SOME EMPIRICAL EVIDENCE ON MODEL SELECTION RULES

Richard E. Quandt
T. James Trussell

Econometric Research Program
Research Memorandum No. 248

June 1979

Econometric Research Program
Princeton University
207 Dickinson Hall
Princeton, New Jersey

"Some Empirical Evidence on Model Selection Rules"

by

Richard E. Quandt
T. James Trussell*

1. Introduction

An extensive literature deals with the question of how to select one model over another on the basis of observable data. The traditional theory deals with the case of nested hypotheses; i.e. the case in which there exists a random variable y, probability density functions $f(y|\theta_1)$ and $g(y|\theta_2)$ and where the set of density functions $\{f(y|\theta_1)\}$ is a proper subset of the set of density functions $\{g(y|\theta_2)\}$. In the case in which the null hypothesis is not nested, various methods have been suggested. Most attention has been devoted to generalized likelihood ratio tests following the pathbreaking work of Cox (1961, 1962). More recent work along these lines is by Atkinson (1970), Pesaran (1974), and Deaton and Pesaran (1978). Embedding techniques in which a composite density function is set up as a convex combination of the two hypotheses have been studied by Quandt (1974).

Recently Chow (1979) has examined two additional and specific criteria for choosing between competing models. According to Chow we should select the model with the least information in the Kullback-Leibler (1951) sense. This criterion is seen to be closely related in the regression case to the minimum expected squared prediction error criterion. In the process of discussing these criteria, Chow corrects the familiar Akaike criterion. The second criterion examined by Chow is the posterior probability criterion of which the Schwarz criterion is a special case. These two broad criteria differ from one another in one basic respect. The information criterion is relevant if an

*The authors are grateful to Gregory C. Chow for helpful comments. Responsibility for errors is ours alone.

estimated model $f(y|\hat{\theta})$ is to be assessed on the basis of new observations. The posterior probability model is relevant if a sample $(y_1,\ldots,y_n)$ is to be used for assessing a model, where "model" now refers to both the density function $f(y|\theta)$ and the prior density $\pi(\theta)$.

The generalized likelihood and the embedding approaches both permit statistical inference of a traditional sort. In the one case the asymptotic distribution of the generalized likelihood ratio can be obtained, at least in principle; in the embedding case, the maximum likelihood estimate for the mixing parameter λ may be used for inference. At the present time neither the information criterion nor the posterior probability criterion seems to permit classical (non-Bayesian) standard statistical inference. Both criteria lead to an all or nothing choice between two competing models. It is of interest to examine how these criteria perform in some concrete cases. The purpose of this paper is to investigate by some tentative Monte Carlo experiments the behavior of several variants of the two principal criteria discussed by Chow.

2. Statement of the Criteria

The present section states formally the various criteria employed. We assume that observations on a random variable y_i will be generated by one of two density functions $f(y_i|\theta_1)$ or $g(y_i|\theta_2)$. The corresponding likelihood functions are $L(y|\theta_1) = \prod_i f(y_i|\theta_1)$, $L(y|\theta_2) = \prod_i g(y_i|\theta_2)$. Maximum likelihood estimates $\hat{\theta}_1$, $\hat{\theta}_2$ are obtained by maximizing the respective likelihood functions.

1. The likelihood criterion. Form

$$\lambda = \frac{L(y|\hat{\theta}_1)}{L(y|\hat{\theta}_2)}$$

where $L(y|\hat{\theta}_i)$ denotes the value of the likelihood function at $\theta_i = \hat{\theta}_i$, and choose the model given by $f(y_i|\theta_1)$ if and only if $\lambda > 1$.

