.1 \times 10^{-6}$, and the Bayes factor $p(z_1^o, z_2^o \mid A) / p(z_1^o, z_2^o \mid B) = 2.3 \times 10^{-4}$. Interestingly, in this example, the evidence against the i.i.d. Bernoulli model A in the prior predictive is somewhat stronger than the "evidence" against A in the posterior predictive, the same incomplete model B being used in both cases.

IV. Conclusion

There is a forceful argument in the statistics literature that Bayesian methods are well suited to the analysis of complete collections of complete models, and in particular to the comparison of fully specified models. The view that frequentist methods of some kind are essential to the analysis of specification error (model validation, in the parlance of this literature) enjoys broad support. The analysis here shows that fully Bayesian methods can be employed if one formally introduces the notion of an incomplete model, and argues that the work of econometricians (and, for that matter, all applied statisticians) utilizes this idea informally on a routine basis.

Section II noted that prior predictive distributions are fully Bayesian, whereas posterior predictive distributions are non-Bayesian at their core. Section III illustrated that comparison of complete and incomplete models using the posterior predictive distribution leads back to prior predictive distributions. This has an important practical implication for work with Bayesian models: specification analysis (model validation) can be conducted without recourse to the posterior distribution at all. The implication is practical, because the construction of posterior simulators for complex models typically consumes considerable time, skill and energy relative to simulation from the prior predictive distribution. In most non-Bayesian methods, specification analysis follows inference, a sequence implicit in studying measures of goodness of fit. These sequences seems driven by the technical steps in inference, and the argument made here suggests that posterior predictive analysis is a portion of professional DNA made redundant by Bayesian simulation methods.

The example used in this essay is extremely simple, driven by considerations of transparency and space. In models of complexity typical of much econometric work the list of checking functions $z_i = g_i(\mathbf{y})$ can be become long, and this presents

interesting practical choices. On the one hand, deficiencies of the model may not be revealed for small collections of z_i, as was the case for the maximum run statistic in the example used here. On the other hand, larger collections of z_i encounter limitations of dimensionality in the simulation-based mechanics of representing prior predictive distributions. It seems plausible that more complex models with more attributes z_i that are deemed essential to model well, will require more experimentation. This entails gaining an understanding of the implications of models for observables, an insight that seems essential to model improvement in any event.

REFERENCES

Berger, J.O. "Bayesian Analysis: A Look at Today and Thoughts for Tomorrow." Journal of the American Statistical Association, 2000, 95(452), pp. 1269-1276.

Box, G.E.P. "Sampling and Bayes Inference in Scientific Modelling and Robustness" (with discussion). Journal of the Royal Statistical Society, Series B, 1980, 143, pp. 383-430.

Dawid, A.P. "The Well-Calibrated Bayesian." Journal of the American Statistical Association, 1982, 77(379), pp. 605-610.

Gelman, A., Meng, X.-L., and Stern, H. "Posterior Predictive Assessment of Model Fitness Via Realized Discrepancies" (with discussion). Statistica Sinica, 1996, 6(4), pp. 733-807.

Geweke, J. Contemporary Bayesian Econometrics and Statistics. Hoboken: Wiley, 2005.

Lancaster, T. An Introduction to Modern Bayesian Econometrics. Blackwell, 2004.

Leamer, E.E. "Specification Searches: Ad Hoc Inference with Nonexperimental Data." New York: Wiley, 1978.

Little, R.J. "Calibrated Bayes: A Bayes/Frequentist Roadmap." The American Statistician, 2006, 60(3), pp. 213-223.

Poirier, D.J. "Frequentist and Subjectivist Perspectives on the Problems of Model Building in Economics." The Journal of Economic Perspectives, 1988, 2(1), pp. 121-144.

Rubin, D.B. "Bayesianly Justifiable and Relevant Frequency Calculations for the Applied Statistician." The Annals of Statistics, 1984, 12(4), pp. 1151-1172.