2. The Akaike criterion. Let k_1 and k_2 be the number of parameters in the vectors θ_1 , θ_2. Then choose the model given by $f(y_i|\theta_1)$ if and only if

$$\log L(y|\hat{\theta}_1) - k_1 > \log L(y|\hat{\theta}_2) - k_2$$

3. The Chow 1 criterion. This criterion is derived from the information criterion and corrects the Akaike criterion. It replaces Akaike's k_1 by $\ell_1 = \text{tr}\{E[\frac{\partial \log f(y|\theta_1)}{\partial \theta_1} \frac{\partial \log f(y|\theta_1)}{\partial \theta_1'}] \cdot [E(-\frac{\partial^2 \log f(y|\theta_1)}{\partial \theta_1 \partial \theta_1'})]^{-1}\}_{\hat{\theta}_1}$ and replaces k_2 by ℓ_2 defined similarly. Then choose $f(y_i|\theta_1)$ if and only if

$$\log L(y|\hat{\theta}_1) - \ell_1 > \log L(y|\hat{\theta}_2) - \ell_2$$

The above expectations are supposed to be taken with respect to the true (unknown) model; hence the expression does not simplify to Akaike's k unless the parameterized model in the brackets (f in the present case) is the true one.[1]

The remaining criteria are more closely related to the posterior probability model. If $p(y|M_j)$ denotes the density of y conditional on the model M_j (f or g) and if $\pi(\theta_j|M_j)$ is the corresponding prior density, one can write (Chow, 1979a)

$$\log p(y|M_j) = \log L(y|\hat{\theta}_j) - \tfrac{1}{2}k_j \log n - \tfrac{1}{2}\log[-\frac{1}{n}\frac{\partial^2 \log L(y|\hat{\theta}_j)}{\partial \theta_j \partial \theta_j'}] + \tfrac{1}{2}k \log 2\pi$$
$$+ \log \pi(\hat{\theta}_j|M_j) + o(n^{-1/2}) \qquad (1)$$

where n is the sample size. We have the following possible criteria.

[1] Of course, when one sets out to evaluate numerically this criterion, the underlying true model is not known. Hence, in the sampling experiments reported below, the expectations were replaced by their sample realizations; i.e. the first expectation is the sample average of $((\partial \log f(y_i|\theta_1)/\partial\theta_1)(\partial \log f(y_i|\theta_1)/\partial\theta_1'))_{\hat{\theta}_1}$ and the second expectation is the negative inverse Hessian evaluated at the $\hat{\theta}_1$ MLE and divided by the sample size.

4. The Chow 2 criterion. Since (1) is not useful if the prior density is diffuse, define the approximation $\log\tilde{p}(y|M_j)$ as $\log p(y|M_j) - \log(\hat{\pi}_j|M_j) - o(n^{-1/2})$ (or alternately as the sum of the first four terms on the right hand side of (1)). Then choose model 1 over model 2 if and only if

$$\log \tilde{p}(y|M_1) > \log \tilde{p}(y|M_2)$$

5. Schwarz criterion. If all terms on the right hand side of (1) are omitted except for the first two, we obtain the Schwarz criterion. Accordingly, choose model 1 if and only if

$$\log L(y|\hat{\theta}_1) - \frac{1}{2}k_1\log n > \log L(y|\hat{\theta}_2) - \frac{1}{2}k_2\log n$$

6. The Chow 3 criterion. Assume that n_1 observations are used to obtain a posterior density for θ_1 and θ_2, leaving $n_2 = n - n_1$ observations for model selection. Then $\pi(\hat{\theta}_j|M_j)$ can be replaced by the posterior pdf for $\hat{\theta}_j$ from the first n_1 observations and the full formula (1) can be used for model selection (except for the term $o(n^{-1/2})$). Accordingly choose model 1 if and only if

$$\log p(y|M_1) > \log p(y|M_2)$$

Note that the idea underlying this set of criteria is that the model which is deemed most likely given the sample will be selected. Since, however, a model with a large number of parameters in general can be expected to perform better, some penalty must be attached to increasing the complexity of the model. Hence, each of the criteria start with the loglikelihood value; they differ only with respect to the penalty imposed. Even the Bayesian criterion can be interpreted in this framework; in particular, the prior can impose a heavy penalty. Nevertheless, since in large samples all criteria are dominated by the loglikelihood term (which is of order n), asymptotically all criteria will pick the true model.

3. Sampling Experiments

Two basic experiments were conducted. In each a pair of density functions was selected. Alternately one and the other in the pair were taken to be the true density and n observations on the random variable were generated from the true density. The various criteria were evaluated and the frequency with which each model was chosen by the several criteria was recorded over 100 replications. The sample size n ranged from 25 to 150 in increments of 25. Values of the appropriate random variables were generated by setting the cumulative distribution function $F(x)$ equal to a uniformly distributed u , distributed on $(0,1)$, i.e. $F(x) = u$, and solving for x . Normally distributed variates were obtained by applying the Box-Muller transformation to uniform variates.

Experiment 1 contrasts the Pareto distribution with pdf

$$f(x) = cx^{-c-1} \qquad 1 < x < \infty \ , \quad c > 0$$

with a shifted exponential distribution with pdf

$$g(x) = be^{-b(x-1)} \qquad 1 < x < \infty \ , \quad b > 0$$

Both densities have exactly one parameter and are nonnested with respect to one another. The actual parameters were $c = 4$, $b = 4$.

Experiment 2 contrasts the normal distribution

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2} \qquad -\infty < x < \infty$$

with a distribution discussed in Goldfeld and Quandt (1979), referred to as the

Sargan distribution, with pdf

$$g(x) = \frac{\alpha e^{-\alpha|x|}}{4}(1+\alpha|x|) \qquad -\infty < x < \infty$$

The Sargan distribution is roughly bell-shaped and has mean zero and variance $4/\alpha^2$. The parameter α was chosen to be 2 so that when the Sargan distribution was chosen to generate the random variables, these would have mean zero and variance 1. When x was normally distributed, the true μ was chosen to be 0 and the true σ^2 to be 1; hence the samples would be roughly comparable in mean and dispersion irrespective of what the "truth" was.

The following interpretations emerge from Table 1. When the truth is Pareto, (1) no criterion is uniformly best, although the likelihood and Chow 2 criteria never are; (2) all criteria tend to improve (nonuniformly) as the sample size increases; (3) for relatively small sample sizes ($n \leq 50$) no criterion is very good, picking the right model at best 4 out of 5 times. When the truth is exponential; (4) the likelihood criterion is the best criterion for all sample sizes; (5) all criteria tend to improve (nonuniformly) as the sample size increases; (6) for small sample sizes no criterion can claim to be acceptable overall. Results not reported in detail indicate that changes in the parameters of the underlying distributions change substantially the absolute performance of the criteria.

From Table 2 we conclude the following. When the truth is normal, (1) the Chow 1, Chow 2, Akaike and Schwarz criteria produce unacceptable results; (2) among these four criteria the Schwarz criterion is always the worst, and the Akaike criterion is always the best, (3) for the likelihood criterion, unlike the other criteria, no improvement occurs as the sample size increases. When the truth is Sargan; (4) all criteria improve with sample size; (5) the Schwarz

Table 1

Comparison of Pareto and Exponential Distributions. Relative Frequency of Choosing the Pareto Distribution

	Truth = Pareto					
	n					
Criterion*	25	50	75	100	125	150
Likelihood	.47	.71	.65	.87	.86	.86
Chow 1	.65	.80	.76	.90	.88	.89
Chow 2	.62	.78	.72	.89	.87	.86
Chow 3**	.69	.71	.81	.83	.85	.92
	Truth = Exponential					
Likelihood	.18	.19	.11	.16	.12	.08
Chow 1	.31	.28	.13	.22	.17	.11
Chow 2	.27	.24	.12	.21	.14	.10
Chow 3**	.39	.32	.22	.19	.17	.15

*The Akaike and Schwarz criteria yield the same answer in the present case as the likelihood criterion.

**Gamma priors were used with different parameters. They all gave substantially similar results. The figures reported are for a gamme prior $\lambda^r x^{r-1} e^{-\lambda x}/\Gamma(r)$ with $r = 3.0$ and $\lambda = 2.5$. The values of n_1 for the various sample sizes were 10, 15, 20, 25, 30, 35.

Table 2

Comparison of Normal and Sargan Distributions. Relative Frequency of Choosing the Normal Distribution

		Truth = Normal				
		n				
Criterion	25	50	75	100	125	150
Likelihood	.93	.89	.88	.92	.86	.95
Chow 1	.53	.49	.63	.69	.73	.84
Chow 2	.62	.47	.63	.66	.66	.80
Akaike	.60	.62	.71	.77	.79	.88
Schwarz	.27	.24	.42	.50	.50	.67
		Truth = Sargan				
Likelihood	.62	.44	.37	.30	.19	.21
Chow 1	.22	.19	.12	.14	.10	.05
Chow 2	.22	.17	.10	.08	.05	.04
Akaike	.23	.23	.20	.16	.12	.09
Schwarz	.13	.10	.04	.06	.03	.02

criterion is uniformly best; (6) the likelihood criterion is unacceptable, picking the wrong model 1 out of every five times even for the largest sample sizes.[2]

4. Conclusions

The sampling experiments were simple. In both basic experiments the two hypotheses were nonnested. In one case the two competing pdf's had the same number of parameters, in the other there was a difference of one. The single most startling conclusion is that when the two experiments are assessed together, no criterion dominates. The performance of the criteria differs between the two experiments and also within an experiment, depending on what the truth is. This lack of symmetry is illustrated by the Schwarz criterion in Table 2 which is worst when the truth is normal but best when it is Sargan.[3] For large samples the Akaike and Chow 1 criteria behave very comparably but not necessarily well; this result suggests that as a practical matter the more easily computed Akaike criterion may be preferred.

Chow has argued forcefully that when selecting a model one might be primarily concerned with prediction; the estimated model which will best predict over future replications is the one selected. Thus selecting the "true" model does

[2] An ideal form (unattainable in practice) of the Chow 1 criterion would be obtained if the expectations ℓ_i are evaluated exactly given knowledge of the true density. This ideal Chow 1 criterion was examined in two cases (truth = Sargan and truth = Pareto) in which the expectation could be computed with reasonable ease. In both cases and for all sample sizes and regardless of whether true or estimated parameter values were employed, the ideal Chow 1 criterion picked the correct model in 89 percent of the cases or more.

[3] This result is perhaps not so surprising when one realizes that the Schwarz criterion imposes the heaviest penalty for extra parameters. In effect, the Schwarz criterion predominantely chooses the smaller model when the sample size is small. As the sample size increases, the loglikelihood term increasingly dominates the $\frac{1}{2}$ k log n term; nevertheless, the smaller model is picked more than under the straight loglikelihood criterion.

not necessarily imply having a model which will predict better. Better prediction may result from dropping parameters from a true model as it reduces the sampling errors of the remaining parameters. Of course, our experiments do not follow the strategy implied by this prediction rule. Instead, we test how well each criterion performs in choosing the true model which generated a particular sample. We feel that in many cases this question is precisely the one which one wants to address. In fact, excepting the special case of regression models which are explicitly worked out by Chow, it is not possible to compute the Chow 1 criterion, since it involves the true unknown model.

References

Atkinson, A.C., "A Method for Discriminating Between Models," Journal of the Royal Statistical Society, Series B, 32(1970), 323-344.

Chow, G.C., "A Reconciliation of the Information and Posterior Probability Criteria for Model Selection," Econometric Research Program, Princeton University, Research Memo No. 234, 1979a.

Chow, G.C., "Selection of Econometric Models by the Information Criterion," Econometric Research Program, Princeton University, Research Memo No. 239, March 1979b.

Cox, D.R., "Further Results on Tests of Separate Families of Hypotheses," Journal of the Royal Statistical Society, Series B, 24(1962), 406-424.

Cox, D.R., "Tests of Separate Families of Hypotheses," Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Vol. 1. Berkeley: University of California Press, 1961, 105-123.

Goldfeld, S.M. and R.E. Quandt, "Recent Problems and Advances in Estimating Disequilibrium Models," paper presented at Western Economic Association Meetings, Las Vegas, June 1979.

Kullback, S. and R.A. Leibler, "On Information and Sufficiency," Annals of Mathematical Statistics, 22(1951), 79-86.

Pesaran, M.H., "On the General Problem of Model Selection," Review of Economic Studies, 41(1974), 153-171.

Pesaran, M.H. and A.S. Deaton, "Testing Non-Nested Nonlinear Regression Models," Econometrica, 46(1978), 677-694.

Quandt, R.E., "A Comparison of Methods for Testing Nonnested Hypotheses," The Review of Economics and Statistics, LVI(1974), 92-99.